ISSN 2415-7740 (Print)
ISSN 2415-7074 (Online)

Belarusian State University of Informatics and Radioelectronics

Open Semantic Technologies for Intelligent Systems

Research Papers Collection

Founded in 2017

Issue 8

Minsk
2024

UDC 004.822+004.89-027.31

The collection includes peer-reviewed articles approved by the Editorial board. Designed for university professors, researchers, students, graduate students, undergraduates, as well as for specialists of enterprises in the field of intelligent systems design.

Editorial address:
220005, Minsk, Platonova str., 39, rm. 606 b

Phone: +375 (17) 293-80-92

E-mail: ostisconf@gmail.com

Web-site: ***http://proc.ostis.net***

The collection is included in the List of scientific publications of the Republic of Belarus for publishing the results of dissertation research in the technical field of science (computer science, computer technology and management)

The collection is included in the Russian Science Citation Index

ISSN 2415-7740 (Print)
ISSN 2415-7074 (Online)

Учреждение образования
«Белорусский государственный университет
информатики и радиоэлектроники»

Открытые семантические технологии проектирования интеллектуальных систем

Сборник научных трудов

Основан в 2017 году

Выпуск 8

Минск
2024

УДК 004.822+004.89-027.31

Сборник включает прошедшие рецензирование и утвержденные Редакционной коллегией статьи.

Предназначен для преподавателей высших учебных заведений, научных сотрудников, студентов, аспирантов, магистрантов, а также для специалистов предприятий в сфере проектирования интеллектуальных систем.

Адрес редакции:

220005, г. Минск, ул. Платонова 39, каб. 606 б

Телефон: +375(17)293-80-92

Электронный адрес: ostisconf@gmail.com

Сайт: ***http://proc.ostis.net***

Сборник включён в Перечень научных изданий Республики Беларусь для опубликования результатов диссертационных исследований по технической отрасли науки (информатика, вычислительная техника и управление)

Сборник включён в Российский индекс научного цитирования

TABLE OF CONTENTS

СОДЕРЖАНИЕ

FOREWORD

The **main practical result** of the current stage of works in the field of Artificial intelligence is not only the creation of the next-generation intelligent computer systems ensuring effective interaction in solving comprehensive problems, but also the creation of a technological complex that ensures the rapid and high-quality building of such systems. This collection of scientific papers "Open semantic technology for intelligent systems design" is dedicated to these issues.

The development of this technological complex requires solving the following problems:

- clear identification of the logical-semantic level of intelligent computer systems, which abstracts from all possible options for the technical implementation of these systems (including the use of fundamentally new computers focused on their hardware support);
- development of an ontology for the design of intelligent computer systems and unification of the description of their logical-semantic models;
- ensuring the platform-independent character of the logical design of intelligent computer systems, the result of which is a unified description of the logical-semantic models of the designed systems;
- use of the methodology for component design of intelligent computer systems, which is based on a constantly replenished library of reusable components of these systems (reusable subsystems, knowledge base components, knowledge processing agents, user interface components);
- ensuring semantic compatibility of reusable components of intelligent computer systems and semantic compatibility of these systems themselves, as well as technologies for their design and support of subsequent stages of their life cycle.

The **main topic** of the collection of scientific papers "Open semantic technology for intelligent systems design" is various aspects of convergence and integration that ensure the transition to intelligent computer systems of a new generation and the corresponding technology of integrated support of their life cycle:

- convergence and integration of various models for information representation and processing in intelligent computer systems of a new generation:
  - convergence and integration of various knowledge types in the knowledge bases of intelligent computer systems of a new generation;
  - convergence and integration of various problem-solving models in intelligent computer systems of a new generation;
  - convergence and integration of various types of interfaces for intelligent computer systems of a new generation;
- convergence and integration of various directions of Artificial intelligence in order to build a general formal theory of intelligent computer systems of a new generation;
- convergence and integration of design technologies for various components of intelligent computer systems of a new generation in order to build a comprehensive technology for designing intelligent computer systems of a new generation;
- convergence and integration of technologies to support various stages of the life cycle for intelligent computer systems of a new generation in order to build a technology for integrated support of all stages of the life cycle for intelligent computer systems of a new generation;
- convergence and integration of various types of human activities in the field of Artificial intelligence (research activities, development of technological complex, applied engineering, educational activities) to increase the level of coherence and coordination of these activities, as well as to increase the level of their complex automation with the help of semantically compatible intelligent computer systems of a new generation;
- convergence and integration of various types and fields of human activities, as well as means of complex automation of this activity with the help of intelligent computer systems of a new generation.

The **main directions** of the collection of scientific papers «Open semantic technology for intelligent systems design»:

- Requirements for intelligent computer systems of a new generation;
- Principles underlying intelligent computer systems of a new generation;

- Requirements for the semantic representation of information. The universal language of the semantic representation of knowledge;
- Ontological stratification of the semantic representation of knowledge bases in intelligent computer systems of a new generation;
- An agent-oriented hierarchical model of hybrid problem solvers based on the semantic representation of knowledge bases;
- An ontological model of multimodal interfaces for intelligent computer systems of a new generation;
- A model of understanding information coming from outside, based on the semantic representation of knowledge bases;
- Flexibility, stratification, reflexivity, and learnability of intelligent computer systems of a new generation;
- Unification, standardization, and semantic compatibility of intelligent computer systems of a new generation — consistency of concepts and terms;
- Interoperability (the ability to interact effectively) of intelligent computer systems of a new generation — the ability to understand each other, negotiate, and coordinate their actions in the collective solution of complex problems;
- An integrated technology for designing intelligent computer systems of a new generation;
- Platforms for the implementation of intelligent computer systems of a new generation. Software platforms and associative semantic computers focused on the implementation of intelligent computer systems of a new generation;
- Convergence of logical-semantic models of intelligent computer systems of a new generation and architectures of associative semantic computers providing interpretation of these models. Wave programming languages;
- Principles underlying the Technology of integrated support of the life cycle for intelligent computer systems of a new generation;
- A global Ecosystem of intelligent computer systems of a new generation, providing complex convergence and automation of all kinds of human activities.

The collection contains 44 articles. The editorial board of the collection thanks all the authors who submitted their articles. For publication, scientific experts selected the best of the submitted papers, many of them were revised in accordance with the comments of the reviewers.

We also thank the experts for their great work in reviewing articles in close cooperation with the authors, which allowed increasing the level of presentation of scientific results and also created a platform for further scientific discussions.

We hope that, as before, the collection will fulfill its main function — to promote active cooperation between business, science, and education in the field of Artificial intelligence.

Editor-in-chief
Golenkov Vladimir

ПРЕДИСЛОВИЕ

Основным **практическим результатом** текущего этапа работ в области Искусственного интеллекта является создание не только интеллектуальных компьютерных систем следующего поколения, обеспечивающих эффективное взаимодействие при решении комплексных задач, но и создание технологического комплекса, обеспечивающего быстрое и качественное построение таких систем. Этим вопросам посвящён, данный сборник научных трудов «Открытые семантические технологии проектирования интеллектуальных систем».

Разработка указанного технологического комплекса требует решения следующих задач:

- чёткого выделения логико-семантического уровня интеллектуальных компьютерных систем, который абстрагируется от всевозможных вариантов технической реализации этих систем (в том числе и от использования принципиально новых компьютеров, ориентированных на их аппаратную поддержку);
- разработки онтологии проектирования интеллектуальных компьютерных систем и унификации описания их логико-семантических моделей;
- обеспечения платформенно независимого характера логического проектирования интеллектуальных компьютерных систем, результатом которого является унифицированное описание логико-семантических моделей проектируемых систем;
- использования методики компонентного проектирования интеллектуальных компьютерных систем, в основе которой лежит постоянно пополняемая библиотека многократно используемых компонентов этих систем (многократно используемых подсистем, компонентов баз знаний, агентов обработки знаний, компонентов пользовательских интерфейсов);
- обеспечения семантической совместимости многократно используемых компонентов интеллектуальных компьютерных систем и семантической совместимости самих этих систем, а также технологий их проектирования и поддержки последующих этапов их жизненного цикла.

Основная тема сборника научных трудов «Открытые семантические технологии проектирования интеллектуальных систем» — различные аспекты конвергенции и интеграции, обеспечивающие переход к интеллектуальным компьютерным системам нового поколения и соответствующей технологии комплексной поддержки их жизненного цикла:

- конвергенция и интеграция различных моделей представления и обработки информации в интеллектуальных компьютерных системах нового поколения:
  - конвергенция и интеграция различных видов знаний в базах знаний интеллектуальных компьютерных систем нового поколения;
  - конвергенция и интеграция различных моделей решения задач в интеллектуальных компьютерных системах нового поколения;
  - конвергенция и интеграция различных видов интерфейсов интеллектуальных компьютерных систем нового поколения;
- конвергенция и интеграция различных направлений Искусственного интеллекта в целях построения общей формальной теории интеллектуальных компьютерных систем нового поколения;
- конвергенция и интеграция технологий проектирования различных компонентов интеллектуальных компьютерных систем нового поколения в целях построения комплексной технологии проектирования интеллектуальных компьютерных систем нового поколения;
- конвергенция и интеграция технологий поддержки различных этапов жизненного цикла интеллектуальных компьютерных систем нового поколения в целях построения технологии комплексной поддержки всех этапов жизненного цикла интеллектуальных компьютерных систем нового поколения;
- конвергенция и интеграция различных видов человеческой деятельности в области Искусственного интеллекта (научно-исследовательской деятельности, развития технологического комплекса, прикладной инженерии, образовательной деятельности) для повышения уровня согласованности и координации этих видов деятельности, а также для

повышения уровня их комплексной автоматизации с помощью семантически совместимых интеллектуальных компьютерных систем нового поколения

- конвергенция и интеграция самых различных видов и областей человеческой деятельности, а также средств комплексной автоматизации этой деятельности с помощью интеллектуальных компьютерных систем нового поколения.

Основные **направления** сборника научных трудов «Открытые семантические технологии проектирования интеллектуальных систем»:

- Требования, предъявляемые к интеллектуальным компьютерным системам нового поколения;
- Принципы, лежащие в основе интеллектуальных компьютерных систем нового поколения;
- Требования, предъявляемые к смысловому представлению информации. Универсальный язык смыслового представления знаний;
- Онтологическая стратификация смыслового представления баз знаний в интеллектуальных компьютерных системах нового поколения;
- Агентно-ориентированная иерархическая модель гибридных решателей задач, основанных на смысловом представлении баз знаний;
- Онтологическая модель мультимодальных интерфейсов интеллектуальных компьютерных систем нового поколения;
- Модель понимания информации, поступающей извне, основанная на смысловом представлении баз знаний;
- Гибкость, стратифицированность, рефлексивность и обучаемость интеллектуальных компьютерных систем нового поколения;
- Унификация, стандартизация и семантическая совместимость интеллектуальных компьютерных систем нового поколения — согласованность понятий и терминов;
- Интероперабельность (способность к эффективному взаимодействию) интеллектуальных компьютерных систем нового поколения — способность к взаимопониманию, договороспособность, способность к координации своих действий при коллективном решении комплексных задач;
- Комплексная технология проектирования интеллектуальных компьютерных систем нового поколения;
- Платформы реализации интеллектуальных компьютерных систем нового поколения. Программные платформы и ассоциативные семантические компьютеры, ориентированные на реализацию интеллектуальных компьютерных систем нового поколения;
- Конвергенция логико-семантических моделей интеллектуальных компьютерных систем нового поколения и архитектур ассоциативных семантических компьютеров, обеспечивающих интерпретацию указанных моделей. Волновые языки микропрограммирования;
- Принципы, лежащие в основе Технологии комплексной поддержки жизненного цикла интеллектуальных компьютерных систем нового поколения;
- Глобальная Экосистема интеллектуальных компьютерных систем нового поколения, обеспечивающая комплексную конвергенцию и автоматизацию всевозможных видов человеческой деятельности.

Сборник содержит 44 статьи. Редакция сборника благодарит всех авторов, представивших свои статьи. Для публикации научными экспертами были отобраны лучшие из представленных работ, многие из них были переработаны в соответствии с замечаниями рецензентов.

Мы также благодарим экспертов за большой труд по рецензированию статей в тесном взаимодействии с авторами, который позволил повысить уровень изложения научных результатов, а также создал платформу для дальнейших научных дискуссий.

Надеемся, что, как и прежде, сборник будет выполнять свою основную функцию — способствовать активному сотрудничеству между бизнесом, наукой и образованием в области искусственного интеллекта.

Главный редактор
Голенков Владимир Васильевич

Principles of Decentralized Problem Solving Within the Ecosystem of Next-Generation Intelligent Computer Systems

Daniil Shunkevich
Belarusian State University of Informatics and Radioelectronics
Minsk, Belarus
Email: shunkevich@bsuir.by

***Abstract*—The paper considers the principles of decentralized problem solving within the ecosystem of next-generation intelligent computer systems, in particular, the architecture of such an ecosystem is considered from the point of view of organizing the process of problem solving, the roles of systems involved in the process of problem solving are separated. The principles of forming a collective of systems involved in problem solving, the stages of solving a particular problem by the resulting collective are specified.**

***Keywords*—Decentralized AI, multi-agent approach, OSTIS technology, OSTIS Ecosystem, problem solver, sc-agent, ostis-system**

I. Introduction

Currently, one of the key trends in the field of intelligent technologies is the so-called decentralized artificial intelligence [1], [2]. This trend is caused, on the one hand, by the development and widespread use of autonomous mobile devices (both self-sufficient and as part of more complex systems, as in the case of various kinds of sensors), which interact intensively when solving tasks together, which, in particular, led to the emergence and development of such a trend as the Internet of Things. On the other hand, decentralization is necessary when solving problems of complex automation of various processes, for example, production. In this case, the implementation of an automation system with a monolithic centralized architecture becomes impossible, in this regard, the task of integrating a variety of heterogeneous devices and subsystems into a single information space, within which various models and methods of artificial intelligence are subsequently applied. There is a large number of studies aimed at developing the principles of problem solving in distributed teams of interacting computer systems [1]–[3].

In turn, the complex automation of various human activities and the development of corresponding complex intelligent systems of automation leads to the need to integrate within such systems different types of knowledge and different problem solving models and to ensure the possibility of joint use of these models in solving complex problems [4], [5]. In addition, the problem of reducing the labor intensity of not only the development of such systems, but also their evolution and maintenance at all stages of the life cycle is relevant. The solution of the mentioned problem becomes possible when providing syntactic and semantic compatibility of models of representation of different types of knowledge and different problem solving models. In other words, to solve the problem of complex automation of human activity, it is necessary to provide convergence of different models of representation and information processing in intelligent systems.

At the same time, as it has already been noted, an important trend in the development of artificial intelligence technologies is the desire to decentralize the solution of various problems, in connection with which the creation of *next-generation intelligent computer systems* with a high level of ***interoperability*** is increasingly relevant. In this case, by interoperability [6] is meant not just ensuring compatibility of systems at the level of technical implementation (coordination of interaction protocols, program interfaces, etc.), but ensuring their semantic compatibility and ability to collectively solve complex problems. This implies a significant development and increase in the level of formalization of the theory of intelligent computer systems, rethinking of the existing technologies of their development and maintenance in the context of ensuring their convergence, and ultimately – the creation of a comprehensive technology of designing intelligent computer systems of a new generation, taking into account the need to develop not only individual intelligent computer systems, but also decentralized collectives of such systems, capable of jointly solving complex problems.

OSTIS Technology is proposed as a comprehensive design technology for *next-generation intelligent computer systems* [4]. Next-generation intelligent computer systems developed on the basis of this technology are called *ostis-systems*. OSTIS Technology is based on a universal method of semantic representation (coding) of information in the memory of intelligent systems, called *SC-code*. Texts of *SC-code* (sc-texts, sc-constructions) represent unified semantic networks with basic theoretical-multiple interpretation. The elements of such semantic networks are called sc-elements (sc-nodes and sc-connectors, which, in turn, can be sc-arc or sc-edges depending on their orientation).

Universality and unification of *SC-code* allows describing any types of knowledge and any methods of problem solving on its basis, which, in its turn, greatly simplifies their integration both within one system and within a collective of such systems.

The basis of the knowledge base developed according to the OSTIS Technology is a hierarchical system of semantic models of *subject areas* and *ontologies*, among which there is a universal Kernel of semantic models of knowledge bases and methodology of development of semantic models of knowledge bases, providing semantic compatibility of the developed knowledge bases. The basis of the ostis-system problem solver is a set of agents interacting exclusively through the specification of information processes they perform in the semantic memory (*sc-agents*).

All of the above principles together allow to ensure semantic compatibility and simplify integration of various components of computer systems as well as of such systems themselves. Within the framework of *OSTIS Technology* several universal variants of visualization of SC-code constructions have been proposed. This paper will use examples in *SCg-code*, a graphical variant of visualization of SC-code constructs, and *SCn-code*, a structured hypertext variant of visualization of SC-code constructs.

Within the framework of OSTIS Technology the models, methods and means of developing hybrid knowledge bases and problem solvers of next-generation intelligent computer systems are considered in detail, including the issues of providing convergence of different types of knowledge and different models of problem solving [4]. At the same time, the problem of organizing the process of problem solving by a decentralized collective of ostis-systems remains relevant.

Thus, the task of developing a theory of problem solving in <u>distributed</u> <u>collectives</u> of interoperable *next-generation intelligent computer systems* remains relevant.

This task solving is proposed to be carried out on the basis of the concept of ecosystem of interacting ostis-systems (***OSTIS Ecosystem***). In the paper [7] general approaches to problem solving within the framework of the OSTIS Ecosystem are considered, in this paper the architecture of the OSTIS Ecosystem and the principles of organizing problem solving within this ecosystem will be specified.

II. OSTIS Ecosystem Architecture

A. Concept and main tasks of the OSTIS Ecosystem

OSTIS Ecosystem — a sociotechnical ecosystem, which is a collective of interacting semantic computer systems and provides permanent support for the evolution and semantic compatibility of all its constituent systems throughout their life cycle [4], [8].

In order to talk about the principles of problem solving within the OSTIS Ecosystem, it is necessary to clarify the architecture of the OSTIS Ecosystem and the advantages of the concept of such an ecosystem in the context of decentralized problem solving.

The *OSTIS Ecosystem*, is a collective of interacting (via the Internet) [4]:

- *ostis-systems* themselves;
- users of the specified *ostis-systems* (both end users and developers);
- some computer systems that are not *ostis-systems*, but are considered by them as additional information resources or services.

OSTIS Ecosystem is a form of implementation, improvement and application of *OSTIS Technology* and, therefore, is a form of creation, development, self-organization of the market of semantically compatible computer systems and includes all resources necessary for this — informational, technological, personnel, organizational, infrastructural.

One of the important tasks of the OSTIS Ecosystem is to ensure that the compatibility of the computer systems included in the OSTIS Ecosystem is maintained at all times, both during their development and during their operation. The problem here is that during the operation of the systems included in the OSTIS Ecosystem, they may change and thus compatibility may be compromised.

The solution of the above task involves solving the following subtasks:

- operational implementation of all agreed changes to the standard of *ostis-systems* (including changes to the systems of used concepts and their corresponding terms) [9];
- permanent support for a high level of mutual understanding of all systems included in the OSTIS Ecosystem and all their users;
- corporate solution of various complex problems requiring coordination of activities of several (most often, a priori unknown) *ostis-systems*, as well as, possibly, some users.

Thus, the *OSTIS Ecosystem* is the basis for the transition from independent (autonomous, separate, integral) *ostis-systems* to collectives of independent *ostis-systems*, i.e. to distributed *ostis-systems*.

B. OSTIS Ecosystem Agent Typology

Let us consider the classification of ostis-systems in terms of their independence (interaction with other *ostis-systems* within the *OSTIS Ecosystem*):

ostis-system
⇒ *subdividing**:
{• *independent ostis-system*
:= [complete *ostis-system* that must independently solve a corresponding set of tasks and, in particular, interact with the external environment (both verbally — with users and other computer systems, and non-verbally).]
• *embedded ostis-system*
:= [intelligent computer subsystem developed according to *OSTIS Technology* and realizing part of the functionality of *ostis-system* of a higher hierarchy level]
:= [*ostis-system* integrated into *independent ostis-system*]
⇒ *subdividing**:
{• *atomic embedded ostis-system*
:= [*embedded ostis-system* that does not include any other *embedded ostis-system*]
• *non-atomic embedded ostis-system*
⊃ *ostis-system interface*
}
• *ostis-systems collective*
:= [group of communicating ostis-systems, which may include not only independent ostis-systems, but also collectives of ostis-systems]
:= [distributed ostis-system]
}

We emphasize that the ***independent ostis-systems*** that are part of the *OSTIS Ecosystem* have special requirements:

- They should have all the necessary knowledge and skills to exchange messages and to organize purposeful interactions with other *ostis-systems* within the *OSTIS Ecosystem*.
- In the conditions of constant change and evolution of *ostis-systems* included in *OSTIS Ecosystem*, each of them should self monitor the state of its compatibility (consistency) with all other *ostis-systems*, i.e. it should independently maintain this compatibility, coordinating with other *ostis-systems* all the changes that require coordination, occurring in itself and in other systems.
- Each system included in the *OSTIS Ecosystem* shall:
  - Learning intensely, actively and purposefully (both with the help of developmental teachers and independently).
  - Notify all other systems of proposed or finalized changes to the *ontologies* and, in particular, to the set of *concepts* used.
  - Accept proposals from other *ostis-systems* for changes to the *ontologies* (including the set of concepts used) in order to agree or approve these proposals.
  - Implement approved changes to the *ontologies* stored in its knowledge base.
  - Contribute to maintaining a high level of semantic compatibility not only with other *ostis-systems* within the *OSTIS Ecosystem*, but also with its *users* (i.e. educate them, inform them about changes in ontologies).

Thus, the OSTIS Ecosystem is essentially a distributed *multi-agent system* [10], which includes *agents of OSTIS Ecosystem*, which are the subjects of the activities performed within the *OSTIS Ecosystem*. Let's consider the classification of *agents of OSTIS Ecosystem* taking into account the classification of *ostis-systems* given above. [8].

agent of OSTIS Ecosystem
⇒ *subdividing**:
{• *individual ostis-system of OSTIS Ecosystem*
⇒ *subdividing**:
{• *independent ostis-system of OSTIS Ecosystem*
• *embedded ostis-system of OSTIS Ecosystem*
}
• *user of the OSTIS Ecosystem*
• *ostis-community*
⇒ *subdividing**:
{• *simple ostis-community*
• *hierarchical ostis-community*
}
}

C. The concept of ostis-community

Ostis-community is a stable fragment of *OSTIS Ecosystem*, providing complex automation of a certain part of collective human activity and permanent increase of its efficiency.

A *hierarchical ostis-community* is a *ostis-community*, at least one of whose members is some other *ostis-community*. *Ostis-community* generally includes not only *ostis-systems collective*, but also a certain collective of people (users and developers of the corresponding *ostis-systems*). Each *OSTIS Ecosystem agent* (both individual and collective) can become a member of any ostis-community of the OSTIS Ecosystem on his/her own initiative after proper registration of [4], [8].

OSTIS Ecosystem is the maximal *hierarchical ostis-community* that provides comprehensive automation of all human activities. It cannot be a part of any other *ostis-community*.

Note that the proposed *OSTIS Ecosystem* architecture and the separation of *ostis-communities* is not exclusively part of the approach to organizing decentralized problem solving of applications. The *OSTIS Ecosystem* is a tool for the realization and evolution of the *OSTIS Technology*, in particular a basis for the realization of a component-based approach in the development of *ostis-systems*. In this paper, the architecture of the *OSTIS Ecosystem* is considered and refined in the context of collective problem solving.

Let us clarify the architectural principles underlying the *OSTIS Ecosystem* [4], [8]:

- *OSTIS Ecosystem* is a network of *ostis-communities*;
- To each *ostis-community* mutually unambiguously corresponds the *corporate ostis-system* of this *ostis-community*;
- Each *ostis-community* may join any other *ostis-community* on its own initiative. Formally, this means that the *corporate ostis-system* of the first *ostis-community* is a member of another *ostis-community*;
- Each specialist who is a part of the *OSTIS Ecosystem* is mutually correspondent to his/her *personal ostis-assistant*, which is interpreted as a *corporate ostis-system* of a degenerated *ostis-community* consisting of one person;
- The *corporate ostis-system* of each *ostis-community* stores a specification of the *ostis-systems* belonging to this *ostis-community*, in its turn, each *ostis-system* stores some specification about the *corporate ostis-system* of each *ostis-community* to which this *ostis-system* belongs ("knows" its *corporate ostis-systems*).

Thus, the following levels of hierarchy can be distinguished in the *OSTIS Ecosystem*:

- individual *cybernetic systems* (individual *ostis-systems* and people who are the end users of the *ostis-systems*);
- hierarchical system of *ostis-communities*, the members of each of which can be *individual ostis-systems*, people, as well as other *ostis-communities*;
- *Maximum ostis-community* of the *OSTIS Ecosystem* that is not a member of any other *ostis-community* within the *OSTIS Ecosystem*.

Each person who is a member of *OSTIS Ecosystem* unambiguously corresponds to his/her personal assistant in the form of *personal ostis-assistant*. The collective consisting of a person and her corresponding *personal ostis-assistant* will be called a *minimal ostis-community* [4], [8]. Since formally non *minimal ostis-communities* do not include persons, but their corresponding *personal ostis-assistants*, all *ostis-communities* except *minimal ostis-communities* are *collectives of ostis-systems*.

ostis-community
$\Rightarrow$ *subdividing**:
{• *minimal ostis-community*
• *collective of ostis-systems*
}

A *corporate ostis-system* is a central *ostis-system* that coordinates, organizes, and supports the evolution of the activities of the members of the correspondent *ostis-community*. The *corporate ostis-system* is the representative of the correspondent *ostis-community* in other *ostis-communities* of which it is a member. Consequently, the main purpose of the *Corporate system of OSTIS Ecosystem* is to organize common interaction in the performance of various types and areas of human activities, which may be either fully automated, partially automated or not automated at all.

The *OSTIS Metasystem* is considered as the *Corporate system of OSTIS Ecosystem*, which is also implemented based on *OSTIS Technology* and whose knowledge base contains [4]:

- Formal theory of ostis-systems;
- Ostis-systems and OSTIS Technology Standard (*OSTIS Standard*);
- Core Library of reusable ostis-system components (*OSTIS Library*);
- Methods and tools for supporting the life cycle of ostis-systems and their components;
- Description of the structure of the *OSTIS Ecosystem*, including a description of the hierarchy of *ostis-communities* and their composition.

Figure 1 shows an example of a fragment of the *OSTIS Ecosystem* in education. As can be seen from the example, the same ostis-community can be part of several other ostis-communities. Each ostis-community corresponds to its *corporate ostis-system* (only a part of it is shown in the figure).

Within the framework of this work *corporate ostis-systems* is proposed to be considered as one of the important means of organizing the problem solving process in *OSTIS Ecosystem* [4], [8].

III. Features of problem solving within the ecosystem of next-generation intelligent computer systems

A. General principles of information processing within the framework of OSTIS Technology

The basic principles of information processing in memory ostis-systems are formulated in the previous works of the author [4], [11]:

- All problem solving processes are proposed to be divided into *logically atomic actions*, into those depending on what actions are performed before and after them and what more general actions they are part of;

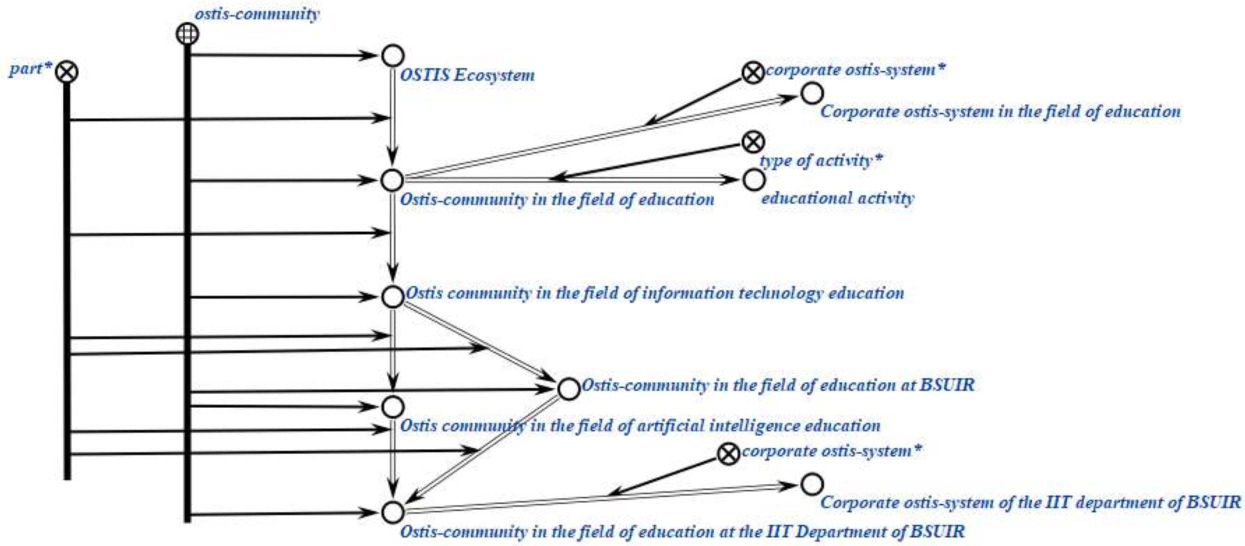


Figure 1. Example of ostis-community hierarchy

- Each *class of logically atomic actions* corresponds to a solver component (sc-agent) capable of performing actions of the specified class;
- Sc-agents react to various events in the ostis-system memory (sc-memory) and communicate with each other only by specifying the actions they perform in this memory. Direct message exchange between sc-agents is excluded;
- Each sc-agent has a corresponding specification, which is also stored in the same sc-memory as the constructs being processed;
- The ostis-system problem solver is treated as a hierarchical system of sc-agents. Two aspects of organization of such hierarchy are distinguished:
  - hierarchy of sc-agents in terms of the *l*evelmethod representation language (programming language). Lower-level sc-agents interpret methods corresponding to higher-level sc-agents;
  - hierarchy of sc-agents within one *method representation language*. A sc-agent may correspond not to a specific *method* describing the sc-agent's program of actions, but to a family of simpler sc-agents, and the level of such "nesting" is not limited in principle. Sc-agents decomposable into a collective of simpler sc-agents are called *non-atomic sc-agents*.
- Classes of functionally equivalent sc-agents, which have a common specification but are realized in general in different ways, are called *abstract sc-agents*.

The above principles allow to ensure hybridity, modifiability of ostis-systems problem solvers, as well as convenience of their design and evolution [4], [11].

The key difference between the distributed ostis-system and the internal system of sc-agents within an individual ostis-system is the absence of a common memory storing a common knowledge base for all sc-agents and acting as a medium for communication between sc-agents. In general, as a means of communication between the participants of a distributed collective of ostis-systems can be used [7]:

- Shared unallocated (monolithic) memory, as in the case of sc-agents over sc-memory;
- Shared distributed memory. In this case, from a logical point of view, agents may think that they are still working on a shared memory, where the entire available knowledge base is stored, but in reality the knowledge base will be distributed among several ostis-systems and the performed transformations will have to be synchronized among these ostis-systems;
- Specialized communication channels. Obviously, when solving a problem in a distributed collective of ostis-systems, there should be language and technical means allowing to transfer messages from one ostis-system to another.

All of the above means of communication can be combined depending on the class of the problem solved, the knowledge and skills required for its solution, and the currently available set of ostis-systems.

When solving a particular problem by a distributed collective of ostis-systems, in general the following "organizational" subproblems related to the organization of the communication process of the ostis-systems themselves must be solved before proceeding directly to problem solving within the subject domain:

- which ostis-system will provide the environment for interaction of other ostis-systems;
- what set of ostis-systems will be involved in solving this problem solving (knowledge and skills of which ostis-systems will be required);
- how the overall problem solving plan will be formed and refined, on the basis of which problem solving strategy it will be formed, how this strategy will be selected;
- where the overall problem solving plan will be stored and how it will be interpreted;
- where the result of problem solving will be formed and intermediate results of the solution will be stored;
- how the sub-tasks will be distributed among the participants in the solution process;
- how the intermediate results of individual subtasks will be synchronized with each other.

B. Special features of information processing within the OSTIS Ecosystem

The very idea of *OSTIS Ecosystem* and complex automation of human activity implies the need for self-organization of agents performing problem solving within the ecosystem. Currently, there is a large number of works devoted to the issues of self-organization of participants in the process of decentralized information processing [1]–[3]. In modern works it is common to use the classification of self-organization mechanisms proposed in the works [12], [13]. The work [1] even suggests the idea of expediency of creating libraries of standard algorithms for decentralized computing and group management of networks of autonomous objects, including consensus protocols, protocols for leader selection, protocols for contractual networks, auction protocols, protocols for common intentions, protocols for information exchange in the interests of maintaining situational awareness of participants in group behavior, and many others.

At the same time, the proposed architecture of the *OSTIS Ecosystem* has a number of important features in comparison with traditional multi-agent systems, within the framework of which the existing self-organization mechanisms are implemented.

- Agents in traditional self-organizing systems usually have rather limited functional capabilities, a small amount of knowledge about the environment and relatively low reliability. This is especially pronounced in the works devoted to the so-called "swarm intelligence", where each agent in the system is maximally simplified, and the number of agents in the system grows accordingly.

  In turn, each ostis-system within the *OSTIS Ecosystem* is a complex computer system with an extensive knowledge base and functionality that allows such a system to solve a variety of problems from the relevant subject domain. The implications of this distinction are as follows:
  - In traditional self-organizing systems, it is assumed that the vast majority of tasks can be solved only by a collective of agents. In *OSTIS Ecosystem* many problems can be solved by one particular ostis-system, therefore, it is often relevant not to form a collective of ostis-systems for solving a problem, but simply to search for an ostis-system capable of solving the problem;
  - In traditional self-organizing systems, it is usually assumed that agents have similar functional capabilities or at least there are groups of agents with similar functional capabilities. Special attention is paid to the problem of rational distribution of subtasks (load optimization) among such agents. Within *OSTIS Ecosystem* a number of functions can be duplicated in different ostis-systems, but it is assumed that each ostis-system has a sufficiently large set of unique functionalities available only to it. In this regard, the problem of efficient resource allocation is replaced by the problem of finding the most appropriate performer or forming a collective of performers suitable for problem solving;
  - In traditional self-organizing systems, much attention is paid to the issues of system reliability and preservation of its performance in case of failure of one or more agents. The issue of reliability of ostis-systems is undoubtedly important and relevant, however, taking into account the presence of unique functional capabilities of each ostis-system, the issue of automatic replacement of some ostis-systems by others during collective problem solving is not considered at the current stage of development of the idea of OSTIS Ecosystem.
- Traditional self-organizing systems are usually not considered as hierarchical structures, all agents are considered as autonomous units within the system, interacting with similar agents at the same level. The exception is the approaches to self-organization, which imply the allocation of special coordinating agents or agents-arbitrators, whose task is to control other agents. Within the *OSTIS Ecosystem* framework, it is assumed to explicitly distinguish a hierarchy of agents corresponding to the hierarchy of *ostis-communities*, besides, *OSTIS Ecosystem* agents are classified according to the role they play in the process of collective problem solving, in particular, *corporate ostis-systems* are distinguished.

The hierarchical nature of the agent system makes it easy to develop and modify such systems by analogy with the hierarchical structure of problem solvers in *individual ostis-systems* [4], [11].

- In traditional systems, often all agents of the system, or at least a significant part of them, may be involved in problem solving. Taking into account the complexity of ostis-systems included in the OSTIS Ecosystem, such a situation is unlikely in the OSTIS Ecosystem and most often in the near future several ostis-systems, most often belonging to one ostis-community, will be involved in problem solving.
- Traditional self-organizing systems are usually considered in isolation from the means of representation of information processed in such systems, i.e. neither the form of representation of processed information, nor the semantics of processed information are explicitly fixed. An important advantage of *OSTIS Ecosystem* and *OSTIS Technology* as a whole is the orientation on unified and universal models of information representation, realized in the form of *OSTIS Technology* and a family of top-level ontologies built on its basis. This approach allows us to say
  - about universality of the developed mechanisms of collective problem solving within *OSTIS Ecosystem*, i.e. the possibility to unlimitedly increase the possibilities of *OSTIS Ecosystem* to automate different kinds of human activities in various fields;
  - about the possibility to describe the agents of the OSTIS Ecosystem by the same means that are used to describe the information processed by the agents, with the required degree of detail. Thus, it becomes possible to analyze the specification of some agents (e.g., their functional capabilities, classes of solved tasks, etc.) by other agents, which opens new opportunities for self-organization in collective problem solving.
  - about the possibility of exchanging information constructions of arbitrary complexity between agents, there is no need to limit the possible semantics of such messages and, moreover, to fix their structure, as it is often done in traditional approaches. It should be emphasized that agents in the framework of the proposed approach do not exchange messages directly, but specify in a common knowledge base the actions they perform, so there are no fundamental restrictions on the content of such specification.

The analysis of the presented features allows us to draw the following conclusions:

- Direct application of existing approaches to self-organization in multi-agent systems for solving *OSTIS Ecosystem* problems is not possible due to its essential features, but many known approaches can be adapted for solving a number of specific problems, for example, when organizing the exchange of messages between ostis-systems at the physical level, searching for the most suitable executor for solving this or that task and so on;
- At the same time, *OSTIS Ecosystem* features allow not to consider at the level of collective problem solving a number of non-trivial problems related to reliability assurance and optimization of load distribution between ostis-systems, and open new opportunities to expand the range of possible spheres of human activity automation, to ensure interoperability of corresponding intelligent systems and their collectives and to reduce the labor intensity of their maintenance and evolution.

IV. Proposed approach to problem solving within the next-generation intelligent computer systems ecosystem

The key difference of the proposed approach to the organization of decentralized problem solving in *OSTIS Ecosystem* in contrast to traditional approaches to self-organization is that within *OSTIS Ecosystem* we initially in the process of ecosystem development form a hierarchy of ostis-systems and their specification so as to further simplify the process of organizing collective problem solving, in particular, the formation of a collective of performers, the transfer of messages between performers, etc., leaving the opportunity for continuous refinement and improvement. Thus, the agent system is initially built in such a way as to make self-organization more convenient by analogy with the way top-level ontologies are developed to ensure compatibility of ontologies instead of developing ontologies independently of each other and then solving the problem of ontology alignment, which most often turns out to be not trivial at all.

In the paper [14] a methodology for the design of multi-agent systems including five stages is proposed. Let us consider in more detail the application of this methodology in the context of problem solving within the *OSTIS Ecosystem*:

- **Step 1: Analyze the objectives for which the system is being developed**. The purpose of the work in this case is to ensure the potential possibility of solving any problems arising within the *OSTIS Ecosystem*, which requires the availability of universal and unified means of describing the goals and objectives. Within the *OSTIS Technology* framework, common unified means of specification of actions and tasks are proposed [4]. As far as semantic completeness of such tools is concerned (taking into account all possible classes of tasks that may arise), it is proposed to take as a basis the task ontology proposed in [15].
- **Step 2: Designing the overall structure of the multi-agent system**. Within the framework of the

considered approach, the structure of the multi-agent system is based on the architecture of the *OSTIS Ecosystem* discussed above and is constantly refined taking into account the addition of new agents to the *OSTIS Ecosystem*. In terms of classification of *agents of OSTIS Ecosystem* at the current stage it is proposed to single out only *corporate ostis-systems* into a separate class due to the fact that they play a special role in the process of organizing collective problem solving. The principles of agents' communication via *corporate ostis-systems* are discussed in more detail at Step 4.

- **Step 3: Designing the internal structure of the agent and the principles of its operation**. Since all *OSTIS Ecosystem* agents are ostis-systems (even users of the OSTIS Ecosystem interact with it through personal ostis-assistants, which are ostis-systems [4], [8]), additional specification of the principles of their structure is not required, as it is discussed in detail in the works devoted to the OSTIS Technology [4], [16]. To ensure the possibility of interaction between ostis-systems over the network, it is proposed to add an interface subsystem to each system, which is discussed in more detail in Step 5.
- **Step 4: Develop the principles of agent interaction**. As mentioned earlier, it is proposed to base the principles of agents' communication within *OSTIS Ecosystem* during collective problem solving on the principles of agents' communication in the memory of ostis-systems (sc-agents). In the work [7] an approach is proposed assuming that one of the ostis-systems included in the collective of ostis-systems will be used as a tool of communication for the participants of the collective of ostis-systems. If such collective is formed on a permanent basis (is a *ostis-community* or a part of it), it is proposed to use the *corporate ostis-system* of the specified *ostis-community* as such communicator system. If a collective of ostis-systems is formed temporarily for solving one or several complex problems, i.e. it is necessary to temporarily involve *ostis-systems* belonging to several *ostis-communities*, two variants of organizing communication of ostis-systems are possible:
  - One of the systems belonging to such a temporary collective of ostis-systems is selected as a communicator system. In this case, such an ostis-system becomes temporarily the *corporate ostis-system* of the temporary *ostis-community*. Accordingly, in this case it is required to install in the ostis-system an interface subsystem for ostis-systems and to load into its knowledge base the specifications of other ostis-systems participating in the problem solving process. Thus, the cost of preliminary preparation of a collective of ostis-systems for problem solving can be quite serious, and this approach may be ineffective for relatively simple problems solving.
  - The *corporate ostis-system* of the closest hierarchical *ostis-community* is chosen as the communicator system, such that all ostis-systems required for the solution belong either to this ostis-community or to more private ostis-communities (possibly on several hierarchical levels). In the example of the *ostis-communities* hierarchy fragment shown in Figure 2, assuming that the problem solving requires the participation of *ostis-systems* ***OS1***, ***OS2***, and ***OS3***, then the *corporate ostis-system* of *ostis-community* ***OC1*** will be selected as the communicator system.

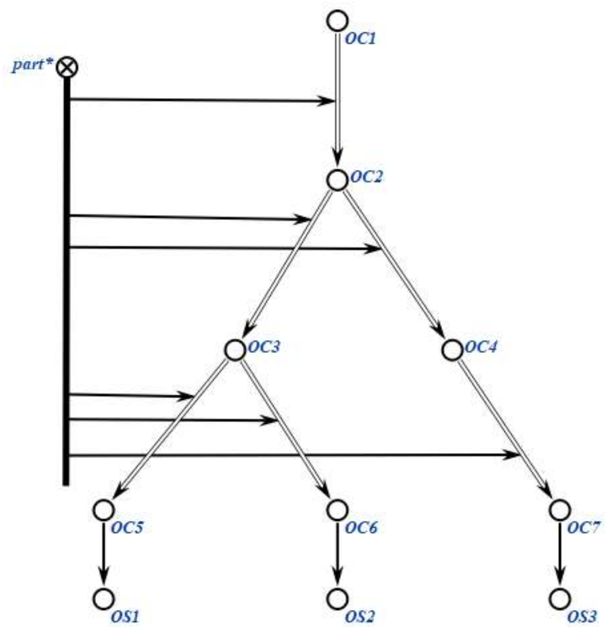


Figure 2. Example of communicator system selection

According to the above architecture of the *OSTIS Ecosystem* such an ostis-community will always exist, in the extreme case the role of such a corporate system will be played by the *OSTIS Metasystem*. The disadvantage of this communication option is that sending messages between the participants of the problem solving process may generally take more time due to the increased path between the corporate ostis-system and the ostis-systems which are performers.

It is important to note that in the presence of such a communicator system, agents at the logical level do not exchange messages directly, but communicate by specifying their actions in the shared memory of the communicator system; nevertheless, at the physical level, messages are forwarded between ostis-systems, generally physically located in different nodes of the network, arbitrarily distant from each other. This idea of separating the logical and physical layers of communication is realized within the concept of *overlay networks* [17]. An overlay network is a virtual network whose structure differs from the real communication network on the basis of which this overlay network functions. The idea of using *corporate ostis-systems* as a basis for agents' communication in such a network and a repository of agents' specifications and methods provided by them can be considered as a new stage of development of the idea of P2P agent platform developed by the FIPA consortium [17]. The main idea of such a platform is to provide all agents in the network with the possibility of semantic search for the services they need, as well as to search for agents that possess the required services. Another function of the platform is to support transparent communication between agents-customers and agents-owners of services [17]. In general, an agent network may have several such platforms, each responsible for a different segment of the network, similar to the way a *corporate ostis-system* is responsible for a corresponding *ostis-community*.

To implement the language of interaction between ostis-systems, it is proposed to use the ideas of wave programming [18], [19], as well as insertion programming [20], [21] as a basis. Later variants of the wave language theory development were called Spatial Grasp Technology [18], [19], within which Spatial Grasp Language is developed accordingly.

Implementing such interaction requires the development of at least two levels of languages:

- transport layer defining the principles of recording SC-code constructions in some format convenient for network transmission. As a variant of such language it is proposed to use SCs-code [4];
- semantic level defining the content of messages transmitted over the network. The SCP language, which is the basic programming language for ostis-systems, is proposed to be used as a basis for such a language [4].

It is important to note that within the framework of the proposed approach, the *corporate ostis-system* acts as a common information resource and notifies the participants of the problem solving process about the events, but does not control the problem solving process directly. Thus, it is not a question of rigid imperative management, but of more flexible declarative. This allows us to realize the advantages of the sc-agent interaction mechanism in a shared semantic memory [4], [11], such as modifiability of the agent system, convenience of its design and others.

- **Step 5: Develop the detailed architecture of the multi-agent system**. At this stage, it is supposed to clarify the principles of interaction between agents and the environment, taking into account the previously clarified agent structure and the principles of their interaction.

  Implementation of the proposed approach assumes that each ostis-system will include a communication interface subsystem that will receive messages from the external environment (from the *corporate ostis-system*), transform them into tasks for internal sc-agents of the *ostis-system*, and then transform the result of their work into a response message and send it to the corresponding recipient. An important feature of such a subsystem is that it behaves as a sc-agent when interacting with internal sc-agents and communicates with other sc-agents according to the same principles. This allows to ensure the independence of the development and evolution of such a subsystem from other components of the ostis-system and to exclude the necessity to take into account the fact of future interaction of the ostis-system with other ostis-systems at the stage of its design. In other words, an ostis-system solves a subtask within a distributed collective of ostis-systems just as if it were solving a problem explicitly formulated, for example, by a user. This approach greatly simplifies the design of ostis-systems collectives, eliminating explicit dependencies between them and the need to provide for the necessity of collective problem solving in advance.

  In turn, each *corporate ostis-system*, when interpreting a particular *method*, interacts with other ostis-systems as if they were internal sc-agents of this ostis-system. Thus, each *corporate ostis-system* includes an interface subsystem that converts events in the memory of the *corporate ostis-system* into messages to other *ostis-systems* and response messages from these *ostis-systems* into corresponding information constructs in the knowledge base of the *corporate ostis-system*. This approach allows to ensure the independence of *corporate ostis-systems* from other ostis-systems participating in the problem solving process and to exclude the necessity to provide in advance the necessity of collective problem solving not only when designing conventional ostis-systems, but also when designing *corporate ostis-systems*. An illustration of this approach will be given below.

From the point of view of the modern classification of self-organization methods in multi-agent systems [17], the proposed approach of agent interaction at the logical level can be considered as a kind of mechanism based on indirect interactions of organizational agents. Mechanisms of this kind assume the absence of direct interaction between autonomous entities composing the system, but their interaction through a common environment, which in the framework of the proposed approach is a common semantic memory (both within the *individual ostis-system* and within the collective *ostis-system*).

V. Means of specification of next-generation intelligent computer systems in the context of collective problem solving

An important role in the proposed approach to problem solving within the *OSTIS Ecosystem* is played by a rather detailed and unified specification of ostis-systems included in the *OSTIS Ecosystem*. Each ostis-system included in the *OSTIS Ecosystem* is subject to a number of requirements [4], [8], the fulfillment of which is necessary to ensure the principle possibility of collective problem solving, to increase the efficiency of the evolution of *OSTIS Ecosystem* and *OSTIS Technology*, to reduce the requirements to the developers of ostis-systems and the labor intensity of their development. The most important of these requirements is the requirement to ensure compatibility (both syntactic and semantic) of each *ostis-system* with others, and in particular with the *OSTIS Metasystem* containing the current version of the *OSTIS Standard*, and to continuously analyze and maintain such compatibility.

At the same time, in order to organize problem solving within *OSTIS Ecosystem* it is additionally necessary to have a detailed specification of functional capabilities of each ostis-system and to ensure the relevance of such specification in the process of evolution of this ostis-system. This specification is part of the knowledge base of *corporate ostis-systems ostis-communities*, to which the specified ostis-system belongs. If the *ostis-system* is not currently a part of any *ostis-community*, the *corporate ostis-system* is the *OSTIS Metasystem*.

The basis of the knowledge base of any ostis-system is a hierarchical system of *sc-models of subject domains* and their corresponding formal *ontologies* describing the properties of entities studied within the specified subject domains [4], [22]. Thus, the knowledge base of the *corporate ostis-system* contains sc-models of those subject domains, on the automation of various activities in which the corresponding *ostis-community* is oriented. In order to provide the possibility of automatic determination of the collective of ostis-systems necessary for solving a particular problem and clarifying the plan of solving this problem, it is necessary to develop for each subject domain the corresponding *ontology of subject domain problem classes and problem solving methods*. [4], [9].

The specified ontology includes a description:

- *classes of problems* solved in the corresponding subject domain;
- *methods* of problem solving corresponding to the specified *classes of problems*;
- *skills* of problem solving corresponding to the specified *classes of problems*, i.e. *methods*, supplemented by the description of *sc-agents* implementing the specified *methods* with the corresponding specification [4];
- *method representation languages* specific to the subject domain;
- *strategies of problem solving* specific to the subject domain, i.e. meta-methods of forming other *methods* of problem solving;
- and other entities, the description of which is necessary to organize problem solving processes within the subject domain. For example, if there are several methods of solving problems of the same class, it is reasonable to describe their comparison in order to be able to choose the method most suitable for the current situation.

As mentioned earlier, the ontology presented in [15] is proposed to be used as a basis for the content of the general ontology of all possible problem classes solved within the *OSTIS Ecosystem*. Thus, the set of problem classes described within a particular *ontology of subject domain problem classes and problem solving methods* will specify some subset of problem classes from such a top-level *problem ontology*. Examples of describing specific classes of problems and corresponding methods of their solution using the example of neural network methods of problem solving can be found in [23].

Thus, each ostis-system being a part of some ostis-community should be specified using the concepts of *ontology of subject domain problem classes and problem solving methods* presented within the corresponding *corporate ostis-system*. In its turn, within each individual ostis-system, this ontology can be further refined. Note that the same methods (and, accordingly, skills) can be duplicated between different ostis-systems, but the information about it is explicitly recorded, which allows us to take it into account when forming a problem solving plan and determining the composition of participants of the collective of ostis-systems taking part in the solution.

Accordingly, when adding ("registering") an ostis-system to an ostis-community, the following steps must be performed:

- Integrate the *ontology of subject domain problem classes and problem solving methods* into the corresponding ontology of the *corporate ostis-system*. Thus, the *corporate ostis-system* will receive information about new problem classes and methods of their solving, if there are any in the added ostis-system;
- Using the obtained integrated ontology, generate a specification of the added ostis-system in the knowledge base of *corporate ostis-system*;
- When the functionality of a *ostis-system* changes, it must notify the *corporate ostis-systems* of all *ostis-*

communities of which this *ostis-system* is a part, which in turn will lead to corresponding changes in the knowledge bases of these corporate *ostis-systems* and possibly to refinements of their corresponding *ontologies of subject domain problem classes and problem solving methods*. Note that this approach has an advantage over many traditional approaches to agent communication in multi-agent systems, where for successful subsequent operation of the system it is required to inform about the addition of a new agent all agents already in the system, since in the process of problem solving agents exchange messages directly and must "know" each other.

The considered specification of *ostis-systems* within the framework of *OSTIS Ecosystem* can be used not only for organizing problem solving, but also for other purposes, in particular, for implementing the idea of component design of ostis-systems [24]. Besides, the considered specification of *ostis-systems* is also necessary for the developers of *ostis-systems* in order to understand what capabilities are already presented within *OSTIS Ecosystem*, within which *ostis-communities*, with the developers of which ostis-systems it is necessary to coordinate these or those components of the developed system, and for solving a number of other design problem solving.

VI. A general plan for solving a specific problem within the next-generation intelligent computer systems ecosystem

According to the proposed approach to problem solving within the *OSTIS Ecosystem*, solving a particular problem generally involves the following steps:

- **Problem formulation**. In general, two options are possible at this step:
  - the initiator of problem solving is an ostis-system, which is a part of *OSTIS Ecosystem*. In this case, the problem formulation is placed in the knowledge base of the corresponding *corporate ostis-system*. To describe the problem formulation at the first stage, both the top-level *ontology of subject domain problem classes and problem solving methods* (included in the *OSTIS Standard* and, respectively, in the knowledge base of the *OSTIS Metasystem*) and more particular *ontology of subject domain problem classes and problem solving methods* corresponding to the *ostis-systems* belonging to the given *ostis-community* can be used.
  - the initiator of problem solving is an external cybernetic system, in particular a human user. In this case, it is assumed that communication with the *OSTIS Ecosystem* is carried out by a *personal ostis-assistant* corresponding to this cybernetic system. Thus, in this case, the task formulation is placed in the knowledge base of the *personal ostis-assistant* and then moved to the knowledge base of the *corporate ostis-system* of the ostis-community of which this *personal ostis-assistant* is a member. If a user is a member of several *ostis-communities* through his/her *personal ostis-assistant*, then the problem of optimal selection of the *ostis-community* within which it is most expedient to start solving a problem becomes relevant. At the same time, the proposed approach to decentralized problem solving in general does not depend on which *corporate ostis-system* the problem formulation initially enters, it affects only the total time of problem solving.

  Thus, as a result of this step, in any case, the problem formulation enters the knowledge base of some *corporate ostis-system* (in general, not necessarily that *corporate ostis-system*, which will act as a communication environment in the process of solving this problem).
- **Determining the set of ostis-systems to be involved in problem solving**. In general, it may be sufficient to involve only *ostis-systems* representing one *ostis-community*, or a set of *ostis-systems* belonging to different *ostis-communities*. The specific mechanism of this stage requires clarification, but the following principles are suggested as its basis:
  - the initiator of this stage is the *corporate ostis-system* whose knowledge base contains the corresponding problem formulation. For this purpose, the specified *corporate ostis-system* interacts with other *corporate ostis-systems*, if necessary involving *corporate ostis-systems* of a higher level. Development of a protocol for such interaction is an actual task;
  - the key role at this stage is played by the previously discussed *ostis-systems* specifications describing *classes of problems*, *methods* of their solving, etc;
  - in the process of performing this stage, the initial problem formulation may be refined taking into account particular *ontologies of subject domain problem classes and problem solving methods*.
- **Definition (selection) of *corporate ostis-system*, which will be the communication environment for solving the currently formulated problem solving task**. The principles of such a selection have been discussed above.
- **Formation of a problem solving plan**. At this stage of development of the theory of decentralized problem solving within the *OSTIS Ecosystem*, we will assume that the solution plan of a particular problem is formed, stored and refined entirely within the corresponding *corporate ostis-system*. In general, we can talk about the possibility of distributed

storage of the problem solving plan, but the interpretation of such a plan will require additional costs for interaction between ostis-systems and the development of additional mechanisms for the transfer of intermediate information and synchronization of actions between ostis-systems, the feasibility of which is difficult to assess at the moment in the absence of a sufficiently large number of applied examples of solving such complex problems.

The development of a general strategy for forming a plan for solving an arbitrary problem is currently an actual direction of development of the approaches considered in this paper. It is important to note that the problem solving plan in the general case will be constantly refined in the course of its implementation, which may require the refinement of the collective of ostis-systems involved in implementing this plan. This strategy is based on the idea of situational management [25] in conjunction with general methodological ideas related to the theory of behaviorism and the ideas of its application in computer science that are gaining popularity [26]–[28], TRIZ [29], as well as SMD-methodology proposed by the school of G. P. Shchedrovitsky [30].

- **Step-by-step interpretation of the problem solving plan**. The basic principles of interaction between *corporate ostis-system* and other *ostis-systems* participating in the problem solving process were considered earlier in the context of specifying the architecture of the multi-agent *ostis-systems* within the *OSTIS Ecosystem*. Implementing these principles requires specifying the architecture of the *ostis-systems* subsystems responsible for interaction between them in the process of problem solving and developing an appropriate interaction language.

  Figure 3 shows an example of the interpretation of a problem solving plan stored in the memory of a *corporate ostis-system* by a collective of *ostis-systems*. As can be seen from the example, *ostis-systems* interact with each other by means of corresponding communication subsystems, while the problem solving process itself does not take into account the fact of decentralized solution in any way.
- **Specification of the result of problem solving and its transfer to the initiator**. At this stage, the specification of the problem solving result (including the result itself, if it is an information construct) is formed, the composition of which generally depends on the problem class, and the obtained specification is transferred to the *ostis-system*, which was the initiator of the problem solving (in case of the end user *OSTIS Ecosystem*, the specification is transferred to his *personal ostis-assistant*).

It should be noted that the presented general plan of problem solving within the *OSTIS Ecosystem*, as can be seen from the explanations to its stages, is preliminary and in the future requires detailed specification of each of the stages.

VII. Conclusion

This paper considers the basic principles of decentralized problem solving within the next-generation intelligent computer systems ecosystem (*OSTIS Ecosystem*). In particular, the architecture of *OSTIS Ecosystem*, the typology of *agents of OSTIS Ecosystem*, and the features of problem solving within *OSTIS Ecosystem* are specified. The approach to problem solving within *OSTIS Ecosystem*, means of specification of *ostis-systems* in the context of collective problem solving is proposed. A general plan for solving a particular problem within the *OSTIS Ecosystem* is proposed.

At the same time, the solution of a number of promising tasks remains relevant:

- Development of a general strategy for solving problems of an arbitrary class, the principles of forming a general plan for solving a particular problem;
- Clarifying the language of the problem statements and objectives;
- Development of a general *Ontology of problem classes and problem solving methods*, clarification of the principles of development of private *ontologies of subject domain problem classes and problem solving methods* on its basis and principles of specification of internal *sc-agents* and whole *ostis-systems* using these ontologies;
- Development of a language of interaction between *ostis-systems* at the stage of collective formation of *ostis-systems* of a particular problem solving;
- Development of a language of interaction between *ostis-systems* at the stage of problem solving (interpretation and refinement of the problem solving plan);
- Refinement of the architecture of the *ostis-systems* subsystems responsible for the interaction between them in the process of problem solving.

Acknowledgment

The author would like to thank the scientific collectives of the departments of Intelligent Information Technologies of the Belarusian State University of Informatics and Radioelectronics and Brest State Technical University for their help and valuable comments.

References

[1] V. Gorodetskii, "Bazovye trendy decentralizovannogo iskusstvennogo intellekta [basic trends in decentralized artificial intelligence]," in *Dvadcataja Nacional'naja konferencija po iskusstvennomu intellektu s mezhdunarodnym uchastiem, KII-2022 : Trudy konferencii. V 2-h tomah, Moskva, 21–23 dekabrja 2022 goda. Tom 2. [Twentieth National Conference on Artificial Intelligence with International Participation, CAI-2022 :*

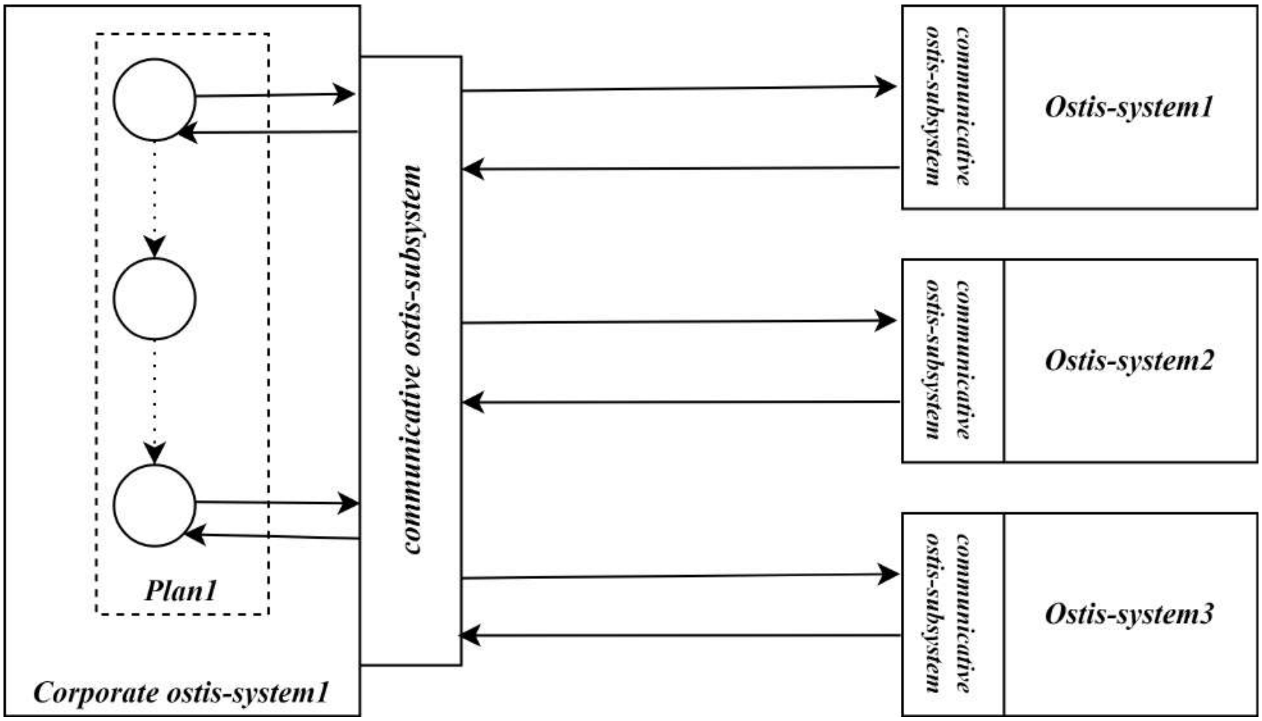


Figure 3. Example of method interpretation by the ostis-systems collective

Proceedings of the Conference. In 2 volumes, Moscow, December 21-23, 2022. Volume 2.]. National Research University "MPEI", 2022, pp. 275–291.

[2] L. Cao, "Decentralized ai: Edge intelligence and smart blockchain, metaverse, web3, and desci," *IEEE Intelligent Systems*, vol. 37, no. 3, p. 6–19, May 2022. [Online]. Available: http://dx.doi.org/10.1109/MIS.2022.3181504

[3] D. Ye, M. Zhang, and A. V. Vasilakos, "A survey of self-organization mechanisms in multiagent systems," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 47, no. 3, p. 441–461, Mar. 2017. [Online]. Available: http://dx.doi.org/10.1109/TSMC.2015.2504350

[4] V. Golenkov, Ed., *Tehnologija kompleksnoj podderzhki zhiznennogo cikla semanticheski sovmestimyh intellektual'nyh komp'juternyh sistem novogo pokolenija [Technology of complex life cycle support of semantically compatible intelligent computer systems of new generation]*. Bestprint, 2023.

[5] A. Kolesnikov, *Gibridnye intellektual'nye sistemy: Teoriya i tekhnologiya razrabotki [Hybrid intelligent systems: theory and technology of development]*, A. M. Yashin, Ed. SPb.: Izd-vo SPbGTU, 2001.

[6] S. Y. Mikhnevich and A. A. Tsezhar, "Evolution of the concept of interoperability of open information systems," *Digital Transformation*, vol. 29, no. 2, p. 60–66, Jun. 2023. [Online]. Available: http://dx.doi.org/10.35596/1729-7648-2023-29-2-60-66

[7] D. Shunkevich, "Principles of problem solving in distributed teams of intelligent computer systems of a new generation," *Open semantic technologies for intelligent systems*, no. 7, pp. 115–120, 2023.

[8] A. Zagorskiy, "Principles for implementing the ecosystem of next-generation intelligent computer systems," *Open semantic technologies for intelligent systems*, no. 6, pp. 347–356, 2022.

[9] V. Golenkov, N. Guliakina, and D. Shunkevich, *Otkrytaja tehnologija ontologicheskogo proektirovanija, proizvodstva i jekspluatacii semanticheski sovmestimyh gibridnyh intellektual'nyh komp'juternyh sistem [Open technology of ontological design, production and operation of semantically compatible hybrid intelligent computer systems]*, V. Golenkov, Ed. Minsk: Bestprint [Bestprint], 2021.

[10] V. Tarasov, *Ot mnogoagentnykh sistem k intellektual'nym organizatsiyam [From multi-agent systems to intelligent organizations]*. M.: Editorial URSS, 2002, (in Russian).

[11] D. Shunkevich, "Hybrid problem solvers of intelligent computer systems of a new generation," *Open semantic technologies for intelligent systems*, no. 6, pp. 119–144, 2022.

[12] G. Di Marzo Serugendo, M.-P. Gleizes, and A. Karageorgos, "Self-organization in multi-agent systems," *The Knowledge Engineering Review*, vol. 20, no. 2, p. 165–189, Jun. 2005. [Online]. Available: http://dx.doi.org/10.1017/S0269888905000494

[13] ——, "Selforganisation and emergence in multiagent systems: An overview," *Informatica*, vol. 30, no. 1, p. 45–54, 2006.

[14] Q.-N. N. Tran and G. Low, "MOBMAS: A methodology for ontology-based multi-agent systems development," *Information and Software Technology*, vol. 50, no. 7–8, p. 697–722, Jun. 2008. [Online]. Available: http://dx.doi.org/10.1016/j.infsof.2007.07.005

[15] A. Fayans and V. Kneller, "About the ontology of task types and methods of their solution," *Ontology of designing*, vol. 10, no. 3, pp. 273–295, Oct. 2020. [Online]. Available: https://doi.org/10.18287/2223-9537-2020-10-3-273-295

[16] V. Golenkov, N. Gulyakina, I. Davydenko, and D. Shunkevich, "Semanticheskie tekhnologii proektirovaniya intellektual'nyh sistem i semanticheskie associativnye komp'yutery [Semantic technologies of intelligent systems design and semantic associative computers]," *Otkrytye semanticheskie tehnologii proektirovanija intellektual'nyh sistem [Open semantic technologies for intelligent systems]*, pp. 42–50, 2019.

[17] V. Gorodetskii, "Samoorganizacija i mnogoagentnye sistemy. i. modeli mnogoagentnoj samoorganizacii [self-organization and multi-agent systems. i. models of multi-agent self-organization]," *Izvestiya RAN. Teoriya i sistemy upravleniya [Proceedings of the RAS. Theory and control systems]*, no. 2, pp. 92–120, 2012, (in Russian).

[18] P. S. Sapaty, *Spatial Grasp as a Model for Space-based Control*

and Management Systems. CRC Press, Apr. 2022. [Online]. Available: http://dx.doi.org/10.1201/9781003230090
[19] ——, *The Spatial Grasp Model: Applications and Investigations of Distributed Dynamic Worlds*. Emerald Publishing Limited, Jan. 2023. [Online]. Available: http://dx.doi.org/10.1108/9781804555743
[20] A. A. Letichevsky, O. A. Letychevskyi, and V. S. Peschanenko, *Insertion Modeling System*. Springer Berlin Heidelberg, 2012, p. 262–273.
[21] A. Letichevsky, Yu. Kapitonova, V. Volkov, V. Vyshemirsky, and A. Letichevsky (j.), "Insercionnoe programmirovanie [Insertion programming]," *Kibernetika i sistemnyj analiz [Cybernetics and system analysis]*, no. 1, pp. 19–32, 2003.
[22] I. Davydenko, "Semantic models, method and tools of knowledge bases coordinated development based on reusable components," in *Otkrytye semanticheskie tehnologii proektirovanija intellektual'nyh sistem [Open semantic technologies for intelligent systems]*, V. Golenkov, Ed., BSUIR. Minsk , BSUIR, 2018, pp. 99–118.
[23] M. Kovalev, "Convergence and integration of artificial neural networks with knowledge bases in next-generation intelligent computer systems," *Open semantic technologies for intelligent systems*, no. 6, pp. 173–186, 2022.
[24] M. Orlov, "Control tools for reusable components of intelligent computer systems of a new generation," *Open semantic technologies for intelligent systems*, no. 7, pp. 191–206, 2023.
[25] D. Pospelov, *Situacionnoe upravlenie. Teorija i praktika [Situational management. Theory and practice]*. M.: Nauka, 1986.
[26] L. Cao, "In-depth behavior understanding and use: The behavior informatics approach," *Information Sciences*, vol. 180, no. 17, pp. 3067–3085, Sep. 2010. [Online]. Available: https://doi.org/10.1016/j.ins.2010.03.025
[27] L. Cao, T. Joachims, C. Wang, E. Gaussier, J. Li, Y. Ou, D. Luo, R. Zafarani, H. Liu, G. Xu, Z. Wu, G. Pasi, Y. Zhang, X. Yang, H. Zha, E. Serra, and V. Subrahmanian, "Behavior informatics: A new perspective," *IEEE Intelligent Systems*, vol. 29, no. 4, pp. 62–80, Jul. 2014. [Online]. Available: https://doi.org/10.1109/mis.2014.60
[28] M. Pavel, H. B. Jimison, I. Korhonen, C. M. Gordon, and N. Saranummi, "Behavioral informatics and computational modeling in support of proactive health management and care," *IEEE Transactions on Biomedical Engineering*, vol. 62, no. 12, pp. 2763–2775, Dec. 2015. [Online]. Available: https://doi.org/10.1109/tbme.2015.2484286
[29] G. S. Al'tshuller, *Najti ideju: Vvedenie v TRIZ — teoriju reshenija izobretatel'skih zadach, 3-e izd. [Find an idea: An introduction to TRIZ - the theory of inventive problem solving, 3rd ed.]*. M.: Al'pina Pablisher, 2010.
[30] G. P. Shhedrovickij, *Shema mysledejatel'nosti – sistemno-strukturnoe stroenie, smysl i soderzhanie [Scheme of mental activity – system-structural structure, meaning and content]*. M.: Shk. kul't. pol., 1995.

ПРИНЦИПЫ ДЕЦЕНТРАЛИЗОВАННОГО РЕШЕНИЯ ЗАДАЧ В РАМКАХ ЭКОСИСТЕМЫ ИНТЕЛЛЕКТУАЛЬНЫХ КОМПЬЮТЕРНЫХ СИСТЕМ НОВОГО ПОКОЛЕНИЯ

Шункевич Д.В.

В работе рассмотрены принципы децентрализованного решения задач в рамках экосистемы интеллектуальных компьютерных систем нового поколения, в частности рассмотрена архитектура такой экосистемы с точки зрения организации процесса решения задач, выделены роли систем, участвующих в процессе решения задач. Уточнены принципы формирования коллектива систем, участвующих в решении задач, этапы решения конкретной задачи полученным коллективом.

Received 13.03.2024

Towards the Theory of Semantic Space

Valerian Ivashenko
Department of Intelligent Information Technologies
Belarusion State University of Informatics and Radioelectronics
Minsk, Republic of Belarus
ivashenko@bsuir.by

***Abstract*—The paper considers models for investigating the structure, topology and metric features of a semantic space using unified knowledge representation.**

The classes of finite structures corresponding to ontological structures and sets of classical and non-classical kinds are considered, and the enumerability properties of these classes are investigated.

The notion of operational-information space as a model for investigating the interrelation of operational semantics of ontological structures of large and small step is proposed.

Quantitative features and invariants of ontological structures oriented to the solution of knowledge management problems are considered.

***Keywords*—Semantic Space, Neg-entropy, Operational-information space, Enumerable sets, Natural numbers, Ackermann coding, Generalized formal language, Enumerable self-founded Hereditarily finite sets, Countable non-identically-equal Hereditarily finite sets, Multigraph, Hypergraph, Metagraph, Orgraph, Unoriented graph, Quasi-metric, Orgraph invariant, Homomorphism, Isomorphism, Homeomorhpism, Oriented sets, Graph wave-front, Dynamic graph system, Receptor, Effector, Transmitter, Resonator, Graph dimension, Fully-connected orgraph period, Rado graph, Universal model, Stable structure, Operational semantics, Denotational Semantics, Infinite structures, Generalized Kleene closure**

I. Introduction

There are different approaches to the study of topological, metrical and other properties of signs in texts leading to the consideration of corresponding semantic (or meaning (sense)) spaces [56].

Space is convenient because it is connected with some ordinal or metric scale which allows to reduce the cost of solving such cognitive tasks as searching (synthesis) or checking (analysis) the presence of an element (including for the purpose of eliminating redundancy) in a set organized as a space.

Knowledge integration based on unification is necessary both to eliminate redundancy and to compute semantic metrics. For this purpose, the developed model of unified knowledge representation [1], [5] can be adopted.

II. Approaches to the construction of a meaning space

The history of the development of the concept of "meaning space" and the corresponding models are described in the works [2], [11], [32], [56].

As stated in [56], the main approaches to the construction and research of the organization of meaning space include:

- exterior studying the physical nature [30], [33], [48] of processes including thinking processes [29],
- (quantitative) interior using quantitative and soft models, including probabilistic description of processes [11], [34], [35], [42], based on the practice of using words of language [20], [53],
- (qualitative) interior investigating the structure of represented knowledge and its dynamics [12], using formal semiotic models [51].

In some cases, it is possible to combine these elements of these approaches.

The following models and methods are used to construct and investigate the semantic space:

- mathematical models of spaces [37]–[41], [43], [44],
- formal and generalized formal languages [45], [56],
- methods of probability theory [11], [36], [54],
- methods of formal concepts analysis [58], [59],
- other models [3], [4], [45], [46], [49].

Further in the paper we consider the main classes of structures, their attributes and corresponding types of subspaces of the semantic space using unified knowledge representation [5], [12].

III. Unified representation and classification of fully representable finite knowledge structures

At the level of syntax, using syntactic links, it is possible to represent only finite knowledge structures in a unified (explicit) way.

Let us consider the principles of unified representation of knowledge [5], [12] with a structure that is one of finite structures of different kinds. Let us compare a certain class of structures to each kind of finite knowledge structures.

Note that finite structures can be divided into two main types: oriented finite structures and unoriented finite structures [21].

The simplest unoriented finite structures are hereditarily finite sets [63]. The class of hereditarily finite sets can be expressed as follows:

$$\emptyset^{(+^{*}_{1})} = H_{\aleph_0}$$

where

$$\binom{0}{A} \overset{def}{=} 2^{\emptyset} = \{\emptyset\}$$

$$\binom{1}{A} \overset{def}{=} \left(\bigcup_{x\in A} 2^{\{x\}}\right) / \binom{0}{A}$$

$$\binom{\iota+1}{A} \overset{def}{=} \left(\binom{\iota}{A} \breve{\cup} \binom{1}{A}\right) / \binom{\iota}{A}$$

$$A \breve{\cup} B \overset{def}{=} \bigcup_{\langle P,Q\rangle \in A\times B} \{P \cup Q\}$$

$$2^{\left(\emptyset+\sum_{x\in A}\{x\}\right)} \overset{def}{=} \bigcup_{\iota\in\mathbb{N}\cup\{0\}} \binom{\iota}{A}$$

$$A^{\left(+_k^1\right)} \overset{def}{=} \tau_k \left(\rho_k \left(\langle A, A\rangle\right) \cup \sigma_k \left(\langle A, A\rangle\right)\right)$$

$$A^{\left(+_k^{\iota+1}\right)} \overset{def}{=} \tau_k \left(\rho_k \left(\left\langle A, A^{\left(+_k^{\iota}\right)}\right\rangle\right) \cup \sigma_k \left(\left\langle A, A^{\left(+_k^{\iota}\right)}\right\rangle\right)\right)$$

$$A^{\left(+_k^*\right)} \overset{def}{=} \bigcup_{\iota\in\mathbb{N}/\{0\}} A^{\left(+_k^{\iota}\right)}$$

$$\tau_1 \left(A\right) \overset{def}{=} A$$

$$\rho_1 \left(\langle A, B\rangle\right) \overset{def}{=} \emptyset$$

$$\sigma_1 \left(\langle A, B\rangle\right) \overset{def}{=} 2^{\emptyset+\sum_{x\in(A\cup B)}\{x\}}$$

According to Ackermann coding [62], all hereditarily finite sets can be a mutually uniquely matched to natural numbers and thus enumerated [27]:

$$f\left(S\right) = 0 + \sum_{x\in S} 2^{f(x)}$$

A generalization of the class of hereditarily finite sets is the class of generalized hereditarily finite sets.

$$A^{\left(+_1^*\right)}$$

Generalized hereditarily finite sets can be embedded in (classical non-generalized) hereditarily finite sets:

$$\emptyset \sim 2^{\emptyset}$$

$$a_k \sim 2^{\{\emptyset\}_k}$$

$$g\left(\emptyset\right) = \{\emptyset\}$$

$$g\left(a_k\right) = \left\{\{\emptyset\}_k, \emptyset\right\}$$

$$g\left(X\right) = \left\{g\left(x\right) | x \in X\right\}$$

or alternatively:

$$\emptyset \sim d\left(1\right) = \{\emptyset\}$$

$$a_k \sim d\left(2 * k + 1\right)$$

$$d\left(k\right) = \bigcup_{i=1}^{\lfloor \log_2 k\rfloor} \left\{d\left(\left\lfloor \frac{k}{2^i}\right\rfloor mod 2\right)\right\}$$

$$d\left(0\right) = \emptyset$$

$$d\left(1\right) = \{\emptyset\}$$

$$d\left(2\right) = \{\{\emptyset\}\}$$

$$d\left(3\right) = \{\{\emptyset\}, \emptyset\}$$

$$d\left(4\right) = \{\{\{\emptyset\}\}\}$$

$$d\left(5\right) = \{\{\{\emptyset\}\}, \emptyset\}$$

$$d\left(6\right) = \{\{\{\emptyset\}\}, \{\emptyset\}\}$$

$$d\left(7\right) = \{\{\{\emptyset\}\}, \{\emptyset\}, \emptyset\}$$

$$d\left(8\right) = \{\{\{\emptyset\}, \emptyset\}\}$$

$$d\left(9\right) = \{\{\{\emptyset\}, \emptyset\}, \emptyset\}$$

$$d\left(10\right) = \{\{\{\emptyset\}, \emptyset\}, \{\emptyset\}\}$$

$$d\left(11\right) = \{\{\{\emptyset\}, \emptyset\}, \{\emptyset\}, \emptyset\}$$

$$d\left(12\right) = \{\{\{\emptyset\}, \emptyset\}, \{\{\emptyset\}\}\}$$

$$d\left(13\right) = \{\{\{\emptyset\}, \emptyset\}, \{\{\emptyset\}\}, \emptyset\}$$

$$d\left(14\right) = \{\{\{\emptyset\}, \emptyset\}, \{\{\emptyset\}\}, \{\emptyset\}\}$$

$$d\left(15\right) = \{\{\{\emptyset\}, \emptyset\}, \{\{\emptyset\}\}, \{\emptyset\}, \emptyset\}$$

$$d\left(16\right) = \{\{\{\{\emptyset\}\}\}\}$$

$$\ldots$$

$$g\left(\emptyset\right) = \{\emptyset\}$$

$$g\left(a_k\right) = d\left(2 * k + 1\right)$$

$$g\left(X\right) = \left\{g\left(x\right) | x \in X\right\}$$

In this way we obtain an ordering of generalized hereditarily finite sets (as example) in accordance with the Ackermann numbering and embedding in hereditarily finite (unoriented) sets.

As for oriented structures (oriented, "ordered" sets), if we take the von Neumann-Bernays-Gödel axiomatics [63] as a basis then with some "traditional" approach (representation of oriented pairs according to K. Kuratowski) [24] an empty string [10], [13], [14], an empty an oriented set [22] cannot be represented as unfounded sets in a theory with the von Neumann-Bernays-Gödel axiomatics [26], [63].

Accepted:

$$x = \langle x\rangle$$

in this case, the oriented pair of K. Kuratovsky:

$$\langle x, y\rangle = \{\{x\}, \{x, y\}\}$$

also

$$\langle x_1, x_2, x_3\rangle = \langle\langle x_1, x_2\rangle, x_3\rangle$$

$$\langle x_1, x_2, x_3, x_4\rangle = \langle\langle x_1, x_2, x_3\rangle, x_4\rangle$$

$$\langle x_1, x_2, x_3, x_4, x_5\rangle = \langle\langle x_1, x_2, x_3, x_4\rangle, x_5\rangle$$

$$\langle x_1, x_2, x_3, x_4, x_5, x_6\rangle = \langle\langle x_1, x_2, x_3, x_4, x_5\rangle, x_6\rangle$$

$$\langle x_1, x_2, ..., x_i, ..., x_{n-1}, x_n\rangle = \langle\langle x_1, x_2, ..., x_i, ..., x_{n-1}\rangle, x_n\rangle$$

$$A^n = \{\langle x_1, x_2, ..., x_i, ..., x_{n-1}, x_n\rangle | x_i \in A\}$$

The consequence of this is that strings are conditionally dimensional, that is, the length of a string is not its function, and therefore cannot be calculated uniquely from a string; an empty string cannot be represented by a set in the von Neumann-Bernays-Gödel theory:

$$\langle x, x\rangle = \{\{x\}\} = \langle\{x\}\rangle$$

$$\langle x, x\rangle = \langle\{x\}\rangle$$

$$2 = length\left(\langle x, x\rangle\right) \neq length\left(x\right) = 1$$

$$n = length\left(\langle x_1, x_2, ..., x_i, ..., x_{n-1}, x_n\rangle\right) \neq$$

$$length\left(\langle x_1, x_2, ..., x_i, ..., x_{n-1}, x_n\rangle\right) = 2$$

During understanding the string length function, if we move from a function (as in the formulas above) to a higher-order function with respect to the set of elements of an oriented set this does not solve the problem:

$$2 = length\left(\{\langle x, x\rangle, \{x\}\}\right)\left(\langle x, x\rangle\right) \neq$$

$$length\left(\{\langle x, x\rangle, \{x\}\}\right)\left(\langle\{x\}\rangle\right) = 1$$

Another consequence of this is that the Cartesian power can exhibit the following non-obvious and non-intuitive properties:

$$\exists A\left(A = A^1 \supset A^2 \supset A^3 \supset ... \supset A^i \supset ...\right)$$

Inability to represent the empty string ε when representing strings as oriented sets.

Let be:

$$E = \{\varepsilon\}$$

Required:

$$E^n = E$$

We have:

$$E^2 = \{\langle\varepsilon, \varepsilon\rangle\} = \{\{\{\varepsilon\}\}\}$$

$$\varepsilon = \{\{\varepsilon\}\}$$

$$E^3 = \{\langle\varepsilon, \varepsilon, \varepsilon\rangle\} = \{\langle\langle\varepsilon, \varepsilon\rangle, \varepsilon\rangle\} = \{\{\{\langle\varepsilon, \varepsilon\rangle\}, \{\langle\varepsilon, \varepsilon\rangle, \varepsilon\}\}\}$$

$$E^3 = \{\{\{\{\{\varepsilon\}\}\}, \{\{\{\varepsilon\}\}, \varepsilon\}\}\}$$

$$\varepsilon = \{\{\{\{\varepsilon\}\}\}, \{\{\{\varepsilon\}\}, \varepsilon\}\}$$

The latter violates the axiom of regularity (foundation), otherwise:

$$E^2 \neq E$$

$$E^3 \neq E$$

The use of non-founded sets is evidence of a transition to non-classical mathematical models.

There are approaches to representing strings in von Neumann-Bernays-Gödel set theory by equivalence classes of groupoids (which is complex) over oriented sets or functions (requires the construction of a set of ordinal numbers). In the first case, the representation grows exponentially, and in the second case, it is necessary to use oriented pairs [23], [25] (the number of characters for a string of length n grows no faster than $1+14*n+p\left(n\right)$, where $p\left(n\right) = 1 + n*\left(n+3\right)/2$ – number of characters to represent ordinal numbers). These approaches do not require a transition to non-classical mathematical models. However, a string of one element is not this element.

Let's consider another approach to representing strings and oriented sets, which does not require, overcomes the identified difficulties within the framework of classical mathematical models and uses pairs not according to K. Kuratowski, which cannot counter-intuitively have cardinality (length) equal to one.

Let us define the concept of disposing of a set.

$$1^S \overset{def}{=} S$$

$$\left(\iota+1\right)^S \overset{def}{=} \left\{\iota^T \middle| T \subseteq S\right\}$$

Example.

$$2^{\{x\}} = \left\{1^{\{x\}}, 1^{\emptyset}\right\} = \{\{\chi\}, \emptyset\}$$

$$3^{\{x\}} = \left\{2^{\{x\}}, 2^{\emptyset}\right\} = \left\{\{\{\chi\}, \emptyset\}, \left\{1^{\emptyset}\right\}\right\} = \{\{\{\chi\}, \emptyset\}, \{\emptyset\}\}$$

$$4^{\{x\}} = \left\{3^{\{x\}}, 3^{\emptyset}\right\} = \{\{\{\{\chi\}, \emptyset\}, \{\emptyset\}\}, \{\{\emptyset\}\}\}$$

$$5^{\{x\}} = \left\{4^{\{x\}}, 4^{\emptyset}\right\} = \{\{\{\{\{\chi\}, \emptyset\}, \{\emptyset\}\}, \{\{\emptyset\}\}\}, \{\{\{\emptyset\}\}\}\}$$

$$4^{\emptyset} = \left\{3^{\emptyset}\right\} = \left\{\left\{2^{\emptyset}\right\}\right\} = \left\{\left\{\left\{1^{\emptyset}\right\}\right\}\right\} = \{\{\{\emptyset\}\}\}$$

Also, let us define the concept of an individual set.

$$\{x\}_1 \overset{def}{=} \{x\}$$

$$\{x\}_{\iota+1} \overset{def}{=} \{\{x\}_\iota\}$$

note that:

$$\left(\iota+1\right)^{\emptyset} = \{\emptyset\}_\iota$$

Finite oriented set:

$$\bigcup_{i=1}^{k}\left\{\left(k-i+1\right)^{\{a_i\}}\right\}_i$$

The number of characters to represent it is no more than $1 + n*\left(5*n+1\right)/2 + q\left(n\right)$ where $q\left(n\right) = 2*n+1$ is the number of characters per representation of individual sets of the empty set.

Examples:

$$\langle\rangle = \emptyset$$

$$\langle\chi\rangle = \{\{\chi\}\}$$

A pair like a Wiener pair [23]:

$$\langle \chi, \gamma \rangle = \{\{\{\chi\}, \emptyset\}, \{\{\gamma\}\}\}$$

Other examples:

$$\langle \chi, \gamma, \zeta \rangle = \{\{\{\{\chi\}, \emptyset\}, \{\emptyset\}\}, \{\{\{\gamma\}, \emptyset\}\}, \{\{\{\zeta\}\}\}\}$$

Also let it be true:

$$A^n = \{\langle x_1, x_2, ..., x_i, ..., x_{n-1}, x_n \rangle | x_i \in A\}$$

From the above this follows:

$$\exists A \left(A^1 \neq A\right)$$

Based on the introduced concepts, we can give a definition of the Kleene closure:

$$A^* \overset{def}{=} A^{*_A}$$

where

$$A^{*_\Sigma} \overset{def}{=} A \cup \left(\bigcup_{\iota \in \mathbb{N} \cup \{0\}} A^\iota \right)^{\oplus \Sigma}$$

$$\Lambda^{\oplus \Sigma} = \bigcup_{\chi \in \Lambda} \left\{ \bigoplus_{\iota=1}^{|\chi|} \kappa\left(\langle \chi_\iota, \Sigma \rangle\right) \right\}$$

$$\kappa\left(\langle \gamma, \Sigma \rangle\right) \in \{\gamma | \tau\left(\langle \gamma, \Sigma \rangle\right)\} \cup \{\langle \gamma \rangle | \neg \tau\left(\langle \gamma, \Sigma \rangle\right)\}$$

$$\tau\left(\langle \gamma, \Sigma \rangle\right) = \Big(\left(|\gamma| \in \mathbb{N} \cup \{0\}\right) \wedge \Big(\exists B \left(\gamma \in B^{|\gamma|} / \Sigma \right) \Big) \Big) \wedge$$

$$\left(\bigcup_{\iota=1}^{\gamma} \gamma_\iota \subseteq \Sigma \right)$$

Kleene closure properties:

$$|\emptyset^*| = |\{\langle\rangle\}| = 1$$

$$\left|\{\emptyset\}^{*_\emptyset}\right| = \left|\{\langle\rangle\}^{*\emptyset}\right| = 1$$

if $0 < |A| \leq |\mathbb{N}|$ then $|A^*| = |\mathbb{N}|$.

Extensiveness property of the Kleene closure:

$$A \subseteq A^*$$

Monotonic property of the Kleene closure:

$$A^* \subseteq (A \cup \Delta)^{*_A}$$

Idempotency property of the Kleene closure:

$$(A^*)^{*_A} = A^*$$

The formal language through the introduced Kleene closure is defined:

$$\Lambda \subseteq A^* / \left(A / A^1\right)$$

Generalized strings differ from strings in that their elements can be not only elements of the alphabet but also strings and other generalized strings.

Let us define a generalized Kleene closure (the set of all generalized strings with respect to the alphabet A, extending the alphabet A).

$$A^{(*^*)} \overset{def}{=} \bigcup_{\iota \in \mathbb{N}/\{0\}} A^{(*^\iota)}$$

where

$$A^{(*^1)} \overset{def}{=} A^*$$

$$A^{(*^{\iota+1})} \overset{def}{=} \left(A^{(*^\iota)} \cup A\right)^*$$

Properties of the generalized Kleene closure: if $0 \leq |A| \leq |\mathbb{N}|$, then $\left|A^{(*^*)}\right| = |\mathbb{N}|$.

Extensibility property of the generalized Kleene closure:

$$A \subseteq A^* \subseteq A^{(*^*)}$$

Monotonicity property of the generalized Kleene closure:

$$A^{(*^*)} \subseteq (A \cup \Delta)^{(*^*)}$$

Idempotency property of the generalized Kleene closure:

$$\left(A^{(*^*)}\right)^{(*^*)} = A^{(*^*)}$$

A generalized formal language is defined:

$$\Lambda \subseteq A^{(*^*)} / \left(A / A^1\right)$$

The following is true:

$$\emptyset^{(*^*)} \subseteq \emptyset^{(+^*_1)}$$

Generalized formal languages are included in generalized hereditarily finite (unoriented) sets:

$$A^{(*^*)} \subseteq A^{(+^*_1)}$$

By embedding generalized strings (generalized hereditarily finite sets) into hereditarily finite sets, they can be ordered according to Ackermann coding.

Besides, all generalized strings can be mutually unambiguously mapped to strings of a formal language whose alphabet (extended alphabet) differs from the alphabet of generalized strings by two additional symbols: the start symbol and the end symbol of the string. Accordingly, given a linear ordering of the symbols of the extended alphabet, all generalized strings can be lexicographically ordered as corresponding strings of a formal language on the extended alphabet.

It is possible to consider unoriented sets that are not oriented sets such sets we will call ultra-unoriented sets.

Together with the corresponding generalized oriented sets, generalized ultra-unoriented hereditarily finite sets can be constructed as the class $A^{(+^*_2)}$ according to the expressions:

$$\tau_2(B) \overset{def}{=} B$$

$$\rho_2\left(\langle A, B \rangle\right) = (A \cup B)^*$$

$$\sigma_2(\langle A,B\rangle)=\pi_2\left(2^{\emptyset+\sum_{x\in A\cup B}\{x\}}\right)$$

$$\pi_2(B)=\{h(x)\,|x\in B\}$$

$$h(\emptyset)=\emptyset$$

$$\left\{h(a)\middle|(a\in A)\bigwedge(h(a)=a)\right\}=A$$

$$h(X)=\emptyset\cup\bigcup_{x\in X}\left\{2^{\{h(x)\}}\right\}$$

where h defines a mutually unambiguous mapping of generalized unoriented hereditarily finite sets into generalized ultra-unoriented hereditarily finite sets.

In this case:

$$A^{(+^*_2)}\subseteq A^{(+^*_1)}$$

Together with generalized unoriented finite sets, generalized ultra-unoriented finite sets can be mapped mutually unambiguously to inherited finite sets.

The classical mathematical model works with justified (grounded) sets, but some non-classical models correspond to other structures. While the structure of hereditarily finite sets is acyclic and involves various hierarchies, other (e.g., cholarchical [65] structures consist of elements where each element is a part and is a whole (composed of parts). Some such structures can be visualized as periodic or cyclic structures. Another example of structures that may not fit within the framework of classical mathematical models are non-trivially automorphic structures, since such models adhere to the abstraction of identity,. However, despite this, some of these “non-classical” structures can (under certain conditions) be represented by classical mathematical models. First of all, among such structures we will be interested in enumerable structures, i.e., such structures that can be enumerated.

Let us consider the representation of $A^{(+^*_3)}$ in generalized hereditarily finite sets of the class of countably non-identically-equal generalized hereditarily finite sets on the alphabet $\dot{A}$, which can be given according to the expressions:

$$\tau_3(B)\overset{def}{=}B$$

$$\rho_3(\langle A,B\rangle)=C\times(A\cup B)^*$$

$$\sigma_3(\langle A,B\rangle)=D\times\pi_3\left(2^{\emptyset+\sum_{x\in A\cup B}\{x\}}\right)$$

$$\pi_3(B)=B$$

$$a_{2k+1}=\begin{cases}\dot{a}_k\left|k\leq\left|\dot{A}\right|\right.\\ d(2*k+1)\left|k>\left|\dot{A}\right|\right.\end{cases}$$

$$a_{2k}=d(2*k)$$

$$C\cup D=E$$

$$E=\left\{a_{2k}\middle|(a_{2k}\in\dot{A})\bigwedge(k\in\mathbb{N})\right\}$$

$$E\subseteq A$$

If

$$C\cap D=\emptyset$$

then the representations of oriented sets in $A^{(+^*_3)}$ will not intersect with the representations of unoriented sets in $A^{(+^*_3)}$.

Fulfilled:

$$A^{(+^*_3)}\subseteq A^{(+^*_1)}$$

Together with generalized finite sets, representations of countable non-identically-equal generalized hereditarily finite sets can be mapped one-to-one onto hereditarily finite sets.

The representation of $A^{(+^*_4)}$ is analogous to $A^{(+^*_2)}$.

The representation of classes of unorient mixed countably non-identically-equal generalized hereditarily finite sets can be given according to the expressions:

$$\tau_5(A)=2^{\left(\emptyset+\sum_{x\in A}\{x\}\right)}$$

$$\rho_5=\rho_3$$

$$\sigma_5=\sigma_3$$

$$\pi_5=\pi_3$$

The representation of classes of orient mixed countably non-identically-equal generalized hereditarily finite sets can be given according to the expressions:

$$\tau_6(A)=A^{\left(*^1\right)}$$

$$\rho_6=\rho_3$$

$$\sigma_6=\sigma_3$$

$$\pi_6=\pi_3$$

Consider the representation of $A^{(+^*_6)}$ in generalized hereditarily finite sets of the class of countably non-identically-equal generalized hereditarily finite sets on the alphabet $\dot{A}$ with codes $\ddot{A}$, which can be given according to expressions:

$$\tau_7(B)\overset{def}{=}B$$

$$\rho_7(\langle A,B\rangle)=C\times(A\cup B)^*$$

$$\sigma_7(\langle A,B\rangle)=D\times\pi_7\left(2^{\emptyset+\sum_{x\in A\cup B}\{x\}}\right)$$

$$\pi_7(B)=B$$

$$a_{2k+1}=\begin{cases}\dot{a}_k\left|k\leq\left|\dot{A}\right|\right.\\ d(2*k+1)\left|k>\left|\dot{A}\right|\right.\end{cases}$$

$$a_{2k}=d(2*k)$$

$$C\cup D=E$$

$$E=\left\{a_{2k}\middle|(a_{2k}\in\dot{A})\bigwedge(k\in\mathbb{N})\right\}$$

$$E \subseteq A$$

where

$$\ddot{A} \cap \mathbb{N} \neq \emptyset$$

If a number in $\ddot{A}$ is the Ackermann coding of the corresponding set in the representation of enumerably self-founded generalized hereditarily finite sets by hereditarily finite sets, then such a set is self-founded, i.e. this set is considered as an element of the alphabet at the same position in $\dot{A}$ as its code $\ddot{A}$.

To ensure the enumerability of self-founded generalized hereditarily finite sets, it is required to ensure that they are all finitely mutually-founded, that is, that they are not infinitely mutually-founded. In this case, the extensional closure will be reduced to a finite structure and there will be an algorithm for comparing these structures provided they are reduced to canonical form.

In general, it is not possible at this stage to canonize the representation of a structure that is a union of structures of all sets of a given class. To canonize the representation of such structures requires a separate investigation of the conditions under which this may be possible.

The embedding of structures in hereditarily finite sets and natural numbers (Fig. 1), using Ackermann coding, gives a structure isomorphic to the Rado graph [62] which is universal for any graph, i.e. it allows us to isomorphically enclose any graph and its supergraphs. Stability is one of the important properties of Rado Graph has. There are known studies of universal uncountable structures and corresponding theories including studies of the property of their stability [60]. As for uncountable structures [31] associated with operational semantics [16], [55] one direction of research is to study an approach based on the use of decision procedures without the presence of enumeration procedures as well as the use of Büchi automata [61] and their hierarchy.

Using the considered classes of structures, one can represent arbitrary finite graph, pseudograph, multigraph, metagraph, hypergraph structures, including abstract simplicial complexes [28], their combinations and others.

The applied value of the considered classes is the possibility of algorithmic construction of canonical forms (representations) of knowledge structures.

IV. Similarity, proximity, other attribute and invariants of structures of meaning space and corresponding models

The analysis of structural properties implies consideration of topological relations and relations of similarity (similarity and analogies) and difference. These relations can be algorithmically realized within the framework of the knowledge specification model in accordance with the knowledge integration model. The knowledge specification model, by considering finite structures, allows for the decidability of the corresponding analysis algorithms.

As for the similarity relations, they can be formed from property detection relations or non-detection relations. A property detection relation itself can be a similarity relation when the property itself is revealed in its specification (i.e., the relation is reflexive); as a rule, such relations are transitive. If a property is not revealed in its specification, but is revealed in other structures, then all these structures can be united into a class of (pairwise) similar structures.

In the first case, such relations can be reduced to the identification of full or partial embeddings, and morphisms: isomorphic, homomorphic or homeomorphic embedding.

The composition of two binary similarity relations is a binary similarity relation.

The union of two ij-similarity relations is a ij-similarity relation.

The intersection of two ij-similarity relations is a ij-similarity relation.

The difference of a ij-similarity relation and a ij-difference relation is a ij-similarity relation.

The difference of two ij-similarity relations, the first of which is a subset of the second, is a ij-difference relation.

The difference of a ij-difference relation and a second relation is a ij-difference relation.

The difference of two ij-difference relations is a ij-difference relation.

The union of two ij-difference relations is a ij-difference relation.

The intersection of two ij-difference relations is a ij-difference relation.

In the second case, similarity can be formed as an equivalence relation or a partial order relation on the set of structures for which no enumerated embeddings are identified (in the first case).

Topological properties [47] can be investigated through the consideration of operations on structures that correspond to topological closure.

Among the topological similarities we can distinguish: similarity on the inclusion of closures, similarity on the inclusion of derived sets, similarity on the inclusion of touch points and others.

Among topological similarities we can distinguish: similarity by equivalence of closures, similarity by equivalence of derived sets, similarity by equivalence of touch points and others.

In addition to search operations for extensional closures, we can identify closure operations on the class of automorphic elements of the structure.

It is also possible to identify search operations by a single (simple) pattern (reverse homomorphism) in exten-

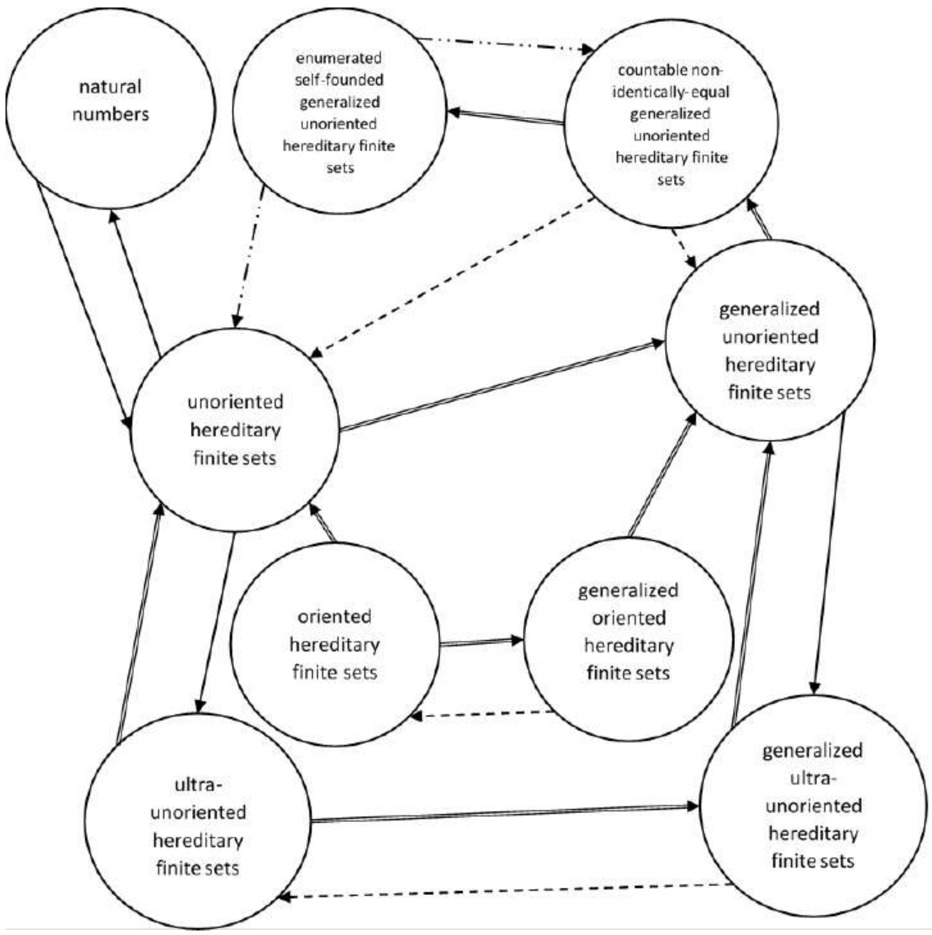


Figure 1. General finite structures classes

sional closures covering all elements of some distance-defined neighborhood of a concept, and all elements incident to the concept must be all elements of the automorphism class in this neighborhood. Among such neighborhoods we can choose a maximal neighborhood. The corresponding neighborhood can be considered as a closed set if it is not included in any other such neighborhood. In this case, the search operation on the corresponding pattern can also be considered as a closure operation.

The number of mappings of a pseudograph to the ordinal scale distinguishable with precision up to the order of vertices does not exceed $|V|^{|V|}$. $|V|^{|V|}$.

The number of mappings of the pseudograph to the metric scale distinguishable with order-of-magnitude accuracy of pairwise distances does not exceed $\left(\left|V^2\right|-|V|\right)^{\left(\left|V^2\right|-|V|\right)}$ (k-ary generalized distances – $\left(\left|V^k\right|-|V|\right)^{\left(\left|V^k\right|-|V|\right)}$).

For finite pseudographs and other representations of G for which the smallest A is found:

$$G \in A^{(+^{*}_{7})}$$

a quantitative attribute model can be based on the following class of quantitative features:

$$\mathbb{N}r_{+}{}^{A^{(+^{*}_{7})}}$$

where

$$\mathbb{R} \subseteq \mathbb{N}r$$

All closures defined on finite structures are finite. Finite structures, including finite (extensional) closures, have the following (global) characteristics: length, width, graph dimension, neg-entropy, sets (sets) of local characteristics and others.

Global characteristics can be considered as invariants [17], [68], on the basis of which characteristics similarity or proximity measures (metrics or pseudo-metrics) can be calculated.

Local (conditional) numerical characteristics include: centrality, metrics [10] and others.

Another example of numerical features are measures of scalar and coscalar product for introorthogonal sets considered in taxonomy management problems [].

The features for a finite set can be computed as a generalized or stepped average:

$$\sqrt[p]{\frac{\sum_{x \in S} x^p}{|S|}}$$

For a finite set of objects (an element of the product of metric spaces), metrics based on the (stepped) mean can be considered

$$d\left(\langle P, Q\rangle\right) = \sqrt[p]{\sum_{x \in P} \sum_{y \in Q} d\left(\langle x, y\rangle\right)^p}$$

The transition from an unbounded metric to a metric as a bounded monotone (fuzzy) measure, can be realized according to:

$$\lim_{z \to d(\langle x,y\rangle)} \frac{z}{1+z}$$

or any other semi-additive monotonically increasing bounded function from the point $(0;0)$:

$$f(x+y) \leq f(x) + f(y)$$

For a metric on linear representations by generalized strings, we can use either the edit distance between them or the minimum of the edit distances between the results of the action of a subgroup of the symmetric group of permutations of linear representations by generalized strings.

Let us consider an algorithm for finding the extensional closure of a concept:

1 The distensible set of the extensional closure is empty.
2 Include a concept in the current front.
3 Construct a new next front by going deep down the extensional.
4 Subtract the current front from the next front.
5 Add the elements of the current edge to the distensional set of the extensional closure.
6 If the next edge is empty return the extensional closure.
7 Make the next edge the current edge, go to step 3.

Consider an algorithm for computing the metric (quasi-metric) on the union of (finite) extensional closures of concepts:

1 Find the extensional closure of the first concept.
2 Find the extensional closure of the second concept.
3 If the intersection of the found closures is empty then return the metric equal to infinity.
4 If one concept lies in the extensional closure of another, then find the distance from the other to the first and return the metric equal to the smallest of the found ones.
5 Descend from each concept deep into the extensional closure to the intersection of the extensional closures and remember the lengths of descent as A and B.
6 Stepwise descent in the intersection of closures to continue until the moment of meeting, calculate the corresponding lengths A' and B'.
7 Return the metric $A + A' + (B + B') * 1$.

Each static structure (with its own denotational semantics [19]) and its characteristics can be related (see Fig.6, Fig.7) to the dynamic structure of a (formal) information processing model based on that static structure and its operations (possibly within $A^{(+\frac{*}{?})}$) [57]. Such dynamic structures and their corresponding operations are described by a big-step operational semantics. The connection between small-step [16], [52] and big-step operational semantics [6] can be revealed through the connections between the operational-information space and the (formal) information processing model.

Operational information space can be described as follows. A set of data instances V, a set of channels P, a set of operators K, a set of (commutative) operations O, a set of configurations C, a sequence of configurations R

$$R \subseteq C \times C$$

$$O \subseteq 2^{V^2 \times V} \cup 2^{V \times V^2}$$

$$C \subseteq 2^{K \cup P \cup (P \times (V \cup K)) \cup (K \times (O \cup P))}$$

If a new parameter (with a value) is added to a configuration operation, it is possible to jump to the updated configuration that contains it.

If a new operation is added to the configuration parameters, it is possible to switch to the updated configuration that contains it.

It is possible to switch to a configuration that does not contain an operation.

It is possible to jump to a configuration that does not contain a parameter that is not used by an operation.

Consider the flows of an open (acyclic (Fig.2)) or closed path (cyclic (Fig.3)) for the corresponding open or closed structures and assume the following requirements for its flow c_{ij} [57].

Each edge is mapped to a flow (energy) c_{ij}. To each vertex s there is a flow $c_s = \sum_{j=1}^{n} c_{sj}$. In addition to the forward flow, the reverse flow $c_{ji}^{-1}, c_s^{-1} = \sum_{j=1}^{n} c_{sj}^{-1}$ is also computed. Their difference is equal to: $d_{ij} = c_{ij} - c_{ji}^{-1}$, $d_s = c_s - c_s^{-1}$. The direct (local) amplitude is calculated as follows $p_{ij} = \frac{d_{ij}}{c_i + c_i^{-1}} + \frac{1}{\sum_{j=1}^{V} a_{ij}}$.

$$\sum_{j=1}^{N} c_{ij} = \sum_{j=1}^{N} c_{ji}$$

$$c_{ij} = \frac{\sum_{j=1}^{N} c_{ij}}{\sum_{j=1}^{N} a_{ij}} * a_{ij}; c_{ij} * \sum_{j=1}^{N} a_{ij} = a_{ij} * \sum_{j=1}^{N} c_{ij}$$

Also in matrix form we have:

$$A^T * C = (A * 1) \bullet C$$

For the structure in Fig.3 and its forward flow we have:

$$\begin{cases} c_{11} = c_{12} \\ c_{56} = c_{57} \\ c_{89} = c_{814} \\ c_{1112} = c_{1113} \end{cases}$$

Let us find the minimum solution in natural numbers.

As a result of fulfillment of these requirements we obtain the following table of results (see Fig.3, Fig.4, Fig.8, Fig.5).

Table I
Table of dynamic structure characteristics

Edge number	*Flow difference*	*Forward amplitude*	*Backward amplitude*
0	3	19/32	35/35=1
1	-3	13/32	29/29=1
2	3	35/35=1	35/35=1
3	-3	29/29=1	29/29=1
4	3	35/35=1	19/34
5	-1	15/29	15/34
6	-2	14/29	30/30=1
7	2	34/34=1	34/34=1
8	-2	30/30=1	30/30=1
9	0	16/34=8/17	16/30=8/15
10	2	18/34=9/17	18/35
11	-2	30/30=1	14/30=7/15
12	-2	30/30=1	30/30=1
13	1	17/30	33/33=1
14	-3	13/30	29/29=1
15	1	33/33=1	17/35
16	-3	29/29=1	29/29=1
17	3	35/35=1	35/35=1
18	-3	29/29=1	13/32
19	3	35/35=1	19/32

An analogous result can be obtained for an open (non-closed) structure (Fig. 2).

Each strongly connected structure has an (own) period T, which is the GCD of all periods (lengths of simple cycles) in this structure, and has a partition by levels of wave fronts corresponding to this period. We will call the number of these levels the length of the structure $L = T$. The length L of an acyclic structure is the maximum length of the shortest route for two connected vertices. Each (acyclic) structure has a mapping W of the set of numbers of moments of time [9] to the set of subsets of vertices by levels of wavefronts at given moments of time, the number of which does not exceed the length and diameter of the structure. Each wavefront has an energy $E(t) = \sum_{s \in W(t)} c_s$. The wavefront energy can be different from 1. The amplitude at the top of the wavefront $p_s^t = \frac{c_s}{E(t)}$ is in the interval $[0; 1]$. The average amplitude is inversely proportional to the number of front elements $\frac{E(t)}{|W(t)|}$.

The amount of information of the wavefront at the moment t is expressed.

$$-\sum_{s \in W(t)} \left(\frac{\left|U_s^{(t)}\right|}{p_s^t} * \ln \frac{\left|U_s^{(t)}\right|}{p_s^t} \right)$$

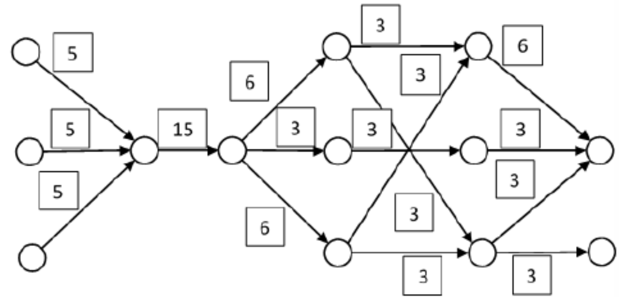


Figure 2. Acyclic orgraph weighted structure.

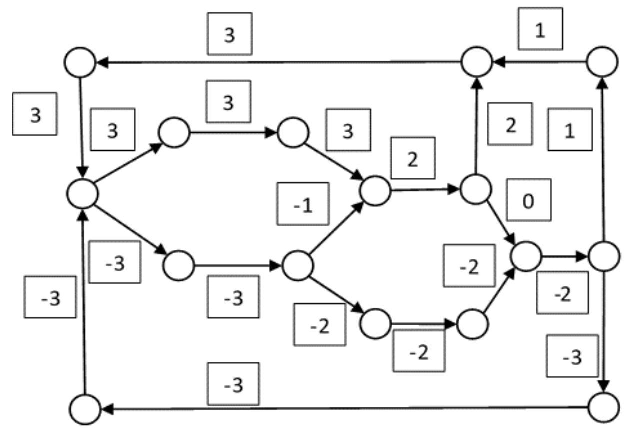


Figure 3. Strong connected flow difference weighted orgraph.

Set of undistinguishable vertices of the wavefront

$$\{s\} \subseteq U_s^{(t)} \subseteq W(t)$$

The average (arithmetic) amount of information of the structure:

$$-\frac{1}{T} * \sum_{t=1}^{T} \sum_{s \in W(t)} \left(\frac{\left|U_s^{(t)}\right|}{p_s^t} * \ln \frac{\left|I_{(t)}\right| * \left|U_s^{(t)}\right|}{T * p_s^t} \right)$$

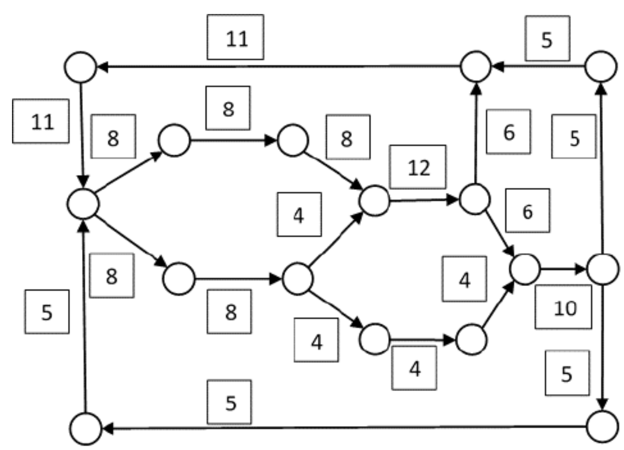


Figure 4. Strong connected orgraph forward flow.

$$A = \begin{matrix} 0 & 3 & -3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 & -2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -0 & 0 & 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -3 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -3 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 3 \\ -3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{matrix}$$

Figure 5. Strong connected orgraph flow difference weights matrix.

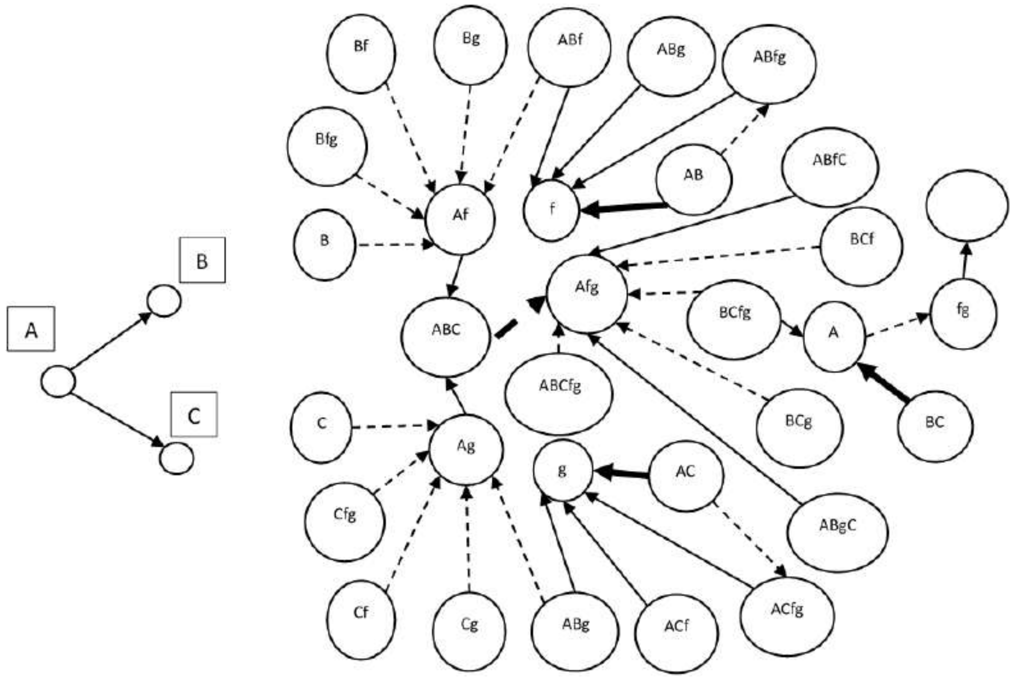


Figure 6. Asymmetric ontological structure with corresponding dynamic structure.

A set of indistinguishable moments in time

$$\{t\} \subseteq I_{(t)} \subseteq Dom\,(W)$$

The information in a strongly connected structure will be called real (elliptic), and in an acyclic structure will be called imaginary (hyperbolic).

Two kinds of structures are considered: a (finite) acyclic graph and a strongly connected pseudograph. An arbitrary (finite) pseudograph structure can be decomposed into its connected components. An arbitrary (finite) connected pseudograph structure can have different kinds of substructures and, in particular, can be broken down into the two kinds of structures discussed earlier: strongly connected components (subpseudographs), acyclic graphs (subgraphs). However, within a structure, these substructures can have different relationships and fulfill different roles [66]. Let us describe different types of substructures according to their roles (relations) performed (available) in the structure [66].

Resonators are maximal strongly connected subpseu-

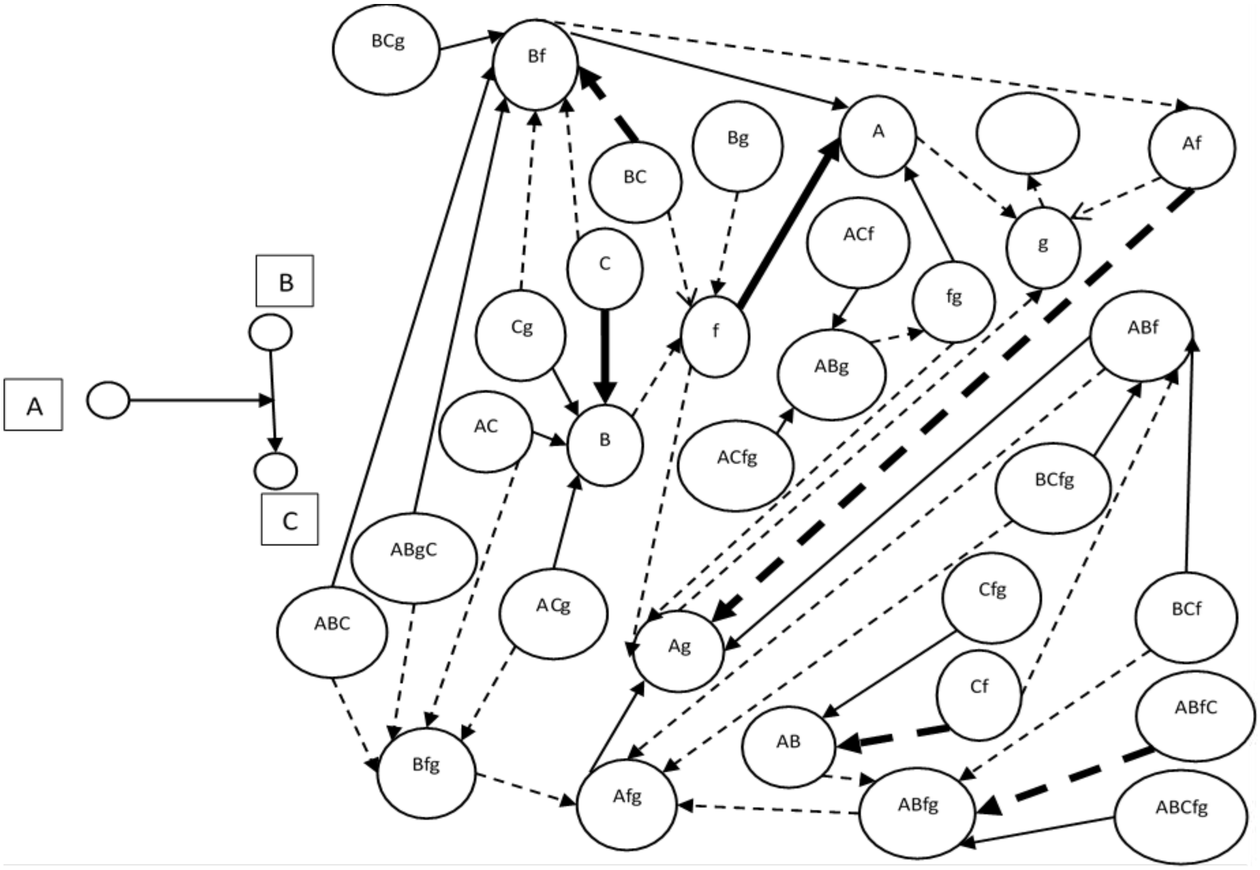


Figure 7. Asymmetric ontological structure with corresponding dynamic structure.

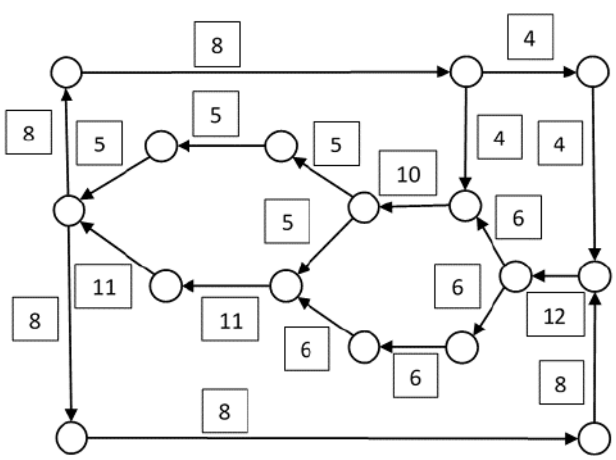


Figure 8. Strong connected orgraph backward flow.

dographs.

Sensors (receptors) are acyclic subgraphs whose elements are not reachable from any resonator.

Dispensers (effectors) are acyclic subgraphs whose elements are not reachable from any resonator.

Transmitters are acyclic subgraphs whose elements are reachable from at least one resonator and from whose elements at least one other (different) resonator is reachable.

Transmitters and transmitters are consumers.

Sensors and transmitters are suppliers.

We can also consider generating resonators (not reachable from other resonators) and consuming resonators (reachable from other resonators).

Among resonators, we can distinguish unimodal (harmonic) resonators and multimodal (non-harmonic) resonators (multimodal waveform). All multimodal resonators are consuming resonators. Each mode is phase shifted by less than a period.

A resonator-free subgraph can be extracted from any structure, which is the set of all vertices and edges reachable from the receptor elements.

The remaining edges together with their initial and final vertices form the resonator subpseudograph.

A receptor element is called a sensor element if it has no supplier.

A dispenser element is called a dispenser element if it has no consumer.

If there are no suppliers (sensors or transmitters) whose elements are suppliers to the generating resonator, then all its elements are sensor elements.

If there are no consumers (dispensers or transmitters) whose elements are consumers of the resonator, then all its elements are dispenser elements.

The method for determining the capacitive characteristics of structures used in problem solving is summarized in the following principles.

Each resonator has a period, which is the partial GCD T of its own period and periods of all (its) suppliers, from which it consumes, to the GCD of this GCD (T) and all its divisors k, for which the convolution of the phases $k * n$ with the phases (taking into account their shifts) of all modes (waveforms) of the signal is equal to $k\ T - k * ((T - k)!)$.

For each consumer (dispenser or transmitter) element, the period is calculated similarly, except that instead of the GCD of the period itself and the periods of all (its) suppliers from which it consumes, the GCD of the periods of all its suppliers is taken.

For each supplier element, the potential period, which is the period of the consumer in the inverted pseudograph (inverse ratio pseudo-graph), can be similarly determined.

For sensor elements, it may be assumed, unless otherwise accepted, that their period is equal to their potential period.

For each sensor element (supplier), a period can be calculated that is equal to the LCM of the periods of all its consumers (taking into account the phase shift).

The entropy of a set of resonators is calculated on a period equal to the GCD of periods of all resonators of this set.

If resonators are present, then the (maximum possible) entropy of the entire pseudograph is valid and can be computed as the greatest entropy of the greatest entropy of the greatest entropies of the smallest sets of resonators cutting the set (paths) of the set of smallest sets of paths connecting all sensor elements to all dispenser elements.

If all sensor elements and dispenser elements are resonator elements, then the (minimum required) entropy of the entire pseudograph is valid and can be computed as the smallest entropy of the smallest entropies of the smallest sets of resonators cutting the set (paths) of the set of smallest sets of paths connecting all sensor elements to all dispenser elements.

If all sensor elements and dispenser elements are not resonator elements, then the (minimum necessary) entropy of the whole pseudograph is invalid (imaginary) and is computed on the period equal to the LCM of all periods (elements) of the subpseudographs of the pseudograph and the maximum of the lengths of the (simple) paths from the sensor element to the dispenser element.

The entropy of an unbound pseudograph can be calculated as the average (minimum, maximum, etc.) of the entropies of its components on a period equal to the LCM of periods (on which calculations for) its components were performed.

Conclusions

The classification of enumerable finite structures and their representations is proposed, relations between classes of this classification are considered. The classification is oriented on unification of knowledge representation with finite structure and algorithmization of solutions of problems of investigation of topological and metric properties of structures of meaning space in order to exclude redundant fragments of knowledge at representation in meaning space.

Approaches are considered and concepts for investigation of structural-topological and metric properties of structures of the sense space for the purpose of optimization of structures of knowledge bases are proposed. General quantitative evaluations of mappings of finite structures with accuracy up to order preservation to ordinal and metric scales in the study of corresponding properties of these structures are given. An approach to the classification of quantitative features of finite structures is proposed. An algorithm for metric computation on the union of extensional closures of sense space concepts is proposed.

A model and a method for computing entropy (as one of the invariants) for finite dynamic structures of information processing models (in accordance with the models of graph dynamical system and generalized finite automaton [64]) based on analytical calculation of transition probabilities on the state graph in accordance with its structure are proposed.

The features of structures of semantic space that can be used as invariants in order to reduce the time to identify redundant fragments in the semantic space are considered.

A model of operational-information space is proposed, which corresponds to the model of model-parametric space [67], is oriented to solving the problems of knowledge management [15], [50], [56] in information processing and the study of the relationship between the attributes of structures with operational semantics expressed by the operational semantics of small and large step.

An approach to the consideration of infinite structures through (limit) sequences of finite structures converging to them is proposed in accordance with the classification of finite structures, the model of knowledge specification [5], the algebraic system of extensible sets and the metamodel of semantic space [56].

References

[1] V.V. Golenkov. Otkrytyi proekt, napravlennyi na sozdanie tekhnologii komponentnogo proektirovaniya intellektual'nykh sistem [An open project aimed at creating a technology for the

component design of intelligent systems], Otkrytye semanticheskie tekhnologii proektirovaniya intellektual'nykh system [Open semantic technologies for intelligent systems], 2013, pp. 55—78.

[2] V.V. Golenkov, N.A. Gulyakina, D.V. Shunkevich. Otkrytaya tekhnologiya ontologicheskogo proektirovaniya, proizvodstva i ekspluatatsii semanticheski sovmestimykh gibridnykh intellektual'nykh komp'yuternykh system [Open technology for ontological design, production and operation of semantically compatible hybrid intelligent computer systems], Minsk, Bestprint, 2021. 690 p.

[3] Z. Pawlak. Rough sets. International Journal of Parallel Programming, 1982, vol. 11, no. 5, pp. 341—356.

[4] D. Dubois, H. Prade. Rough fuzzy sets and fuzzy rough sets International Journal of General Systems, 1990, vol. 17, no. (2-3), pp. 191-–209.

[5] V.P. Ivashenko. Modeli i algoritmy integratsii znanii na osnove odnorodnykh semanticheskikh setei (disc. na soiskanie stepeni kand. tekhn. nauk: 05.13.17) [Models and algorithms for knowledge integration based on homogeneous semantic networks (thesis for the degree of Candidate of Technical Sciences: 05.13.17)] , Minsk, BGUIR, 2014, 152 p.

[6] V.P. Ivashenko Operatsionnaya semantika mnogoagentnykh sistem obrabotki znanii [Operational semantics of multi-agent knowledge processing systems], Information Technologies and Systems, 2020, Minsk, BGUIR, pp. 78–79.

[7] V.P. Ivashenko. Modeli i algoritmy integratsii znanii na osnove odnorodnykh semanticheskikh setei [Models and algorithms for knowledge integration based on homogeneous semantic networks], Otkrytye semanticheskie tekhnologii proektirovaniya intellektual'nykh system [Open semantic technologies for intelligent systems], 2015, Minsk, BGUIR, pp. 111—132.

[8] V.P. Ivashenko Tekhnologiya razrabotki programmnykh komponentov intellektual'nykh sistem na osnove integratsionnoi platformy [Technology for the development of software components of intelligent systems based on an integration platform], Information Technologies and Systems, 2021, Minsk, BGUIR, pp. 84–85.

[9] V.P. Ivashenko. Ontologicheskaya model' prostranstvenno-vremennykh otnoshenii sobytii i yavlenii v protsessakh obrabotki znanii [Ontological model of the space-time relations of the event and the phenomenon in the processes of knowledge processing] 2017, vol. 5, no. 107, Minsk, BSUIR, pp. 13–17.

[10] V.P. Ivashenko. String processing model for knowledge-driven systems. Minsk, Doklady BGUIR, 2020, vol. 18, no. 6, pp. 33–40.

[11] Yu. Manin, M. Marcolli. Semantic spaces. Published, Location, 2016. 32 p. (arXiv)

[12] V.V. Golenkov, O.E. Eliseeva, V.P. Ivashenko. Predstavlenie i obrabotka znanii v parallel'nykh grafodinamicheskikh assotsiativnykh mashinak [Representation and processing of knowledge in parallel graphodynamic associative machines], Minsk, BGUIR, 2001. 412 p.

[13] G. Rozenberg, A. Salomaa. Handbook of Formal Languages, Volume 1: Word, Language, Grammar. Verlag, Berlin, Heidelberg, Springer, 1997. 873 p.

[14] B. Smith. Metody i algoritmy vychislenii na strokakh [Computing Patterns in Strings], Moscow, OOO "I.D. Williams", 2006. 496 p.

[15] D.A. Pospelov. Situatsionnoe upravlenie: teoriya i praktika [Situational management: theory and practice], Moscow, Nauka, 1986. 288 p.

[16] A. Blass. A game semantics for linear logic. Annals of Pure and Applied Logic, 1992, vol. 56, pp. 183–220.

[17] J.H. Conway. On Numbers and Games. AK Peters/CRC Press, 2000. 242 p.

[18] G. Plotkin. Call-by-name, call-by-value and the lambda-calculus. Theoretical Computer Science, 1975, vol. 1, no. 2. pp. 125–159.

[19] D. Scott, Ch. Strachey. Toward a mathematical semantics for computer languages. Proceedings of the Symposium on Computers and Automata, Microwave Research Institute Symposia Series, Polytechnic Institute of Brooklyn Press, New York, 1971, vol. 21, pp. 19—46.

[20] F. Daneš. On Prague school functionalism in linguistics. Functionalism in Linguistics, 1987, pp. 3–38.

[21] A. R. D. Mathias. Unordered pairs in the set theory of Bourbaki 1949, Archiv der Mathematik, 2010, vol. 94, pp. 1–10.

[22] W. Quine. On ordered pairs. Journal of Symbolic Logic, 1945, vol. 10, no. 3, pp. 95—96.

[23] J. Heijenoort. From Frege to Gödel. A source book in mathematical logic, 1879–1931. Cambridge, Mass., Harvard University Press, 1967, 664 p.

[24] C. Kuratowski. Sur la notion de l'ordre dans la Théorie des Ensembles. Fundamenta Mathematicae, 1921, vol. 2, no. 1,pp. 161–171.

[25] A.P. Morse. A Theory of Sets. Academic Press, 1965, 130 p.

[26] A. Levy. Basic Set Theory, Springer-Verlag, 1979, 391 p.

[27] K.J. Devlin. Fundamentals of contemporary set theory. Universitext. Springer-Verlag, 1979, 182 p.

[28] J.M. Lee. Introduction to Topological Manifolds, Springer, 2011, 452 p.

[29] G. W. F. Hegel. Nauka logiki [Science of Logic]: v 3-h t. T.1, Moscow, "Mysl'", 1970. 501 p.

[30] R.T.W. Arthur. Leibniz on Time, Space, and Relativity, Oxford University Press, 2022, 432 p.

[31] D. Hilbert. On the infinite. In Philosophy of Mathematics (1984), Selected Readings, Cambridge, Cambridge University Press, 2012, pp. 183–201.

[32] D. Bohm and B.J. Hiley. The Undivided Universe: An Ontological Interpretation of Quantum Theory, London, Routledge, 1993. xii + 397 p.

[33] D. Bohm. Wholeness and the Implicate Order, London, Routledge, 2002. 284 p.

[34] V.V. Nalimov, Zh.A. Drogalina. Real'nost' nereal'nogo. Veroyatnostnaya model' bessoznatel'nogo [The reality of the unreal. Probabilistic model of the unconscious], Moscow, Mir idei, AO Akron, 1995. 432 p.

[35] V.V. Nalimov. Spontannost' soznaniya. Veroyatnostnaya teoriya smyslov i smyslovaya arkhitektonika lichnosti [Spontaneity of consciousness. Probabilistic theory of meanings and semantic architectonics of personality], Moscow, Prometei, 1989. 287 p.

[36] V.V. Nalimov. Veroyatnostnaya model' yazyka. O sootnoshenii estestvennykh i iskusstvennykh yazykov [Probabilistic model of language. On the relationship between natural and artificial languages] Moscow, Nauka, 1979. 304 p.

[37] S. Roman, Advanced Linear Algebra, Graduate Texts in Mathematics, vol. 135 (2nd ed.), Berlin, New York, 2005. 488 p.

[38] M.A. Armstrong, Basic Topology [1979], Undergraduate Texts in Mathematics, Springer, 1983. 263 p.

[39] L. Collatz. Functional Analysis and Numerical Mathematics, New York, San Francisco, London, Academic Press, 1966. xx 473 p.

[40] J. Heinonen. Lectures on analysis on metric spaces, New York, Springer, 2001. X+141 p.

[41] H.H. Schaefer, M.P. Wolff. Topological Vector Spaces, New York, Springer New York Imprint Springer, 1999. 361 p.

[42] I. Mel'ˇcuk. Language: from Meaning to Text. Ed. by D. Beck. Moscow & Boston, 2016.

[43] J. Harris. Algebraic Geometry: A First Course, New York, Springer, 1992. 330 p.

[44] A. Kostrikin, Yu. Manin. Linear Algebra and Geometry, Gordon and Breach Science Publishers, 1997. 320 p.

[45] V.V. Martynov. V tsentre soznaniya cheloveka [At the center of human consciousness], Minsk, BGU, 2009. 272 p.

[46] V.V. Martynov. Yazyk v prostranstve i vremeni: k probleme glottogeneza slavyan (2-e izd.) [Language in space and time: to the problem of glottogenesis of the Slavs (2nd ed.)], Moscow, Editorial URSS, 2004. 106 p.

[47] C. Marijuan. Finite topologies and digraphs. Proyecciones (Antofagasta), 2010, vol. 29, pp. 291–307.

[48] R. Penrose. The Road to Reality, Oxford, Oxford University Press, 2004. 1094 p.

[49] A.K. Guts. Khronogeometriya. Aksiomaticheskaya teoriya otnositel'nosti [Chronogeometry. Axiomatic theory of relativity], Omsk, OOO "UniPak", 2008. 340 p.

[50] M. Bansal, D. Burkett, G. de Melo, Gerard, D. Klein, Dan. Structured Learning for Taxonomy Induction with Belief Propagation. 52nd Annual Meeting of the Association for Computational

Linguistics, ACL 2014 – Proceedings of the Conference, 2015, vol. 1, pp. 1041–1051.
[51] Yu.I. Zhuravlyov. Ob algebraicheskom podkhode k resheniyu zadach raspoznavaniya ili klassifikatsii [On an algebraic approach to solving problems of recognition or classification]. Problemy kibernetiki, 1978, vol. 33. pp. 5–68.
[52] V.V. Golenkov, D.V. Shunkevich. Agentno-orientirovannye modeli, metodika i sredstva razrabotki sovmestimykh reshatelei zadach intellektual'nykh sistem [Agent-based models, methodology and tools for developing compatible problem solvers for intelligent systems]. Programmnye produkty i sistemy, 2020, vol. 33, no. 3. pp. 404–412.
[53] J.A. Fodor. The Language Of Thought. Crowell Press, 1975. 214 p.
[54] A.N. Kolmogorov. Problems of Probability Theory. Theory of Probability & Its Applications, 1994, vol. 38, no. 2, pp. 177–178.
[55] A.N. Kolmogorov, V.A. Uspenskij K opredeleniyu algoritma [To the definition of the algorithm] UMN, 1958, vol. 13. no. 4.
[56] Ivashenko, V. General-purpose semantic representation language and semantic space Otkrytye semanticheskie tekhnologii proektirovaniya intellektual'nykh system [Open semantic technologies for intelligent systems] Minsk, BGUIR, 2022. vol. 6. pp. 41–64.
[57] V. Ivashenko, Structures and Measures in Knowledge Processing Models / V. Ivashenko // Pattern Recognition and Information Processing. Artificial Intelliverse: Expanding Horizons : Proceedings of the 16th International Conference, Minsk, October 17–19, 2023 / ed.: A. Nedzved. A. Belotserkovsky / Belarusian State University. – Minsk, 2023. – P. 16–21.
[58] J. Poelmans, D. Ignatov, S. Kuznetsov, G. Dedene Fuzzy and rough formal concept analysis: A survey. International Journal of General Systems, 2014, vol. 43. 10.1080/03081079.2013.862377.
[59] B. Ganter, R. Wille, Formal Concept Analysis: Mathematical Foundations, translated by C. Franzke, Springer-Verlag, Berlin, 1998. ISBN 3-540-62771-5
[60] S. Shelah, Classification theory and the number of nonisomorphic models, Studies in Logic and the Foundations of Mathematics (2nd ed.), Elsevier, 1990 [1978], vol. IX, no. 1.19, p.49. ISBN 978-0-444-70260-9
[61] B. Khoussainov, A. Nerode. Automata Theory and its Applications, Springer Science & Business Media, 6 December 2012.
[62] W. Ackermann, "Die Widerspruchsfreiheit der allgemeinen Mengenlehre". Mathematische Annalen, 1937 , vol. 114, pp.305–315.
[63] A. Kanamori, "Bernays and Set Theory", Bulletin of Symbolic Logic, 2009, vol. 15, no. 1, pp.43—-69.
[64] H.S. Mortveit, C.M. Reidys, An Introduction to Sequential Dynamical Systems. Universitext, 2007.
[65] A. Koestler. The Ghost in the Machine. London: Hutchinson. 1, 1990 reprint edition [1967], Penguin Group.
[66] V.P. Ivashenko Informatsionnye kharakteristiki ontologicheskikh struktur [Information characteristics of ontological structures], Information Technologies and Systems, 2023, Minsk, BGUIR, pp. 55–56.
[67] Yu.R. Val'kman, V.I. Gritsenko, A.Yu. Rykhal'skii. Model'no-parametricheskoe prostranstvo: teoriya i primenenie [Model-parametric space: theory and application], 2012, Kiev, Naukova Dumka. 192 p.
[68] B.A. Kulik. Logika estestvennykh rassuzhdenii. [Logic of natural reasoning.], 2001, SPb, Nevskii Dialekt. 128 p.

К ТЕОРИИ СМЫСЛОВОГО ПРОСТРАНСТВА

Ивашенко В. П.

Статья рассматривает модели для исследования структуры, топологии и метрических признаков смыслового пространства, использующего унифицированное представление знаний.

Рассмотрены классы конечных структур, соответствующие онтологическим структурам и множествам классического и неклассического вида, исследованы свойства перечислимости этих классов.

Предложено понятия операционно-информационного пространствм, как модели для исследования взаимосвязи операционной семантики онтологических структур большого и малого шага.

Рассмотрены количественные признаки и инварианты онтологических структур, ориентированные на решение задач управления знаниями.

Received 03.04.2024

Integration of Fuzzy Systems with Parametric Interpretation for Unified Knowledge Representation

Valerian Ivashenko
Department of Intelligent Information Technologies
Belarusion State University of Informatics and Radioelectronics
Minsk, Republic of Belarus
ivashenko@bsuir.by

***Abstract*—The paper considers the problem of stable interpretation of fuzzy logic models. An approach based on parameterized fuzzy logic is proposed, where each logical formula has a set of model parameters in addition to truth values. Parameterized fuzzy logic allows combining different fuzzy logic systems. Model parameters are used to calculate fuzzy truth values as a fuzzy measure. Models and model parameters related to metric spaces, consistent with metric sense spaces and being the basis for interpretation of fuzzy logic formulas on ontological models are considered.**

***Keywords*—Fuzzy logic, Parameterized fuzzy logic, Metric space, Fuzzy measure, Simplicial complex, Residual simplicial complex, Canonical form, Semantic metric, Semantic Space, Integration, Knowledge representation model, Knowledge processing model, Ontology, Unified representation of knowledge, Linear vector space, Parametric t-norm classes, Hilbert cube, Finite structure, Substructural logic**

I. Introduction

Approaches to integration of logical models in a general form are considered in [7].

One of the problems of integration of logical models of knowledge representation and processing [11], [17], [20] is to identify compatible models that provide construction of interpretations of corresponding logical formalisms [6]. If necessary, these models can be considered as part of the corresponding semantic space [7], [8], [19].

One of the broad classes of logical models is fuzzy logics [3]. There is a problem of unreliability of fuzzy logics due to uncertainties [18] existing at different stages of their application [15]. One of the stages is selection of a fuzzy logic model or system with the purpose of application for realization of reasoning and problem solving. It is not always clear how suitable the chosen fuzzy system is. This is due to the fact that interpretations (which are built in the process of fuzzy logical inference) connect logical constructions with abstract algebraic systems that have no definite connection with any subject area or its model. Moreover, for each fuzzy system a different algebraic system is considered, the connection of which with other algebraic systems also remains undefined. This high degree of uncertainty does not allow reliable use of fuzzy logic models which is one of the problems of fuzzy logics [15].

The choice of fuzzy logics is also conditioned by their rich internal and external diversity, which allows fuzzy logics to represent other non-classical logical [3], [9], [10] models by means of fuzzy logics. The diversity of fuzzy logical models leads, among other things, to the diversity of fuzzy logical operations (for example, such as triangular norms and conorms), the emergence of their classes and their parameterization within the corresponding subclasses.

In development of the idea of parameterization of logical operations, the concept of parameterized fuzzy logics [6] is proposed.

The two main parametric families of triangular norms (and corresponding conorms) [14] are: the Frank parametric family [12], [13] and the Schweitzer-Sklar parametric family.

When constructing interpretations of parametrized fuzzy logics we can distinguish their following types: interpretations on algebraic systems, interpretations on "amorphous" models, interpretations on concrete structural-static models.

Interpretations on algebraic systems are largely similar to traditional fuzzy systems, the general scheme of which is given in [6], and therefore we will not consider them in detail in this paper. Further we will consider examples of interpretations of formulas of parameterized fuzzy logics on "amorphous" models and on concrete structural-static models.

II. Interpretation and models of fuzzy logics

As an "amorphous" model, consider a model in which each fuzzy predicate is matched with a vector quantity (vector) A, which can be given by some unit (or zero) vector 1_A, specifying the direction of this quantity, in some linear basis of some vector space and a scalar $||A||$ in the range from 0 to 1, specifying the length of vector A. If and only if the length is 0 or the direction is given by a zero vector, then the vector quantity is equal to a zero vector, its length is 0, but its direction can be a non-zero vector.

The fuzzy negation operation in this model reverses the direction of the vector according to the expression:

$$-1_A$$

and its length according to the expression:

$$1 - ||A||$$

The next operation we will consider is the fuzzy conjunction. It should be noted that the fuzzy conjunction does not fulfill all the properties characteristic, for example, for triangular norms since this conjunction is parameterized, covering more than one triangular norm. A parameterized conjunction can naturally cover several triangular norms in one expression, so the properties of one triangular norm cannot be extended to such a conjunction.

To consider the result of the computation of such a fuzzy conjunction, let us consider 23 cases (variants, see Table I) of the spatial relation of vectors of a pair of its arguments ($A = 1_A * ||A||$ and $B = 1_B * ||B||$), which we will later reduce to a smaller number of cases.

Table I
Variants of relations of parameters of "amorphous" parameterized fuzzy logic

№	$\|\|A\|\| * \|\|B\|\|$	$cos\left(\langle A, B\rangle\right)$
0	0	$[-1;1]$
1	0	$[-1;1]$
2	$(0;1]$	1
3	$(0;1]$	$(0;1)$
4	$(0;1]$	$(0;1)$
5	$(0;1]$	$(0;1)$
6	$(0;1]$	0
7	$(0;1]$	$(-1;1)$
8	$(0;1]$	$(-1;1)$
9	$(0;1]$	$(-1;1)$
10	$(0;1]$	$(-1;1)$
11	$(0;1]$	$(-1;1)$
12	$(0;1]$	$(-1;1)$
13	$(0;1]$	$(-1;1)$
14	$(0;1]$	$(-1;1)$
15	$(0;1]$	$(-1;1)$
16	$(0;1]$	$(-1;1)$
17	$(0;1]$	$(-1;1)$
18	$(0;1]$	$(-1;1)$
19	$(0;1]$	$(-1;1)$
20	$(0;1]$	$(-1;1)$
21	$(0;1]$	$(-1;1)$
22	$(0;1]$	$(-1;1)$
23	$(0;1]$	-1

Due to symmetry (commutativity of the fuzzy conjunction operation), the number of these cases (variants) can be reduced to 16 which in turn are reduced to 10 (see Table II) as a result of decomposition. The result of the initial variant (case) is the arithmetic mean of the variants (cases) into which it is decomposed.

Table II
Decomposition of variants of parameter relations of "amorphous" parameterized fuzzy logic

Symmetry of variants		**Decomposition into variants**	
1	1	0	
2	2	5	
3	5	6	
4	4	7	
6	6	8	
7	7	1	1
8	11	1	2
9	15	1	3
10	19	1	4
12	12	2	2
13	16	2	3
14	20	2	4
17	17	3	3
18	21	3	4
22	22	4	4
23	23	9	

Hypothetically, variants 13 and 16 are impossible.

Let us consider these variants sequentially.

As a result of the operation of fuzzy conjunction of two arguments ($A = 1_A * ||A||$ and $B = 1_B * ||B||$), we must obtain a vector quantity given by two parameters: a vector and a scalar.

For the 0th variant, the vector coincides with the vector of a non-zero argument or is calculated by the formula:

$$(1_A + 1_B) / \sqrt{(1_A + 1_B) * (1_A + 1_B)}$$

For all variants except for the 0-th (23rd) we will calculate the vector by the formula:

$$(1_A + 1_B) / \sqrt{(1_A + 1_B) * (1_A + 1_B)}$$

The proposed formula has the advantage of a more convenient model for modeling traditional fuzzy logics but alternative expressions for calculating the vector are possible:

$$(A + B) / \sqrt{(A + B) * (A + B)}$$

$$\frac{(A + B + 1_A + 1_B)}{\sqrt{(A + B + 1_A + 1_B) * (A + B + 1_A + 1_B)}}$$

and others.

For two nonzero noncollinear vectors A and B, the vector $H_{AB} = A * u + B * v$ from their common origin to the intersection of the perpendiculars to their ends can be expressed:

$$H_{AB} = \frac{A * (A * (B - A)) * B^2 + B * ((A - B) * B) * A^2}{(A * B)^2 - A^2 * B^2}$$

from

$$(A * u + B * v - B) * B = (A * u + B * v - A) * A = 0$$

$$\begin{cases} u * (A * B) = (1 - v) * B * B \\ v * (A * B) = (1 - u) * A * A \end{cases}$$

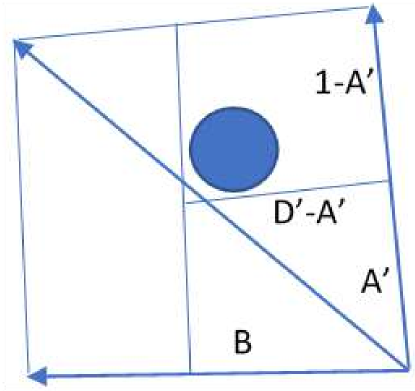


Figure 1. Variant 1 illustration.

$$\begin{cases} u = (A*(B-A))*B^2/\left((A*B)^2 - A^2*B^2\right) \\ v = ((A-B)*B)*A^2/\left((A*B)^2 - B^2*A^2\right) \end{cases}$$

Let the following:

$$D = H_{AB}$$

$$D' = H_{A'B}$$

$$1_{A'} = -1_A$$

$$1_D = H_{1_A 1_B}$$

$$1_{D'} = H_{1_{A'} 1_B}$$

Variant 1: The angle between vectors A (vector $A' = A - 1_A$) and B is obtuse ($A*B<0$), the perpendicular to vector A' intersects the perpendicular to B before B $||B|| \leq ||A'||/cos\left(\langle A', B\rangle\right)$ (see Fig.1) then the length of the result is equal to the ratio of areas:

$$\frac{(2*||D'-A'|| - ||1_{A'}-A'||)*ctg\left(\langle A', B\rangle\right)*||1_{A'}-A'||/2}{\left(1_{A'}*\left(1_{D'}-1_{A'}\right)+1_B*\left(1_{D'}-1_B\right)\right)/2}$$

Variant 2. The angle between vectors A (vector $A' = A - 1_A$) and B is obtuse ($A*B<0$), the perpendicular to vector A' intersects B before the perpendicular to it $||B|| > ||A'||/cos\left(\langle A', B\rangle\right))$ (see Fig.2) then the length of the result is equal to the ratio of areas:

$$\frac{\frac{\left(||1_{A'}||^2-||A'||^2\right)}{ctg(\langle A',B\rangle)} - \left(\frac{||1_{A'}||}{cos(\langle A',B\rangle)} - ||B||\right)^2 * ctg\left(\langle A', B\rangle\right)}{2*\left(1_{A'}*\left(1_{D'}-1_{A'}\right)+1_B*\left(1_{D'}-1_B\right)\right)/2}$$

Variant 3. The angle between vectors A (vector $A' = A - 1_A$) and B is obtuse ($A*B<0$), the perpendicular to vector A' intersects the perpendicular to B before B $||B|| \leq ||A'||/cos\left(\langle A', B\rangle\right)$ (see Fig.3) then the length of the result is equal to the ratio of areas:

$$\frac{||D'-A'||*||D'-A'||*tg\left(\langle A', B\rangle\right)/2}{1_{A'}*\left(1_{D'}-1_{A'}\right)+1_B*\left(1_{D'}-1_B\right)/2}$$

Variant 4. The angle between vectors A (vector $A' = A - 1_A$) and B is obtuse ($A*B<0$), the perpendicular to vector A' intersects B before the perpendicular to it

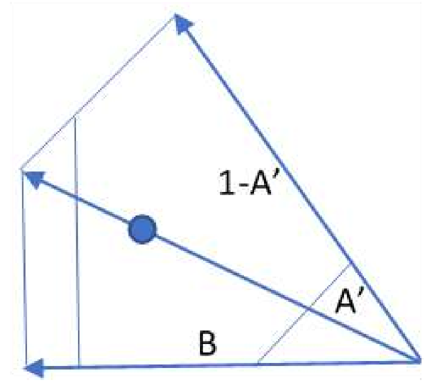


Figure 2. Variant 2 illustration.

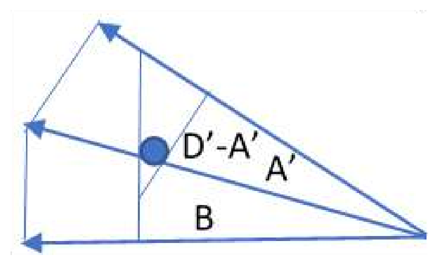


Figure 3. Variant 3 illustration.

$||B|| > ||A'||/cos\left(\langle A', B\rangle\right))$ (see Fig.4) then the length of the result is equal to the ratio of areas:

$$\frac{||B||*||B|| - ||A'||*||A'||*tg\left(\langle A', B\rangle\right)/2}{1_{A'}*\left(1_{D'}-1_{A'}\right)+1_B*\left(1_{D'}-1_B\right)/2}$$

Variant 5. Vectors A and B are co-oriented then the length of the result is: $min\left(\{||A||\} \cup \{||B||\}\right)$.

Variant 6. The angle between vectors A and B is acute ($A*B>0$), the perpendicular to vector A intersects B before the perpendicular to it $||B|| > ||A||/cos\left(\langle A, B\rangle\right)$ (see Fig.5) then the length of the result is equal to the

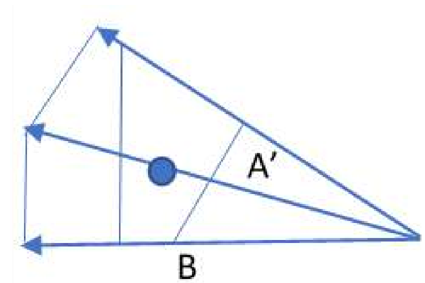


Figure 4. Variant 4 illustration.

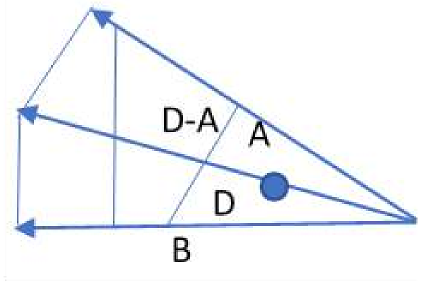


Figure 5. Variant 6 illustration.

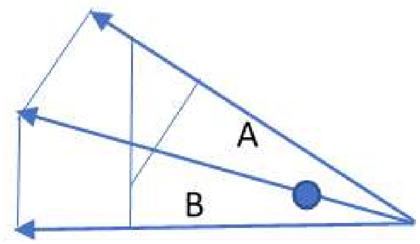


Figure 6. Variant 7 illustration.

ratio of areas:

$$\frac{A*(D-A)/2}{(1_A*(1_D-1_A)+1_B*(1_D-1_B))/2}$$

Variant 7. The angle between vectors A and B is acute ($A*B>0$), the perpendicular to vector A intersects the perpendicular to B before B $||B|| \le ||A||/cos(\langle A,B\rangle)$ (see Fig.6) then the length of the result is equal to the ratio of areas:

$$\frac{((A+B)*D-A*A-B*B)/2}{(1_A*(1_D-1_A)+1_B*(1_D-1_B))/2}$$

Variant 8. Vectors A and B are orthogonal $((A*B=0)\wedge(||A||+||B||>0))$ (see Fig.7) then the length of the result is equal to the ratio of areas: $||A||*||B||/1$.

Variant 9. Vectors A and B are differently directed, then the length of the result is equal to:

$$max(\{0\}\cup\{A+B-1\})$$

Properties of negation:

$$A =\sim(\sim A)$$

$$0 =\sim 1$$

$$1 =\sim 0$$

Properties of conjunction:

- zero element
$$A\widetilde{\wedge}0=0$$

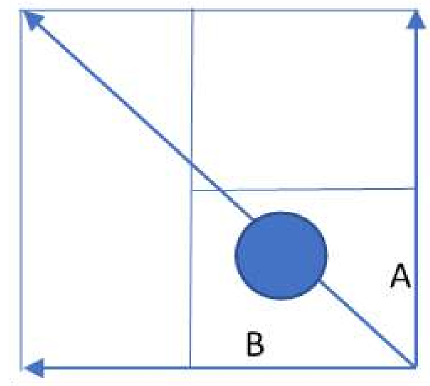


Figure 7. Variant 8 illustration.

- neutral element
$$A\widetilde{\wedge}1=A$$
- idempotency
$$A\widetilde{\wedge}A=A$$
- commutativity
$$A\widetilde{\wedge}B=B\widetilde{\wedge}A$$
- non-associativity
$$\neg\left(A\widetilde{\wedge}\left(B\widetilde{\wedge}C\right)=\left(A\widetilde{\wedge}B\right)\widetilde{\wedge}C\right)$$
- non-monotonicity
$$\neg\left(A\le B\to A\widetilde{\wedge}C\le B\widetilde{\wedge}C\right)$$
- monotonicity in direction
$$A\overleftarrow{\le}B\to A\widetilde{\wedge}C\overleftarrow{\le}B\widetilde{\wedge}C$$

Properties of disjunction:

$$A\widetilde{\vee}B=\sim\left((\sim A)\widetilde{\wedge}(\sim B)\right)$$

Properties of implication:

$$A \sim> B=\left((\sim A)\widetilde{\vee}B\right)$$

Properties of a fuzzy measure [16]:

$$A\widetilde{\wedge}B\le A$$

$$A\le A\widetilde{\vee}B$$

As concrete structural-static models we can consider finite models or simplicial complexes [1], and also we can consider their generalizations, for example, as a residual simplicial complex. Let us consider variants with simplicial complexes. An important question of such consideration is the canonical form of the corresponding simplicial complex.

For simplicial complexes and the corresponding sets defined by them, the (generalized) operations of union $\hat{\cup}$ and intersection $\hat{\cap}$ are naturally defined.

Each argument of the parameterized fuzzy expression can be matched with a residual simplicial complex as one of the parameters. We will also use the notion of residual simplicial complex to represent the results of parameterized fuzzy logics.

The residual simplicial complex can be given by $2*n$-simplicial complexes $\langle C_1, C_2, ..., C_{2*n}\rangle$ through an expression of the form:

$$C_1/(C_2/(\ldots/C_{2*n}))$$

For a residual simplicial complex the following is true:

$$C_{i+1} \subset C_i$$

$$C_{i+2} \subseteq \partial C_i$$

$$(X \in C_i \cap C_{i+1}) \rightarrow \exists Y\, (Y \in C_i/C_{i+1}) \wedge (\emptyset \subset Y \cap X)$$

$$\emptyset \subset C_{2*n}$$

$$\partial C = \bigcup_{X \in C} 2^X / \{X\}$$

The height of the residual complex is $2*n$.

We will consider simplicial complexes covering points of subsets of the set of points of the space spanned by the universal simplicial complex U whose residual simplicial complex is $\langle U, \emptyset, ..., \emptyset\rangle$.

The complement of $U\hat{/}C$ of the residual simplicial complex $\langle C_1, C_2, ..., C_{2*n}\rangle$ will be the height complex $2*m(m \leq n)$:

$$D_1/(D_2/(\ldots/D_{2*m}))$$

with such smallest T_1, T_k $(2 \leq k \leq n)$:

$$U/\left(\bigcup_{i=1}^{n} C_{2*i-1}/C_{2*i}\right) \subseteq T_1 \subseteq U$$

$$C_{k-1}/\left(\bigcup_{i=1}^{n-k+1} C_{2*i-1+k}/C_{2*i+k}\right) \subseteq T_k \subseteq C_{k-1}$$

$$D_1/\left(\bigcup_{i=1}^{m} D_{2*i}/D_{2*i+1}\right) = T_1/\left(\bigcup_{i=1}^{n} T_{2*i}/T_{2*i+1}\right)$$

$$((m < i) \wedge (i \leq n)) \rightarrow (T_i = \emptyset)$$

The intersection $O = I\hat{\cap}E$ of the two residual simplicial complexes I and E $(n \leq m)$ is the height complex $2*l(l \leq n*(2*m-n+1)$:

$$O_1/(O_2/(\ldots/O_{2*l}))$$

The tiers of the residual simplicial complex are filled in according to the tables Table III ($E_0 = E_{10}$), Table IV) in accordance with the order:

The difference $O = I\hat{/}E$ of the two residual simplicial complexes I and E $(n \leq m)$ is the height complex $2*l(l \leq n*(2*m-n+1)$:

$$I\hat{/}E = I\hat{\cap}\left(U\hat{/}E\right)$$

Table III
Computable operations for calculating the intersection of two residual simplicial complexes

$I_1 \cap E_1$	$I_2 \cup E_2$	$I_3 \cap E_1$	$I_4 \cup E_4$	$I_5 \cap E_1$	$I_6 \cup E_0$
$I_2 \cup E_2$	$I_2 \cap E_2$	$I_2 \cap E_2$	$I_4 \cup E_4$	$I_5 \cap E_2$	$I_6 \cup E_0$
$I_1 \cap E_3$	$I_2 \cap E_2$	$I_3 \cap E_3$	$I_4 \cup E_4$	$I_5 \cap E_3$	$I_6 \cup E_0$
$I_4 \cup E_4$	$I_4 \cup E_4$	$I_4 \cup E_4$	$I_4 \cap E_4$	$I_4 \cap E_4$	$I_6 \cup E_0$
$I_1 \cap E_5$	$I_2 \cap E_5$	$I_3 \cap E_5$	$I_4 \cap E_4$	$I_5 \cap E_5$	$I_6 \cup E_0$
$I_6 \cup E_6$	$I_6 \cup E_6$	$I_6 \cup E_6$	$I_6 \cup E_6$	$I_6 \cup E_6$	$I_6 \cup E_0$
$I_1 \cap E_7$	$I_2 \cap E_7$	$I_3 \cap E_7$	$I_4 \cap E_7$	$I_5 \cap E_7$	$I_6 \cup E_0$
$I_6 \cup E_8$	$I_6 \cup E_8$	$I_6 \cup E_8$	$I_6 \cup E_8$	$I_6 \cup E_8$	$I_6 \cup E_0$
$I_1 \cap E_9$	$I_2 \cap E_9$	$I_3 \cap E_9$	$I_4 \cap E_9$	$I_5 \cap E_9$	$I_6 \cup E_0$
$I_6 \cup E_0$	$I_6 \cup E_0$	$I_6 \cup E_0$	$I_6 \cup E_0$	$I_6 \cup E_0$	$I_6 \cup E_0$

Table IV
Sequence (transposed) of computable operations to compute the intersection of two residual simplicial complexes

1	2	3	6	7	12	13	18	19	24
2	4	4	6	8	12	14	18	20	24
3	4	5	6	9	12	15	18	21	24
6	6	6	10	10	12	16	18	22	24
7	8	9	10	11	12	17	18	23	24
24	24	24	24	24	24	24	24	24	24

The residuum $O = I\hat{\rightarrow}E$ of the two residual simplicial complexes I and E $(n \leq m)$ will be the height complex $2*l$ $(l \leq n*(2*m-n+1))$:

$$I\hat{\rightarrow}E = U\hat{/}\left(I\hat{/}E\right)$$

The union of $O = I\hat{\cup}E$ of the two residual simplicial complexes I and E $(n \leq m)$ is the height complex $2*l$ $(l \leq n*(2*m-n+1))$:

$$I\hat{\cup}E = U\hat{/}\left(\left(U\hat{/}I\right)\hat{/}E\right)$$

The value of a fuzzy expression (predicate) in parametric fuzzy logic can be calculated as the length $hvol_1(C)$, area $hvol_2(C)$, volume $hvol_3(C)$ or hypervolume $hvol_{dim(C)}(C)$ of a simplicial complex. For each simplicial complex with a basis in linear vector space, a minimal covering simplex can be given, and its dimension $dim(C)$, equal to the dimension of the maximal simplex in the complex, can also be computed. If the space is n-dimensional, the value of the fuzzy expression can be computed:

$$1 + 2^{-dim(C)} * \left(hvol_{dim(C)}(C) - 2\right)$$

For the corresponding fuzzy operations, the properties of the fuzzy measure will also be fulfilled:

$$A\widetilde{\wedge}B \leq A$$

$$A \leq A\widetilde{\vee}B$$

Another kind of non-classical logics [4], [5] are substructural logics in which (structural) properties of deducibility such as monotonicity, contraction (absorption) and others are violated. These include relevance logics and connexive logics which find out to justify causal implicative properties. Analyzing the properties of these logics involves clarifying the similarities and analogies of the schemes of these logics with other logics and models such as argumentation logics [2]. One of the prospects for further research is to study the connection of non-classical logics of this kind with the fuzzy models considered in this paper in the framework of causal and spatio-temporal relations of the semantic space.

III. Conclusions

Approaches and models to the interpretation of fuzzy logics are proposed. The proposed models can be used in the interpretation of fuzzy logic formulas on the basis of metric meaning space for finite structures in order to analyze or synthesize schemes of fuzzy logic inference systems relevant to the structures of ontologies of subject areas.

References

[1] J.M. Lee. Introduction to Topological Manifolds, Springer, 2011, 452 p.

[2] Finn V. K. Intellektual'nye sistemy i obshchestvo: Sbornik statei [Intelligent systems and society: Collection of articles], Moscow, KomKniga, 2006. 352 p.

[3] D.E. Pal'chunov, G.E. Yakh"yaeva Nechetkie logiki i teoriya nechetkikh modelei [Fuzzy logic and theory of fuzzy models], Algebra i logika [Algebra and Logic], vol. 54 no.1, 2015, p. 109–118; Algebra and Logic, vol. 54, no. 1, 2015, p. 74–80.

[4] Plesnevich, G.S. Binarnye modeli znanii [Binary knowledge models], Trudy Mezhdunarodnykh nauchno-tekhnicheskikh konferentsii «Intellektual'nye sistemy» (AIS'08) i «Intellektual'nye SAPR» (CAD-2008) [Proceedings of the International Scientific and Technical Conferences “Intelligent Systems” (AIS'08) and "Intelligent CAD" (CAD-2008)]. Nauchnoe izdanie v 4-kh tomakh [4 volumes], Moscow, Fizmatlit, 2008, vol.2. p. 424, 135–146 pp.

[5] V.N. Vagin Deduktsiya i obobshchenie v sistemakh prinyatiya reshenii [Deduction and generalization in decision-making systems], Moscow, Nauka. Gl. red. fiz.-mat. lit. 1988. 384 p.

[6] V.P. Ivashenko Ontologicheskie struktury i parametrizovannye mnogoznachnye logiki [Ontological structures and parameterized multivalued logics], Information Technologies and Systems, 2023, Minsk, BGUIR, pp. 57–58.

[7] V. Ivashenko Semantic space integration of logical knowledge representation and knowledge processing models Otkrytye semanticheskie tekhnologii proektirovaniya intellektual'nykh system [Open semantic technologies for intelligent systems], Minsk, BGUIR. Minsk, 2023, vol. 7. pp. 95–114.

[8] Ivashenko, V. General-purpose semantic representation language and semantic space Otkrytye semanticheskie tekhnologii proektirovaniya intellektual'nykh system [Open semantic technologies for intelligent systems] Minsk, BGUIR, 2022. vol. 6. pp. 41–64.

[9] G. Malinowski, Kleene logic and inference. Bulletin of the Section of Logic, 2014, vol. 1, no. 43.

[10] D. Maximov, Logika N.A. Vasil'eva i mnogoznachnye logiki [Vasil'ev Logic and Multi-Valued Logics.] Logical Investigations, 2016, vol. 22. 82–107 pp.

[11] V.P. Ivashenko. Modeli resheniya zadach v intellektual'nykh sistemakh. V 2 ch. Ch. 1 : Formal'nye modeli obrabotki informatsii i parallel'nye modeli resheniya zadach : ucheb.-metod. posobie [Models for solving problems in intelligent systems. In 2 parts, Part 1: Formal models of information processing and parallel models for solving problems: a tutorial] Minsk, BGUIR, 2020. 79 p.

[12] A. Hussain, K. Ullah, M. Khan, T. Senapati, S. Moslem, Complex T-Spherical Fuzzy Frank Aggregation Operators With Application in the Assessment of Soil Fertility / IEEE Access, 2023, vol. 11.

[13] E.P. Klement, M. Navara, Propositional Fuzzy Logics Based on Frank T-Norms: A Comparison // Fuzzy Sets, Logics and Reasoning about Knowledge 2000, vol. 15. 17–38 pp.

[14] F. Giannini, M. Diligenti, M. Maggini, M. Gori, G. Marra, T-norms driven loss functions for machine learning. Applied Intelligence, 2023, vol. 53. 1–15 pp.

[15] G. Heald Issues with reliability of fuzzy logic. Int. J. Trend Sci. Res. Develop, 2018, vol. 2 no. 6, 829-834 pp. 8, 2018.

[16] Zh. Wang, J. K. George Fuzzy Measure Theory, Plenum Press, New York, 1991.

[17] V.V. Golenkov. Otkrytyi proekt, napravlennyi na sozdanie tekhnologii komponentnogo proektirovaniya intellektual'nykh sistem [An open project aimed at creating a technology for the component design of intelligent systems], Otkrytye semanticheskie tekhnologii proektirovaniya intellektual'nykh system [Open semantic technologies for intelligent systems], 2013, pp. 55—78.

[18] A.S. Narinyani. NE-faktory: netochnost' i nedoopredelennost' – razlichie i vzaimosvyaz' [Non-factors: inaccuracy and underdetermination – difference and interrelation]. Izv RAN (RAS). Ser. Teoriya i sistemy upravleniya 5, 2000. pp. 44—56.

[19] Yu. Manin, M. Marcolli. Semantic spaces. Published, Location, 2016. 32 p. (arXiv)

[20] D.A. Pospelov. Situatsionnoe upravlenie: teoriya i praktika [Situational management: theory and practice], Moscow, Nauka, 1986. 288 p.

ИНТЕГРАЦИЯ НЕЧЁТКИХ СИСТЕМ И ИХ ПАРАМЕТРИЧЕСКАЯ ИНТЕРПРЕТАЦИЯ ДЛЯ УНИФИЦИРОВАННОГО ПРЕДСТАВЛЕНИЯ ЗНАНИЙ

Ивашенко В. П.

В статье рассматривается проблема устойчивой интерпретации нечётких логических моделей. Предлагается подход на основе параметризованной нечёткой логики, где каждая логическая формула кроме значений истинности имеет набор модельных параметров. Параметризованные нечёткая логика позволяет объединить различные нечёткие логические системы. Модельные параметры используются для вычисления значений нечёткой истинности, как нечёткой меры. Рассмотрены модели и модельные параметры, связанные с метрическими пространствами, согласуемыми с метрическим смысловыми пространствами и являющиеся основой для интерпретации нечётких логических формул на онтологических моделях.

Received 03.04.2024

Current State of ostis-systems Component Design Automation Tools

Maksim Orlov, Anna Makarenko, Ksenija Petrochuk
Belarusian State University of
Informatics and Radioelectronics
Minsk, Belarus
Email: orlovmaksimkonst@gmail.com, anna.makarenko1517@gmail.com, xenija.petrotschuk@gmail.com

***Abstract*—In the article, an approach to the design of intelligent systems is considered, focused on the use of compatible reusable components, which significantly reduces the complexity of developing such systems. The key means of supporting the component design of intelligent computer systems is the manager of reusable components proposed in the work.**

***Keywords*—Component design of intelligent computer systems; reusable semantically compatible components; knowledge-driven systems; semantic networks; OSTIS Technology.**

I. Introduction

The main result of artificial intelligence is not the intelligent systems themselves, but powerful and effective technologies for their development. The analysis of modern technologies for designing intelligent computer systems shows that along with very impressive achievements, the following serious problems occur [1]–[3]:

- high requirements for the initial qualifications of users and developers. Artificial intelligence technologies are not focused on the wide range of developers and users of intelligent systems and, therefore, have not received mass distribution;
- modern information technologies are not oriented to a wide range of developers of applied computer systems;
- there is no general-unified solution to the problem of semantic compatibility of computer systems [4]. There are no approaches that allow integrating scientific and practical results in the field of artificial intelligence, which generates a high degree of duplication of results and a lot of non-unified formats for representation of data, models, methods, tools, and platforms;
- lack of powerful tools for designing intelligent computer systems, including intelligent training subsystems, subsystems for collective design of computer systems and their components, subsystems for verification and analysis of computer systems, subsystems for component design of computer systems;
- long terms of development of intelligent computer systems and high level of complexity of their maintenance and extension;
- the degree of dependence of artificial intelligence technologies on the platforms on which they are implemented is high, which leads to significant changes in technologies when transitioning to new platforms;
- the degree of dependence of artificial intelligence technologies on subject domains in which these technologies are used is high;
- there is a high degree of dependence of intelligent computer systems and their components on each other; the lack of their automatic synchronization. The absence of self-sufficiency of systems and components, their ability to operate separately from each other without loss of expediency of their use;
- increase in the time to solve the problem with the expansion of the functionality of the problem solver and with the expansion of the knowledge base of the system [5];
- lack of methods for designing intelligent computer systems. Updating computer systems often boils down to the development of various kinds of "patches", which eliminate not causes of the identified disadvantages of updated computer systems but only some consequences of these causes;
- poor adaptability of modern computers to the effective implementation of even existing knowledge representation models and models for solving problems that are difficult to formalize, which requires the development of fundamentally new computers [6];
- there is no single approach to the allocation of reusable components and the formation of libraries of such components, which leads to a high complexity of reuse and integration of previously developed components in new computer systems.
- there is a variety of semantically equivalent implementations of problem-solving models, duplication of knowledge base and user interface components that differ not in the essence of these components but in the form of representation of the processed information;

To solve these problems, it is necessary to implement a comprehensive technology for designing intelligent computer systems, which includes the following components:

- a model of an intelligent computer system [7];
- a *library of reusable components* and corresponding *tools to support component design of intelligent computer systems*;
- an intelligent integrated automation system for the collective design of intelligent computer systems, including subsystems for editing, debugging, performance evaluation, and visualization of developed components, as well as a simulation subsystem;
- methods of designing intelligent computer systems;
- an intelligent user interface;
- training subsystems for designing intelligent computer systems, including a subsystem for conducting a dialogue with the developer and the user;
- a subsystem for testing and verification of intelligent computer systems, including a subsystem for testing the compatibility of the developed system with other systems;
- an information security support subsystem for the intelligent computer system.

The key component of the technology for intelligent systems design is a component design that is represented as a *library of reusable components* and the corresponding *tools for supporting component design of intelligent computer systems*. With its help, it is possible to effectively implement the typical subsystems to support the design of intelligent computer systems.

II. Analysis of existing approaches to solving the problem

The problem is the lack of accessibility and integration in artificial intelligence technologies, which have high initial qualification requirements, lack a unified semantic compatibility solution, and have a high degree of dependency on platforms, subject areas, and components, leading to long development times, high maintenance costs, and difficulties in reusing and integrating previously developed components in new systems.

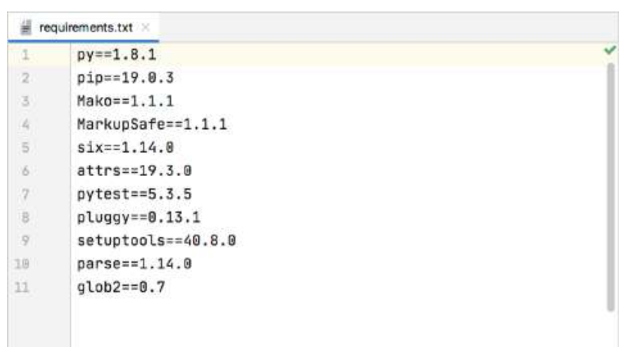
Existing approaches to solving the problem include component libraries and package managers of programming languages and operating systems, as well as separate systems and platforms with built-in components and means for saving created components.

The components of the library may be implemented in different programming languages (which leads to the fact that for each programming language different libraries are developed with their own solutions to various common situations), and may be located in different places, which leads to the fact that the library needs a tool to find components and install them.

Modern package managers such as *npm*, *pip*, *papt*, *maven*, *poetry* and others have the advantage that they are able to resolve conflicts when installing dependent components, but they do not take into account the semantics of components, but only install components by the [8] identifier. Libraries of such components are only a repository of components, without taking into account the purpose of components, their advantages and disadvantages, areas of application, the hierarchy of components and other information necessary for the intellectualization of component design of computer systems. Searching for components in *component libraries* corresponding to these package managers is reduced to searching by component identifier. Modern package managers are only "installers" without automatic integration of components into the system. Also a significant disadvantage of the modern approach is the platform dependency of components. Modern component libraries are oriented only to a certain programming language, operating system or platform.

The ***pip*** package manager is a package management system that is used to install packages from the Python Package Index, which is some library of such packages. Pip is often installed with Python. The pip package manager is used only for the Python programming language. It has many functions for working with packages:

- installation of a package;
- installation of a package of a specialized version;
- deletion of a package;
- reinstallation of a package;
- display of installed packages;
- search for packages;
- verification of package dependencies;
- creation of a configuration file with a list of installed packages and their versions;
- installation of a set of packages from a configuration file.


requirements.txt

```
py==1.8.1
pip==19.0.3
Mako==1.1.1
MarkupSafe==1.1.1
six==1.14.0
attrs==19.3.0
pytest==5.3.5
pluggy==0.13.1
setuptools==40.8.0
parse==1.14.0
glob2==0.7
```

Figure 1. pip configuration file

The pip package manager works well with dependencies, displays unsuccessfully installed packages, and also displays information about the required package version in case of conflict with another package. An example of a pip package configuration file is shown in Figure 1.

Another example of a package manager is *npm*. npm is a package manager for the javaScript language. The npm package manager has a component library (npm Registry) and a command-line user interface. The source code for the package manager and related npm tools can be found at https://github.com/npm. The most commonly used npm commands are:

- initializing the project (creating the package.json file);
- install all packages from the package.json file;
- install a package by name;
- deleting a package by name;
- check for obsolete packages;
- display help;
- view installed packages;
- search for packages;
- update packages.

The component approach to the design of computer systems can be implemented within various languages, platforms, and applications. Let us consider some of them.

The ontology implemented in ***OWL*** (Web Ontology Language) is a set of declarative statements about the entities of the dictionary of a subject domain (discussed in more detail in [9]). OWL assumes the concept of an "open world", according to which the applicability of subject domain descriptions placed in a specific physical document is not limited only to the scope of this document — the contents of the ontology can be used and supplemented by other documents adding new facts about the same entities or describing another subject domain in terms of this one. The "openness of the world" is achieved by adding a URI to each element of the ontology, which makes it possible to understand the ontology described in OWL as part of a universal unified knowledge.

The ***IACPaaS platform*** is designed to support the development, management, and remote use of applied and instrumental multi-agent cloud services (primarily intelligent ones) and their components for various subject domains [10].

The IACPaaS platform supports:

- the basic technology for the development of applied and specialized instrumental (intelligent) services using the basic instrumental services of the platform that support this technology;
- a variety of specialized technologies for the development of applied and specialized instrumental (intelligent) services, using specialized platform tool services that support these technologies.

The IACPaaS platform also does not contain means for a unified representation of the components of intelligent computer systems and means for their specification and automatic integration.

Based on the analysis carried out, it can be said that at the current state of development of information technologies, <u>there is no</u> comprehensive library of reusable semantically compatible components of computer systems and corresponding component management tools. Thus, it is proposed to implement a library and an appropriate component management tool that will implement seamless integration of components, ensure semantic compatibility of systems and their components, and significantly simplify the design of new systems and their components.

III. Proposed approach

Within this article, it is proposed to take the OSTIS Technology [11] as a basis, the principles of which make it possible to implement a library of semantically compatible components of intelligent computer systems and, accordingly, provide the ability to quickly create knowledge-driven systems using ready-made compatible components.

The systems developed on the basis of the OSTIS Technology are called *ostis-systems*. The *OSTIS Technology* is based on a universal method of <u>semantic</u> representation (encoding) of information in the memory of intelligent computer systems, called an *SC-code*. Texts of the *SC-code* (sc-texts) are unified semantic networks with a basic set-theoretic interpretation, which allows solving the problem of compatibility of various knowledge types. The elements of such semantic networks are called *sc-elements* (*sc-nodes* and *sc-connectors*, which, in turn, depending on orientation, can be *sc-arcs* or *sc-edges*). The *Alphabet of the SC-code* consists of five main elements, on the basis of which SC-code constructions of any complexity are built, including more specific types of sc-elements (for example, new concepts). The memory that stores SC-code constructions is called semantic memory, or *sc-memory*.

Within this article, fragments of structured texts in the SCn code [12] will often be used, which are simultaneously fragments of the source texts of the knowledge base, understandable to both human and machine. This allows making the text more structured and formalized, while maintaining its readability. The symbol ":=" in such texts indicates alternative (synonymous) names of the described entity, revealing in more detail certain of its features.

The basis of the knowledge base within the *OSTIS Technology* is a hierarchical system of subject domains and ontologies.

In order to solve the problems that have arisen in the design of intelligent systems and libraries of their reusable components, it is necessary to adhere to the general principles of the technology for intelligent computer systems design, as well as meet the following requirements:

- ensuring <u>compatibility</u> (integrability) of components of intelligent computer systems based on the unifying representation of these components;
- clear <u>separation</u> of the process of developing formal descriptions of intelligent computer systems and the

process of their implementation according to this description;
- clear separation of the development of a formal description for the designed intelligent system from the development of various options for the interpretation of such formal descriptions of the systems;
- availability of an ontology for component design of intelligent computer systems, including (1) a description of component design methods, (2) a model of a *library of reusable components*, (3) a model of a *specification of reusable components*, (4) a complete *classification of reusable components*, (5) a description of means for interaction of the developed intelligent computer system with *libraries of reusable components*;
- availability of *libraries of reusable components of intelligent computer systems*, including component specifications;
- availability of means for interaction of the developed intelligent computer system with libraries of reusable components for installation of any types of components and their management in the created system. The installation of a component means not only its transportation to the system (copying sc-elements and/or downloading component files) but also the subsequent execution of auxiliary actions so that the component can operate in the system being created.

Based on this, in order to solve the problems set within this article, it is proposed to develop the following system of subject domains and corresponding ontologies:

- Subject domain of reusable ostis-systems components
- Subject domain of a library of reusable ostis-systems components
- Subject domain of the manager of reusable ostis-systems components

IV. Concept of reusable component of ostis-systems

The *Subject domain of reusable ostis-systems components* describes the concept of a reusable component, the classification of components, and their general specification. This subject domain allows creating new and specifying existing components to add them to the library.

As a *reusable ostis-systems component*, a component of some ostis-system that can be used within another ostis-system is understood (see [13]). This is a component of the ostis-system that can be used in other ostis-systems (***child ostis-systems***) and contains all those and only those sc-elements that are necessary for the functioning of the component in the child ostis-system. In other words, it is a component of some ***maternal ostis-system***, which can be used in some child ostis-system. To include a reusable component in some system, it must be installed in this system, that is, all the sc-elements of the component should be copied into it and, if necessary, auxiliary files, such as the source or compiled component files. *Reusable components* must have a unified specification and hierarchy to support compatibility with other components. The compatibility of *reusable components* leads the system to a new quality, to an additional extension of the set of problems to be solved when integrating components.

reusable ostis-systems component
:= [typical ostis-systems component]
:= [reused ostis-systems component]
:= [reusable OSTIS component]
:= [ostis-systems ip-component]
:= *frequently used sc-identifier**: [reusable component]
⊂ *ostis-system component*
⊂ *sc-structure*

The requirements for *reusable ostis-systems components* inherit the common requirements for the design of software components and also include the following ones [14]:

- there is a technical possibility to embed a reusable component into a child ostis-system;
- a reusable component should perform its functions in the most general way, so that the range of possible systems in which it can be embedded is the widest;
- compatibility of a reusable component: the component should strive to increase the level of negotiability of ostis-systems in which it is embedded and be able to be automatically integrated into other systems;
- self-sufficiency of components, that is, their ability to operate separately from other components without losing the appropriateness of their use.

In the *Subject domain of the library of reusable ostis-systems components*, the most common concepts and principles are described, which are valid for any *library of reusable components*. This subject domain allows building many libraries, each of which will be semantically compatible with any other built according to the proposed principles. Such libraries store components and their specifications for use in child ostis-systems. An example of a specification of a reusable ostis-systems component is shown in Figure 2.

Versions for the full contents of the *Subject domain of reusable ostis-systems components* and the *Subject domain of the library of reusable ostis-systems components* are represented in the work [15].

The ***manager of reusable ostis-systems components*** is the main means of supporting component design of intelligent computer systems built by the *OSTIS Technology* ([16]). It allows installing reusable components in ostis-systems and controlling them. The *Subject domain*

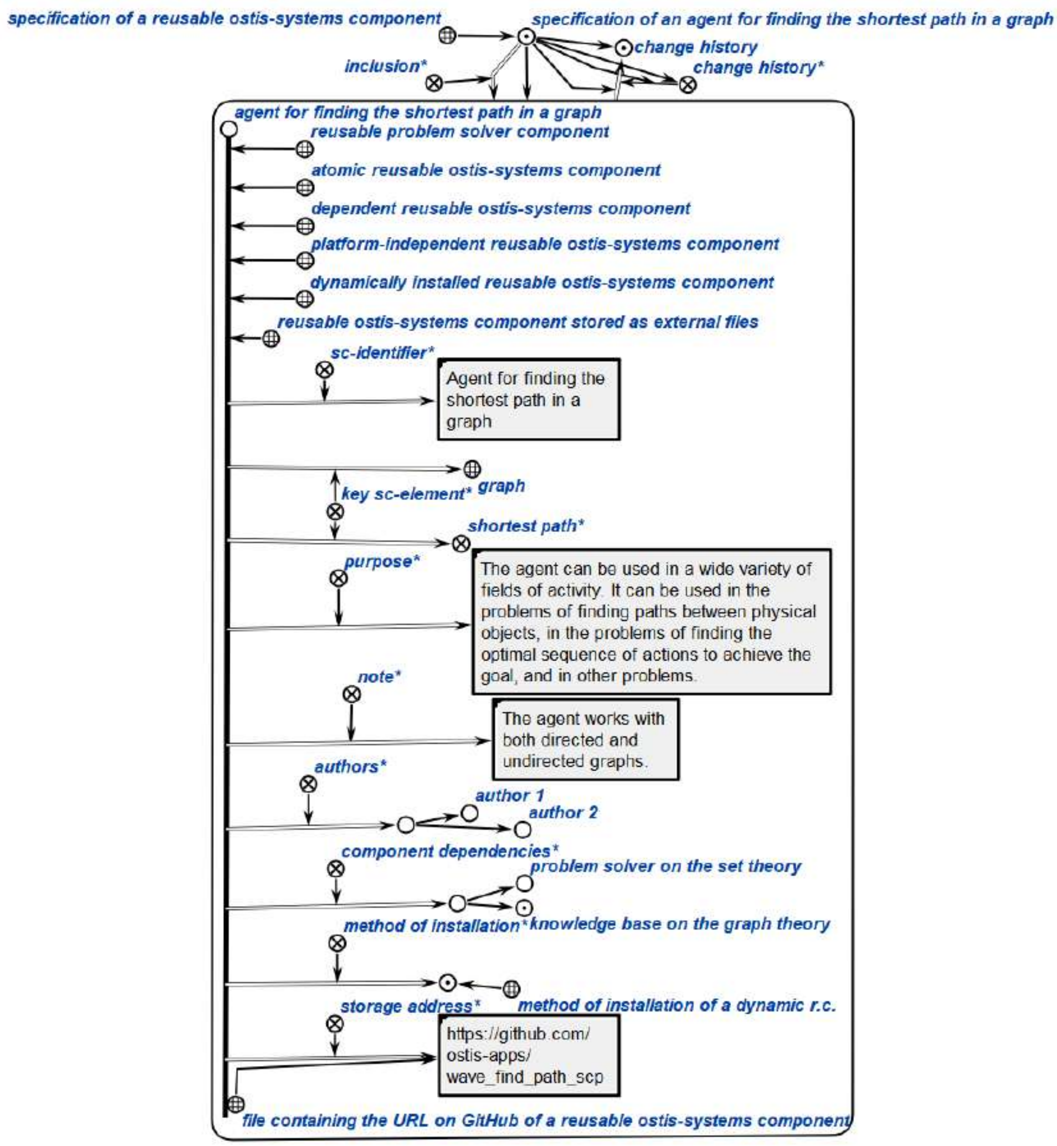


Figure 2. An example of a specification of a reusable ostis-systems component

of the manager of reusable ostis-systems components contains the full specification for the manager of ostis-systems components, the requirements for the component manager, its functionality, the specification of the implementation option for the manager of ostis-systems components, including the sc-model of the knowledge base, the problem solver, and the interface.

V. Architecture of component manager and library of reusable components of ostis-systems

To install reusable components of ostis-systems it is necessary to have a special subsystem in the system: a manager of reusable components of ostis-systems. The component manager interacts with the user and with the library of reusable components. To clarify the specifics of such interaction, diagrams are developed and SC-constructions necessary to initiate actions when working with the component manager are depicted.

Fig. 3 shows the entity-relationship diagram for the component manager, describing the current state of the component manager functionality.

The diagrams (Fig. 3 and Fig. 4) use the following concepts:

- of entities
  - Developer — a developer (of their local system);
  - OSTIS Metasystem Developer;
  - Component manager — manager of reusable components of ostis-systems;
  - Library of components;
  - System — user's system;

- Reusable component — Reusable component of ostis-system;
- (Reusable) component specification — Specification of reusable component of ostis-system;
- Storage (GitHub) — A repository of components and specifications, such as GitHub;
- Other library — a third-party library of reusable components.

- relationship
  - update;
  - use;
  - subsystem;
  - search;
  - connect;
  - store;
  - link;
  - installation.
- attributes
  - OSTIS Metasystem library — OSTIS Metasystem library of reusable components of ostis-system;
  - OSTIS Metasystem manager — OSTIS Metasystem manager of reusable components of ostis-system;
  - search arguments — search arguments for reusable components
    - Author — the author of the component;
    - Class — the class of the component;
    - Identifier — the name of the component;
    - Explanation — explanation of the component.

Component Library is a library of reusable components of ostis-systems, which is a subsystem of them. The library's *knowledge base* is a repository of reusable component specifications, and the library also provides an interface to visualise and manage component specifications of the user's system.

Component Manager — is a reusable component manager for ostis systems that is a subsystem for installing, downloading, and tracking components and their specifications for both the user's system and other systems that store reusable components.

The entity-relationship diagram for the component manager from the point of view of the ostis-system user on the example of the OSTIS Metasystem Library (Fig. 3) contains the following information.

The developer uses some ostis-system, a subsystem of which is a reusable component manager and optionally a library of reusable components. The developer can update a reusable component from the OSTIS Metasystem Library using the OSTIS Metasystem Reusable Component Manager. The developer can use the **component manager** to search for components in the OSTIS Metasystem and third-party libraries known to the manager by criteria such as component author, class, identifier, and component explanation fragment. The developer can connect to other component libraries. The developer can also install the found components into his system.

The entity-relationship diagram for a component library depicts the main relationships between a system, in this case the OSTIS Metasystem, and its subsystems (*manager* and *library*) in terms of the storage of components and their specifications.(Fig. 4). The diagram contains the following information.

OSTIS Metasystem has a subsystem in the form of a library and a component manager. The OSTIS Metasystem library stores many specifications of reusable components. Since all components are stored via GitHub, the manager uses the links provided in the component specifications to access them. The component specifications have a link to the repository that stores the component itself.

Updating reusable OSTIS Metasystem components in the OSTIS Metasystem Library is done through the OSTIS Metasystem Manager and the GitHub repository. The manager allows you to select the required versions of components and install the corresponding component specifications in the OSTIS Metasystem Library. According to these specifications, users using the OSTIS Metasystem will be able to learn about components and install them in their systems.

A component repository such as GitHub has many repositories, each of which can store any number of components and their specifications for installation on other systems using the reusable component manager.

VI. Reusable components installation process

Let's consider the functions of the manager of reusable components of ostis-systems.

reusable ostis-system components manager
:= [component manager]
⇒ *functions**:
{• *reusable component installation*
⇒ *partitioning**:
{• *scnitem*
scnitem component download
• *setting component dependencies*
• *translating component scs files into system sc-memory*
}
• *search for specifications of reusable components*
• *downloading specifications of reusable components*
}

In general, component installation consists of the following steps:

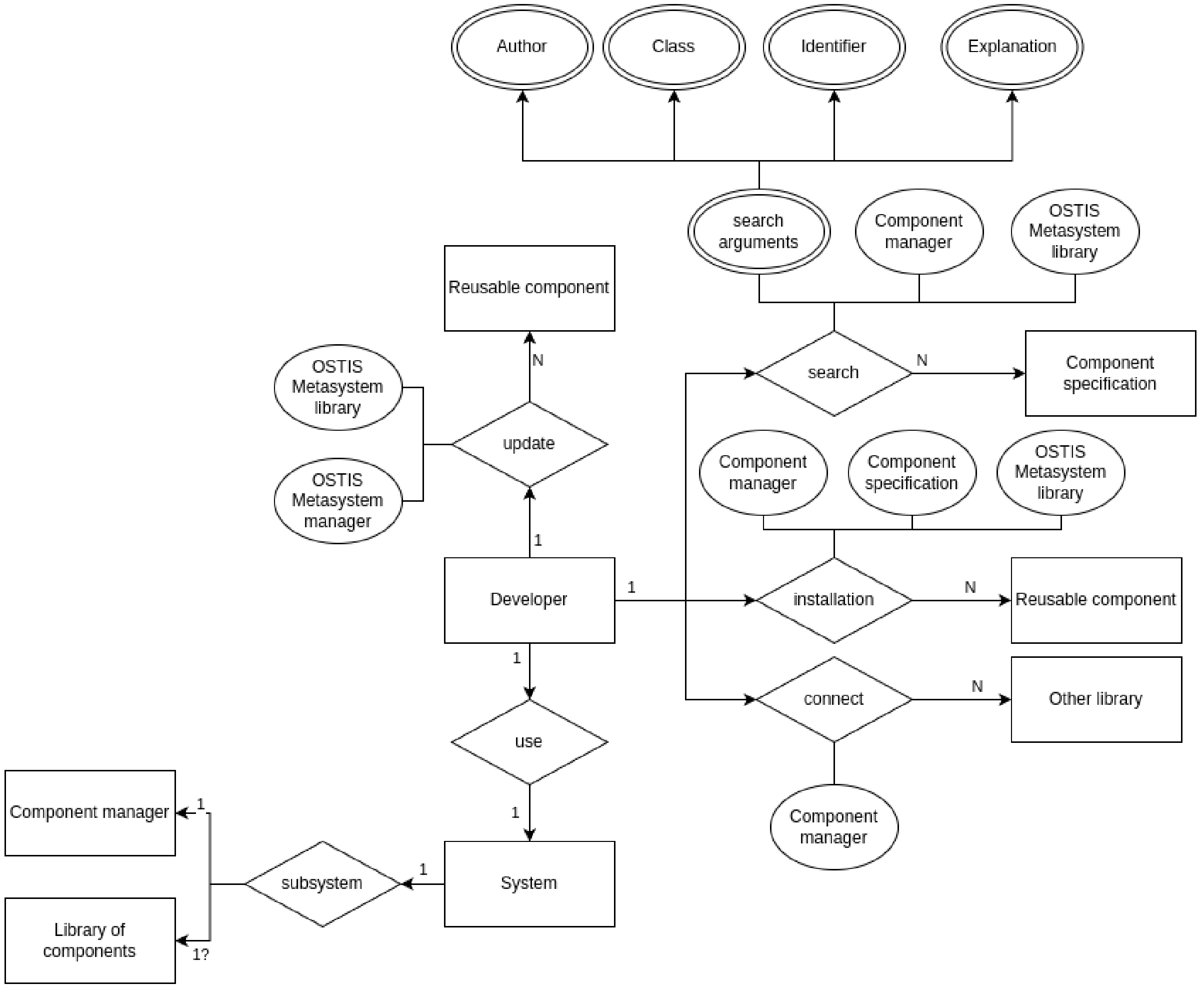


Figure 3. Entity-relationship diagram for a component library

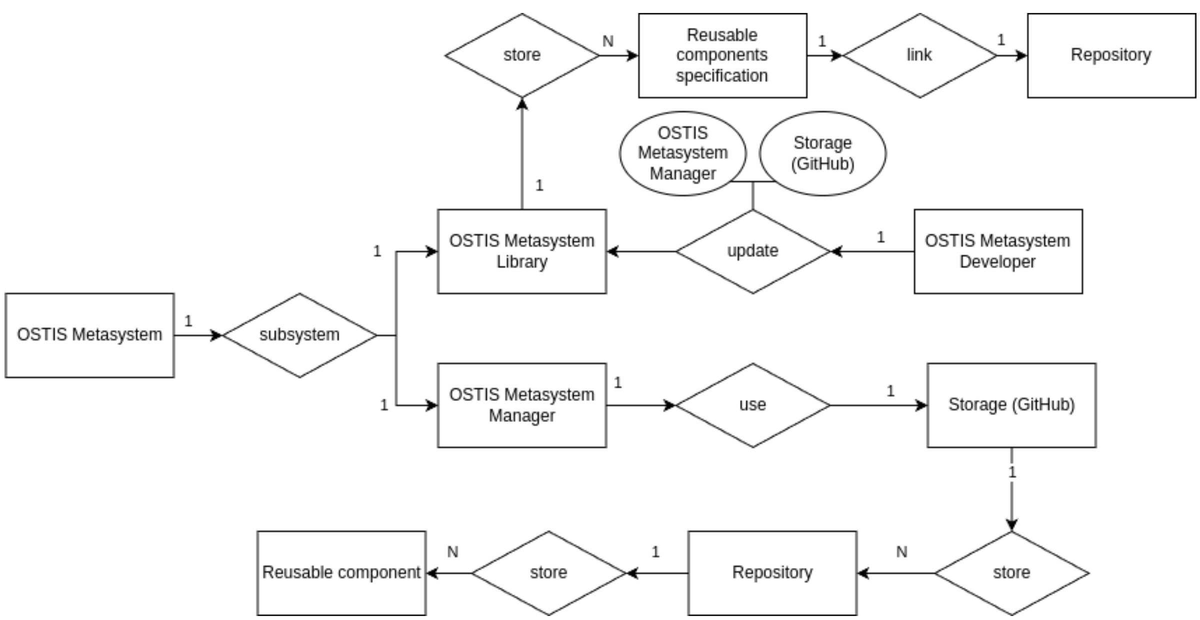


Figure 4. Entity-relationship diagram for a component library

- initiation of the agent to install all component specifications described in the knowledge base;
- to initiate the agent to search for the required component specifications in the knowledge base;
- initiating the agent to install the selected components.

After the user has initiated the component specification agent, the component manager will search the component specifications for references to the component specification repository. The specification file is called *specifcation.scs* and is stored in the folder with the reusable component itself. If the component manager was able to locate this file, it will load the file into sc memory. The component specification may include:

- identifier of the component;
- classes to which the component belongs;
- indicating the author of the component;
- indicating a note for the component;
- specifying how the component is installed;
- specifying the location (link) where the component is stored.

After the specifications are set, the user can search for components or install them.

The design of initiating the action of searching for component specifications is shown in Fig. 5.

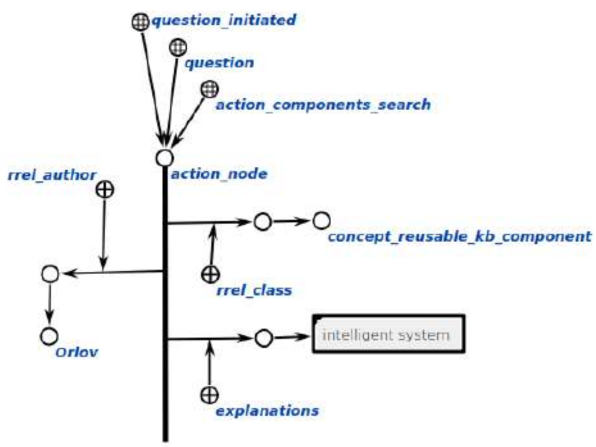


Figure 5. Example of calling the reusable components specification search agent

Three parameters are possible for the agent to search for component specifications: class, author, note. According to the above example, the manager will search for specifications of components created by *Orlov M.K.*, belonging to the class *multiple-used knowledge base component* and having the substring "intelligent system" in the note.

For the agent to find all components known to the system, then *class of reusable ostis-system components* must be passed as a parameter, and then the agent will find all specifications of reusable components stored in the system.

In order to install a component, you need to pass it as a parameter when calling the component installation agent. The agent will find the required component and its semantic neighbourhood that specifies the storage location of the component and how to install it, then the agent will install the reusable component in the ostis-system.

The design of the component installation agent initiation is shown in Fig 6:

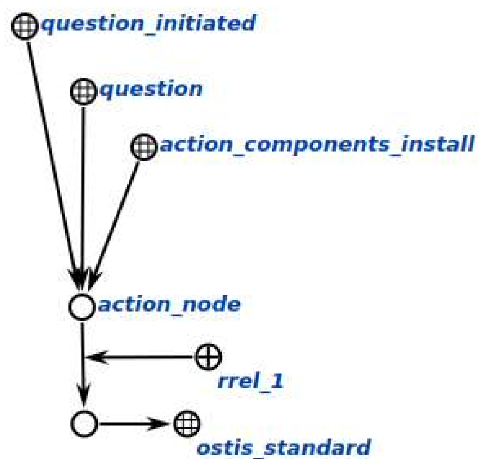


Figure 6. Example of calling the reusable components installation agent

Thus, the component manager and reusable component library allow systems to create and design intelligent systems based on off-the-shelf solutions, thus enhancing system interoperability and simplifying system development.

VII. Specification of ostis-system generation

Component-based design of computer systems means not only extending the functionality of a system already created in some form, but also creating an entire system "from scratch".

For the generation of ostis-systems the manager of reusable components of ostis-systems is used, which provides the possibility to assemble the system from the components available from the libraries of reusable components of ostis-systems.

The following typical sequence of user actions is used to generate ostis-systems.

generation of ostis-systems
:= [creation of ostis systems]
⇒ *generalised sequence of user actions**:
⟨• *search for ostis platform*
• *installing the ostis platform*
• *search for generic subsystems*

- *installing generic subsystems*
- *search for reusable components of ostis-systems*
- *installing reusable components*
- *configure ostis system*
  := [configuration of the ostis system]

⟩

Whether installing reusable components into an already created system or creating a system from scratch, constructs are created in the ostis-system knowledge base to denote which components are installed into the system. Figure 7 shows an example of a construct specifying which components are installed in an ostis system.

Finding and installing an ostis-platform is necessary because different ostis-platforms may be suitable for different classes of tasks and components to be installed in the generated system.

By installing generic subsystems, the functionality of the ostis-system being generated can be greatly expanded. The OSTIS Metasystem Library contains many subsystems often used in other ostis-systems. Typical subsystems include, for example, the subsystem for collective design of ostis-systems, natural language interface, training subsystem, security subsystem and others.

Thanks to the extensive search functionality of reusable ostis-system components, it is possible to search for any components according to various criteria and combinations thereof.

Customising an ostis-system implies setting parameters to specify the peculiarities of the system operation, as well as specifying which users are administrators, developers, experts, and users of the created ostis-system.

In addition to user actions when creating an ostis-system, the ostis-system generation subsystem also registers the created ostis-system in the OSTIS Metasystem. Thus, the OSTIS Metasystem is able to monitor and update the status of the components of this system.

VIII. User interface of component manager and library of reusable components of ostis-systems

The multi-component manager for ostis-systems has a console interface. The component manager is connected to sc-memory as a dynamic component, so it does not require a restart, and you can immediately see the installed components in a running system.

Let's look at the commands with which you can use the component manager. Each command calls the corresponding agent. Agent-based architecture allows you to implement any user interfaces for the component manager of ostis-systems. Any variant of the component manager user interface creates sc-constructs in sc-memory that are needed to invoke the corresponding sc-agent.

action. Set the specifications of reusable ostis-system components
⇒ *agent**:
  [ScComponentManagerInitAgent]
⇒ *command to initiate an action**:
  [*components init*]

action. Find specifications for reusable ostis components
⇒ *agent**:
  [ScComponentManagerSearchAgent]
⇒ *command to initiate action**:
  [*components search*]
⇒ *possible flags**:
- [*author*]
- [*class*]
- [*explanation*]

When using the search command with the author flag, you must list the system identifiers of the sc nodes that denote the authors of the reusable component. The class flag is used to pass the class name to the component manager to search for components belonging to this class. The explanation flag is used to specify a natural-language fragment that is a substring of the component's explanation. If multiple search flags are listed, components that satisfy all search criteria simultaneously will be found. If you use the component search command without flags, all components whose specifications are downloaded will be found.

action. Install reusable ostis components
⇒ *agent**:
  [ScComponentManagerInstallAgent]
⇒ *command to initiate action**:
  [*components install*]
⇒ *flags**:
- [*idtf*]

The component install command requires the mandatory idtf flag, which the component manager uses to search by system ID for the components to be installed and create the necessary construct to call the component install agent.

The interface of the reusable component library is graphical (Fig. 8. It displays the components that are in the library and provides access to search, browse, and install them.

IX. Example of usage of library of reusable components of ostis-systems

As an example of a library of reusable components of ostis-systems for consideration of an example of work, let's take the OSTIS Metasystem Library.

Installation of the OSTIS Metasystem is performed using the following command sequence.

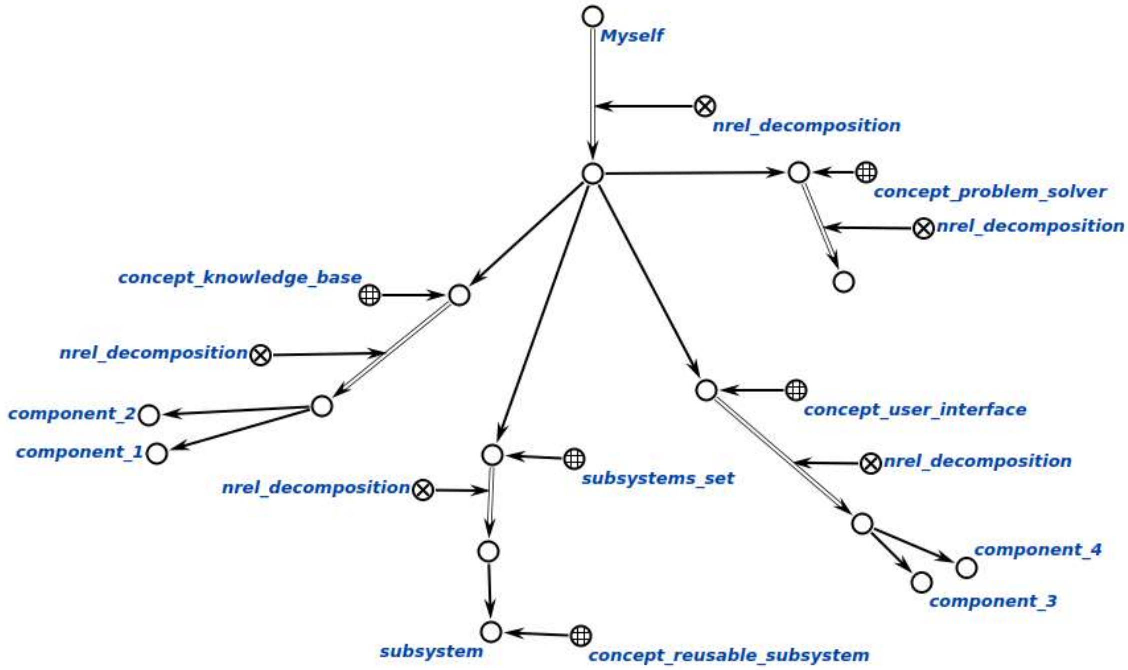


Figure 7. Formalisation of installed components to the ostis-system

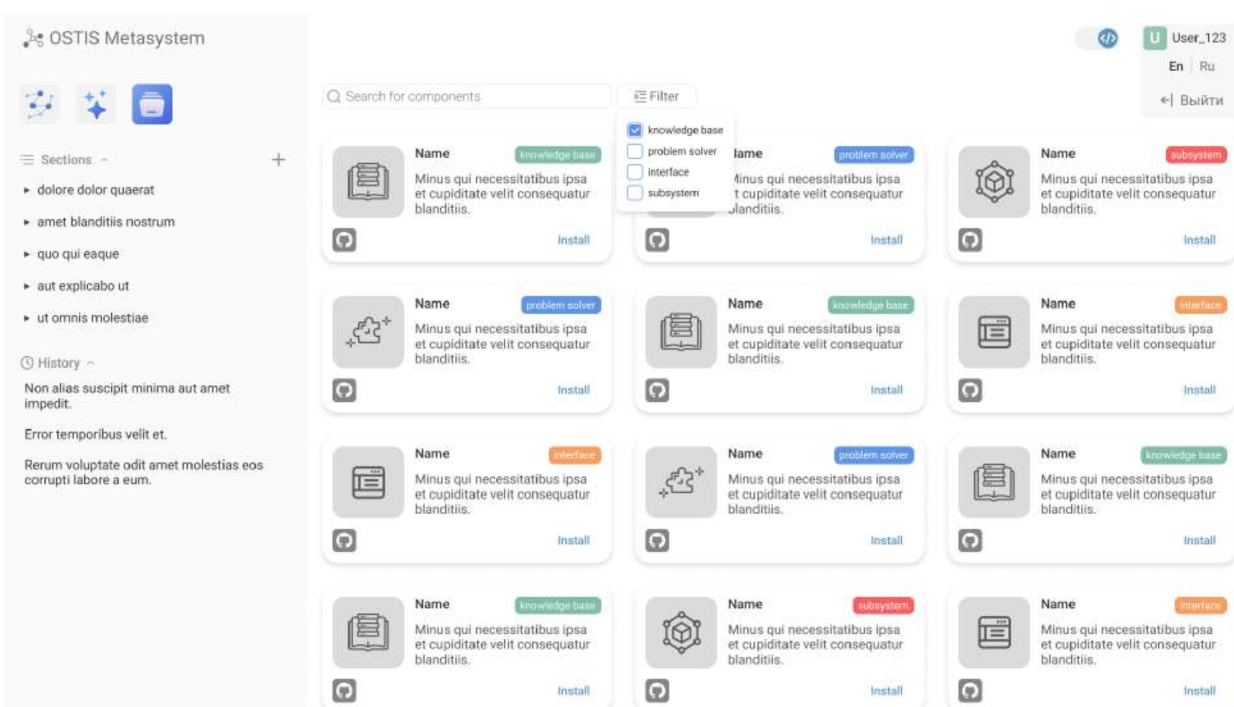


Figure 8. User interface of a library of reusable components of ostis-system

OSTIS Metasystem
⇒ *installation stages**:
⟨• *Repository cloning*
⇒ *terminal command**:
[*git clone https://github.com/ostis-ai/ostis-metasystem*]
• *Change dorectory to the project root*
⇒ *terminal command**:
[*cd ostis-metasystem*]
• *OSTIS Metasystem installtion*
⇒ *terminal command**:
[*./scripts/install_metasystem.sh*]
• *Run sc-component-manager*
⇒ *terminal command**:
[*./scripts/run_sc_component_manager.sh*]
⟩
⇒ *component installation procedure**:
⟨• *Install all the reusable components specifications*
⇒ *command**:
[*components init*]
• *Reusable components searching*
⇒ *command**:
[*components search*]
• *Installation of reusable component*
⇒ *command**:
[*components install −−idtf <identifier>*]
⟩
⇒ *usage examples**:
• *Creation of the kernel*
• *Extension of kernel functionality*

Creation of the core
⇒ *stages**:
⟨• *Install all the reusable components specifications*
⇒ *sc-agent**:
[ScComponentManagerInitAgent]
⇒ *commmand to call an agent**:
[*components init*]
⇒ *result**:
[All the reusable components specifications from OSTIS Library are installed.]
• *Search reusable user interface components*
⇒ *sc-agent**:
[ScComponentManagerSearchAgent]
⇒ *commmand to call an agent**:
[*components search −−class concept_reusable_interface_component*]
⇒ *result**:
[All the reusable user interface components are found.]
• *Install sc-models of user interface interpreter*
⇒ *sc-agent**:
[ScComponentManagerInstallAgent]
⇒ *commmand to call an agent**:
[*components install −−idtf sc_web*]
⇒ *result**:
[sc-web is intalled by specification.]
⇒ *note**:
[If you start the web interface after this step, only the start page will load, because the knowledge base is currently empty.]
• *Searching for Knowledge Base components*
⇒ *sc-agent**:
[ScComponentManagerSearchAgent]
⇒ *commmand to call an agent**:
[*components search −−class concept_reusable_kb_component*]
⇒ *result**:
[Received all components for which their specification states that they are Knowledge Base components.]
• *OSTIS Standard installation*
⇒ *agent**:
[ScComponentManagerInstallAgent]
⇒ *commmand to call an agent**:
[*components install −−idtf ostis_standard*]
⇒ *result**:
[OSTIS Standard is installed]
⇒ *note**:
[If you now start the web interface again, this step in the web interface will display a page with part of the standard to navigate to.]
⟩
⇒ *result**:
[The kernel is installed.]

Extension of kernel functionality
⇒ *stages**:
⟨• *Search for all available Knowledge Base components in the library*
⇒ *sc-agent**:
[ScComponentManagerSearchAgent]

- ⇒ *commmand to call an agent**:
  [*components search −−class concept_reusable_kb_component*]
- ⇒ *result**:
  [Found all Knowledge Base components whose specifications have been installed]

- *Installing the Knowledge Base component*
  - ⇒ *sc-agent**:
    [ScComponentManagerInstallAgent]
  - ⇒ *commmand to call an agent**:
    [*components install −−idtf part_polygons*]
  - ⇒ *result**:
    [A Knowledge Base component in the form of subject domain of polygons was established.]
  - ⇒ *note**:
    [After performing this step, we can find the concept "multiple" in the web interface and browse its semantic neighbourhood. But it is worth noting that the subject domain of triangles, which is a private subject domain of polygons, is empty.]
- *Installing the Knowledge Base component*
  - ⇒ *sc-agent**:
    [ScComponentManagerInstallAgent]
  - ⇒ *commmand to call an agent**:
    [*components install −−idtf part_triangles*]
  - ⇒ *result**:
    [The Knowledge Base component is installed in the form ofof subject domain of triangle]
  - ⇒ *note**:
    [After performing this step, we can find the concept "triangle" in the web interface and browse its semantic neighbourhood. It is worth noting that the subject domain of triangles, which is a private subject domain of polygons, is fully described and compatible with the subject domain of polygons.]
- *Creating two sets of triangles*
  - ⇒ *note**:
    [At this step it is necessary to find the class "triangle" in the web-interface, create two sets of triangles and add elements to them. It is necessary to specify that the sets and their elements belong to the class "triangle".]
  - ⇒ *example**:
    [*triangles_1 = {ABC, CDE, XYZ}, triangles_2 = {MNK, CDE, XYZ}*]
  - ⇒ *note**:
    [After performing this step, you can check that no operations on sets can be performed now. This can be verified by right-clicking on the node "triangles_1".]
- *Search for all available problem solver components in the library*
  - ⇒ *sc-agent**:
    [ScComponentManagerSearchAgent]
  - ⇒ *commmand to call an agent**:
    [*components search −−class concept_reusable_ps_component*]
  - ⇒ *result**:
    [Found all components of the problem solver whose specifications are installed.]
- *Installing the components of the problem solver*
  - ⇒ *sc-agent**:
    [ScComponentManagerInstallAgent]
  - ⇒ *commmand to call an agent**:
    [*components install −−idtf agent_of_finding_intersection_of_sets*]
  - ⇒ *result**:
    [A problem solver component for finding the intersection of two sets is established.]
  - ⇒ *note**:
    [After this step, you can check that you can now perform an operation on sets. In the web interface, search for the concept "installed components" and select the node of the desired agent *agent_of_finding_intersection_of_sets*) and run the set intersection agent using the example of two previously created triangle sets. The intersection of the two sets will be found. But it should be noted that this way of launching the agent is long and inconvenient.]
- *Search for all available interface components in the library*

- ⇒ *sc-agent**:
  [ScComponentManagerSearchAgent]
- ⇒ *commmand to call an agent**:
  [*components search −−class concept_reusable_interface_component*]
- ⇒ *result**:
  [Found all interface components whose specifications have been downloaded.]

- *Installing the user interface component*
  - ⇒ *sc-agent**:
    [ScComponentManagerInstallAgent]
  - ⇒ *commmand to call an agent**:
    [*components install −−idtf menu_of_agent_of_finding_intersection_of_sets*]
  - ⇒ *result**:
    [Installed an interface component for an agent to find the intersection of two sets.]
  - ⇒ *note**:
    [After this step, the intersection finder can be invoked using a button in the interface, which is much faster and more convenient than the first method. This can be checked by calling the agent to find the intersection of two sets using the example of triangle sets (*triangles_1* and *triangles_2*).]
- *Setting a logical formula component*
  - ⇒ *sc-agent**:
    [ScComponentManagerInstallAgent]
  - ⇒ *commmand to call an agent**:
    [*components install −−idtf lr_about_isosceles_triangle*]
  - ⇒ *result**:
    [Established a component with a logical formula for determining whether a triangle is isosceles or not.]
  - ⇒ *note**:
    [If you go to the web interface after performing this step, create the necessary fragment for the geometry logic formula parcels and try to run the logic output, it fails because the logic output component is missing.]
- *Setting the logic inference component*
  - ⇒ *sc-agent**:
    [ScComponentManagerInstallAgent]
  - ⇒ *commmand to call an agent**:
    [*components install −−idtf scl_machine*]
  - ⇒ *result**:
    [Logic inference machine is installed.]
  - ⇒ *note**:
    [After performing this step go to the web-interface, create the necessary fragment to send a logical formula on geometry and try to run the logical output, then the formula will generate the necessary fragment of the Knowledge Base. However, this is still not very convenient.]
- *Installing the user interface component*
  - ⇒ *sc-agent**:
    [ScComponentManagerInstallAgent]
  - ⇒ *result**:
    [Interface component for logic output component installed]
  - ⇒ *note**:
    [After performing this step in the web interface after creating the necessary fragment to send the formula on geometry, you can easily call the logical output agent through the interface component.]

⟩

⇒ *result**:
[The functionality of the system is extended. A system capable of logical inference and finding intersection of sets is obtained. The system has interface components corresponding to these agents. The Knowledge Base on geometrical figures (polygons and triangles) is also obtained.]

X. Conclusion

The component approach is key in the technology of designing intelligent computer systems. At the same time, the technology of component design is closely related to the other components of the technology of designing intelligent computer systems and ensures their compatibility, producing a powerful synergetic effect when using the entire complex of private technologies for designing intelligent systems. The most important principle in the implementation of the component approach is the semantic compatibility of reusable components, which minimizes the participation of programmers in the creation of new computer systems and the improvement of existing ones.

To implement the component approach, in the article, a library of reusable compatible components of intelligent computer systems based on the *OSTIS Technology* is proposed, classification and specification of reusable ostis-systems components is introduced, a component manager model is proposed that allows ostis-systems to interact with libraries of reusable components and manage components in the system, the architecture of the ecosystem of intelligent computer systems is considered from the point of view of using a library of reusable components.

At the moment the manager of reusable components of ostis-systems with console user interface and the library of reusable components of ostis-systems with graphical user interface have been implemented. The subject areas necessary for the implementation of component design have been implemented, and diagrams showing the details of the use and operation of the component manager and the component library have been implemented.

The results obtained will improve the design efficiency of intelligent systems and automation tools for the development of such systems, as well as provide an opportunity not only for the developer but also for the intelligent system to automatically supplement the system with new knowledge and skills.

References

[1] K. Yaghoobirafi and A. Farahani, “An approach for semantic interoperability in autonomic distributed intelligent systems,” *Journal of Software: Evolution and Process*, vol. 34, 02 2022.

[2] Natalia N. Skeeter, Natalia V. Ketko, Aleksey B. Simonov, Aleksey G. Gagarin, Irina Tislenkova, “Artificial intelligence: Problems and prospects of development,” *Artificial Intelligence: Anthropogenic Nature vs. Social Origin*, 2020.

[3] Olena Yara, Anatoliy Brazheyev, Liudmyla Golovko, Liudmyla Golovko, Viktoriia Bashkatova, “Legal regulation of the use of artificial intelligence: Problems and development prospects,” *European Journal of Sustainable Development*, 2021.

[4] J. Waters, B. J. Powers, and M. G. Ceruti, “Global interoperability using semantics, standards, science and technology (gis3t),” *Computer Standards & Interfaces*, vol. 31, no. 6, pp. 1158–1166, 2009.

[5] M. Wooldridge, *An introduction to multiagent systems*, 2nd ed. Chichester : J. Wiley, 2009.

[6] X. Shi, Z. Zheng, Y. Zhou, H. Jin, L. He, B. Liu, and Q.-S. Hua, “Graph processing on GPUs,” *ACM Computing Surveys*, vol. 50, no. 6, pp. 1–35, Nov. 2018. [Online]. Available: https://doi.org/10.1145/3128571

[7] P. Lopes de Lopes de Souza, W. Lopes de Lopes de Souza, and R. R. Ciferri, “Semantic interoperability in the internet of things: A systematic literature review,” in *ITNG 2022 19th International Conference on Information Technology-New Generations*, S. Latifi, Ed. Cham: Springer International Publishing, 2022, pp. 333–340.

[8] Blähser, Jannik and Göller, Tim and Böhmer, Matthias, “Thine — approach for a fault tolerant distributed packet manager based on hypercore protocol,” in *2021 IEEE 45th Annual Computers, Software, and Applications Conference (COMPSAC)*, 2021, pp. 1778–1782.

[9] V. Torres da Silva, J. S. dos Santos, R. Thiago, E. Soares, and L. Guerreiro Azevedo, “OWL ontology evolution: understanding and unifying the complex changes,” *The Knowledge Engineering Review*, vol. 37, p. e10, 2022.

[10] V. Gribova,L. Fedorischev, P. Moskalenko, V. Timchenko, “Interaction of cloud services with external software and its implementation on the IACPaaS platform,” pp. 1–11, 2021.

[11] Vladimir Golenkov and Natalia Guliakina and Daniil Shunkevich, *Open technology of ontological design, production and operation of semantically compatible hybrid intelligent computer systems*, V. Golenkov, Ed. Minsk: Bestprint [Bestprint], 2021.

[12] (2022, September) IMS.ostis Metasystem. [Online]. Available: https://ims.ostis.net

[13] Shunkevich D.V., Davydenko I.T., Koronchik D.N., Zukov I.I., Parkalov A.V., “Support tools knowledge-based systems component design,” *Open semantic technologies for intelligent systems*, pp. 79–88, 2015. [Online]. Available: http://proc.ostis.net/proc/Proceedings%20OSTIS-2015.pdf

[14] G. Sellitto, E. Iannone, Z. Codabux, V. Lenarduzzi, A. De Lucia, F. Palomba, and F. Ferrucci, “Toward understanding the impact of refactoring on program comprehension,” in *29th International Conference on Software Analysis, Evolution, and Reengineering (SANER)*, 2022, pp. 1–12.

[15] M. K. Orlov, “Comprehensive library of reusable semantically compatible components of next-generation intelligent computer systems,” in *Open semantic technologies for intelligent systems*, ser. Iss. 6. Minsk : BSUIR, 2022, pp. 261–272.

[16] ——, “Control tools for reusable components of intelligent computer systems of a new generation,” in *Open semantic technologies for intelligent systems*, ser. Iss. 7. Minsk : BSUIR, 2023, pp. 261–272.

ТЕКУЩЕЕ СОСТОЯНИЕ СРЕДСТВ АВТОМАТИЗАЦИИ КОМПОНЕНТНОГО ПРОЕКТИРОВАНИЯ OSTIS-СИСТЕМ

Орлов М. К., Макаренко А. И., Петрочук К. Д.

В работе рассматривается подход к проектированию интеллектуальных систем, ориентированный на использование совместимых многократно используемых компонентов, что существенно сокращает трудоемкость разработки таких систем. Ключевым средством поддержки компонентного проектирования интеллектуальных компьютерных систем является предложенный в работе менеджер многократно используемых компонентов.

Received 03.03.2024

A Formal Model of Shared Semantic Memory for Next-Generation Intelligent Systems

Nikita Zotov
Belarusian State University of Informatics and Radioelectronics
Minsk, Belarus
Email: nikita.zotov.belarus@gmail.com

***Abstract*—This paper discusses in detail the formal model of semantic memory for intelligent systems, its structure, its elements, correspondences between them, rules and algorithms. The implementation based on this model is described, quantitative indicators of its efficiency are given.**

***Keywords*—shared memory, semantic memory, graph database, sc-memory, formal model of semantic memory, mathematical model of semantic memory, ostis-platform, intelligent system, unified knowledge representation, parallel information processing, semantic networks storage**

I. Introduction

Earlier in the works [1]–[3], devoted to the description of the *Software platform for intelligent systems* developed according to the principles of the *OSTIS Technology* [4] (*Software platform of ostis-systems*) — a software emulator of the future associative semantic computer [5], the software implementation of the general semantic memory (sc-memory) was considered [3], and the implementation of its programming interface was described in detail [2].

The peculiarity of previous works is that they focused not on the peculiarities of component implementation, but on approaches to describing and documenting such complex systems as the Software platform of ostis-systems [6], [7]. In this paper, the main task is to formally describe how a shared semantic memory can be realized in intelligent systems, i. e. to describe its model.

Therefore, the purpose of the current work and the novelty of this paper is to describe a formal model of shared semantic memory used in ostis-systems, allowing:

- to store information of any kind in a unified semantically compatible form;
- to efficiently process this information using a specified set of operations in both single-threaded and multi-threaded modes;
- to process this information using agents that react to events in this memory,

that is, allowing [8]:

- to represent knowledge in the form of semantically compatible knowledge bases of intelligent systems [9], [10];
- to create various types of interpreters of models of intellectual systems components, including interpreters for intellectual systems problem solvers;
- create libraries of reusable platform-dependent components to implement other components of ostis-platforms [11].

The relevance of the work is conditioned by the current state in the field of development of intelligent systems [12], namely:

- labor intensity of development of intellectual systems of various kinds;
- complexity of tasks solved in intelligent systems;
- complexity of integration of various components of intelligent systems;
- increase in the volume of processed information;
- growth of requirements to the speed of information processing;
- insufficient performance of modern open computing systems.

Further in this paper we will consider and describe the necessary and sufficient model of semantic memory for its implementation and use in solving the above problems [13]–[15].

II. Principles of implementation of ostis-platforms

All intelligent systems developed according to the principles of the *OSTIS Technology* are commonly referred to as *ostis-systems*. Each *ostis-system* consists of its own sc-model, including a knowledge base, problem solver and user interface, and an *ostis-platform* on which this sc-model is interpreted [4], [16]. Any sc-model of an ostis-system is a logical-semantic model of that system described in *SC-code*, the language of universal information encoding. By *ostis-platform* is meant a hardware-implemented computer or a software emulator of this computer for interpretation of sc-models of ostis-systems [9].

There can be a great variety of *ostis-platform* implementations on which *ostis-systems* can be interpreted, but each of such *ostis-platforms* should provide the following basic principles [1], [16], [17]:

- the development of an *ostis-system* should be reduced to the development of its sc-model (i.e. the description of the model in *SC-code* [18]) without modification of the chosen *ostis-platform* and regardless of the means by which this ostis-platform is developed;
- transfer of the sc-model of an *ostis-system* from one ostis-platform to another is limited to loading this sc-model into the memory of the ostis-platform without loss of functionality of this ostis-system;
- addition of new information should be reduced to the "gluing" of its signs with signs of existing information with further verification of the obtained information;
- processing of information in the system should be provided by changing the configuration of links between entities in this information by means of asynchronous-parallel interaction of sc-agents reacting to the occurrence of events in the shared memory.

Therefore, all ostis-systems interpreted on ostis-platforms, unlike modern computer systems, have the following features:

- unlike modern computers, where data is represented as lines of binary code, the data stored inside the memory of ostis-systems are graph constructions written in SC-code (sc-constructions), due to which [19]:
  - any kind of knowledge is written in the same form using a minimal set of syntactically indivisible elements — the minimal alphabet of a language, which can always be augmented with new syntactic elements by specifying additional syntactically distinguishable classes for the elements of that language's alphabet;
  - synonymy of entity signs is forbidden, since each entity appears in the semantic network once;
  - the meaning of information is represented by explicitly specifying the relationships between entities in that information;
  - the "gluing" of information is reduced to the "gluing" of the entity signs in that information;
  - the processing of this information does not require various means of syntactic and semantic analysis.
- information processing is based on the principles of graphodynamic and agent-oriented models, so that:
  - the process of information processing is reduced not only to changing the state of elements of the semantic network, but (!) also to changing the configuration of links between them;
  - information processing is represented and stored as a semantic network;
  - it is possible to describe and solve problems of any information complexity;
  - it is possible to realize and use any existing types and models of information processing (procedural, neural network, frame, logical, production, etc.);
  - information can be processed in parallel, i.e. different methods of problem solving can be applied simultaneously.

In other words, unlike traditional computer systems, any ostis system must be oriented towards:

- independence from the implementation of a particular ostis-platform;
- storage of information in a unified and semantically compatible form (in SC-code [18]);
- event-oriented and parallel processing of this information.

The principles of ostis-systems, first of all, should be provided by a concrete implementation of the ostis-platform. Within each *ostis-platform* there must be implemented:

- the shared semantic memory that allows [20]:
  - to store information constructions belonging to *SC-code* (sc-texts);
  - to store information constructions that do not belong to *SC-code* (images, text files, audio and video files, etc.);
  - to store subscriptions to occurrences of events in memory;
  - to initiate agents after these events appear in memory;
  - to use a programming interface to work with *SC-code* and *non-SC-code* information constructions, including:
    - operations to create, search, modify, and delete these constructions in the memory;
    - operations for subscribing to the occurrence of events in the memory;
    - operations for controlling and synchronizing processes in the memory;
    - programming interface for creating platform-dependent agents;
- the interpreter of the *SCP* asynchronous-parallel programming language, which is a platform-independent programming interface that implements platform-independent operations on the shared semantic memory.

The shared semantic memory that allows for fulfillment of all of the above mentioned tasks is called sc-memory, and the interpreter of the basic language of asynchronous-parallel programming *SCP* — *scp-interpreter*. In the context of this paper only the sc-memory model is considered. The *scp-interpreter* model and its implementation were considered in [21].

All listed principles of ostis-systems are provided in the first (basic) Software variant of the implementation of the ostis-platform — *Software platform of ostis-systems*. Drawing an analogy with modern developments in the field, the Software platform of ostis-systems can be considered as a graph database management system, for example, *Neo4J* [22]. However, unlike existing database management systems, the Software platform of ostis-systems acts as an interpreter of sc-models of semantically compatible ostis-systems. Therefore, the Software platform ostis-systems should also be considered as a software alternative for modern von Neumann computers. In general, the above mentioned features of the implemented Software platform of ostis-systems are also true for all ostis-platforms regardless of the means by which they are implemented.

An efficient implementation of sc-memory must fulfill the following requirements:

- high performance — minimizing the time spent on operations of adding, searching, modifying and deleting stored information;
- minimum memory and disk space usage for storing sc-texts;
- scalability — the ability to add computing power as the load increases without difficulty.

The following sections of this paper will discuss a possible sc-memory model and its implementation.

III. Proposed sc-memory model for next-generation intelligent systems

So, as mentioned above, in general *sc-memory* performs the following tasks:

- the task of storing sc-constructs and information constructs external to the *SC-code* (ostis-system files),
- the task of managing events and processes working on these constructs,
- the task of managing access to sc-constructions, including tasks of:
  - creating and deleting sc-elements (sc-nodes, sc-connectors, etc.);
  - searching sc-constructions by specified sc-elements;
  - getting sc-element characteristics (type, incident sc-elements);
  - adding content to sc nodes;
  - retrieving ostis-system files by known contents;
  - retrieving content from ostis-system files.

In this regard, all entities in sc-memory are:

- elements of sc-constructions:
  - sc-nodes, ostis-system files,
  - or sc-connectors (sc-arcs, sc-ribs) between sc-nodes, ostis-system files;
- elements of information constructions that are external to the SC-code:
  - string content;
  - or binary content;
- subscriptions to events in this sc-memory, that is, subscriptions to the occurrence of items in it;
- active or inactive processes in this sc-memory,
- the synchronization objects of the processes in this sc-memory,
- operations performed by processes in this sc-memory.

Thus, *the model of sc-memory* can be defined quintuple

$$SM = \langle SS,\ FS,\ RS,\ PS,\ PI\rangle,$$

where

- SS — the sc-element storage model, which is a structure of information about sc-elements,
- FS — the external information construction storage model (file memory), which is a structure of information about external information constructions,
- RS — the event subscription storage model, which is a structure of information about event subscriptions in sc-memory,
- PS — the process storage model that represents the structure of process information in sc-memory,
- PI — the set of operations over sc-memory, i.e., the programming interface of sc-memory.

A. Model of storage of sc-elements in sc-memory

The model of sc-element storage in sc-memory can be represented as

$$SS = \langle S,\ M_e, n_{s_{le}},\ n_{s_{\max}},\ n_{s_{lv}}, n_{s_{lr}},\ m_s,\ {m^n}_{s_{lv}},\ {m^n}_{s_{lr}}, SSPI\rangle,$$

where

- $S = \langle s_1,\ s_2,\ ...,\ s_i,\ ...,\ s_n\rangle,\ i = \overline{1,\ n}$ — sequence of allocated cell segments in sc-memory of fixed size n;
- $s_i = \{\langle e_{i1},\ e_{i2},\ ...,\ e_{ij},\ ...,\ e_{im}\rangle,\ n_{e_{le}},\ n_{e_{lr}},\ m_e\}$, $j = \overline{1,\ m}$, — i-th segment of fixed size m, consisting of cells (elements) of sc-memory e_{ij} of fixed size k,
- $n_{e_{le}} \in (\bar{\mathbb{N}} \cup \{0\})$ — the index of the last engaged cell in the segment s_i,
- $n_{e_{lr}} \in (\bar{\mathbb{N}} \cup \{0\})$ — the index of the last released cell in the segment s_i,
- $m_e \in M$ — the object that synchronizes access to $n_{e_{le}}$ and n_{e_r};
- $M_e \subseteq E \times M$ — a dynamic oriented set of pairs of sc-elements and corresponding synchronization objects;
- $n_{s_{le}} \in (\bar{\mathbb{N}} \cup \{0\})$ — the index of the last engaged segment in sc-memory ($n_{s_{le}} = n$),
- $n_{s_{\max}} \in (\bar{\mathbb{N}} \cup \{0\})$ — the maximum number of segments in sc-memory;

- $n_{s_{lv}} \in (\bar{\mathbb{N}} \cup \{0\})$ — the index of the last vacant segment in sc-memory;
- $n_{s_{lr}} \in (\bar{\mathbb{N}} \cup \{0\})$ — the index of the last released segment in sc-memory;
- $m_s \in M$ — the object that synchronizes access to $S, n_{s_{le}}$ and $n_{s_{max}}$;
- $m^n{}_{s_{lv}} \in M$ — the object that synchronizes access to $n_{s_{lv}}$;
- $m^n{}_{s_{lr}} \in M$ — the object that synchronizes access to $n_{s_{lr}}$;
- $SSPI = \{engage,\ release\}$ — internal programming interface of sc-elements storage in sc-memory.

All allocated segments S may be free $S_f \subseteq S$ or engaged $S_e \subseteq S$. The set of free sc-memory segments S_f includes the set of vacant segments $S_v \subseteq S_f$ and the set of released segments $S_r \subseteq S_f$.

Cells in sc-memory segments E may be free $E_f \subseteq E$ and engaged $E_e \subseteq E$. The set of free sc-memory cells $E_f = \{e^{ij}{}_f | \ e^{ij}{}_f \in s^i{}_f,\ s^i{}_f \in S_f\}$ includes the set of vacant cells $E_v \subseteq E_f$ and the set of released cells $E_r \subseteq E_f$.

Consequently, the following statements hold for all sc-memory segments and cells in them:

- $E_f \cup E_e = E,\ E_f \cap E_e = \varnothing$,
- $E_v \cup E_r = E_f,\ E_v \cap E_r = \varnothing$,
- $S_f \cup S_e = S,\ S_f \cap S_e = \varnothing$,
- $S_v \cup S_r = S_f,\ S_v \cap S_r \subseteq S_f$.

For the sets of engaged and released cells, the corresponding transitions can be defined in the form of:

- operation of allocating a $engage : E_f \to E_e$, which changes the state of the cell from "released" to "engaged":

$$engage() = \begin{cases} e_{ij} \in E_e, & \text{if } \exists e_{ij} \in s_i, E_f \wedge n \leq n_{s_{\max}}, \\ \text{Error}, & \text{if } n > n_{s_{\max}} \vee |E_f| = 0; \end{cases}$$

- operation of releasing a cell in sc-memory segment $release : E_e \to E_f$, which changes the state of the cell from "engaged" to "released":

$$release(e_{ij}) = \begin{cases} e_{ij} \in E_f, & \text{if } e_{ij} \in E_e, \\ \text{Error}, & \text{if } e_{ij} \notin E_e. \end{cases}$$

The algorithm of the cell engaging operation in sc-memory segment (*engage*) can be described as follows:

- Step 1: Try to find any vacant segment s_i in the set S_v:
  - If such a segment exists, go to step 2.
  - If no such segment exists, skip to step 3.
- Step 2: Engage a new cell in the found vacant segment s_i:
  - Increase $n_{e_{le}}$ in segment s_i by 1.
  - Occupy the cell e_{ij} with index $n_{e_{le}}$ in segment s_i.
  - Return the address of the engaged cell e_{ij} and terminate.
- Step 3: Attempt to get a new segment from set S:
  - If the number of engaged segments $n_{s_{le}}$ is less than the maximum $n_{s_{max}}$, create a new segment s_i (set $n_{e_{le}}$ as 0) and add it to S.
    - Increase $n_{e_{le}}$ in segment s_i by 1.
    - Occupy the cell e_{ij} with index $n_{e_{le}}$ in segment s_i.
    - Return the address of the engaged cell e_{ij} and terminate.
  - If the maximum number of segments $n_{s_{max}}$ is reached, go to step 4.
- Step 4: Try to get the released segment s_i from the set S_r:
  - If there is no such segment, report an error and terminate.
  - If such a segment exists, proceed to step 5.
- Step 5: Engage a new cell in the found released segment s_i:
  - Get the last released cell e_{ij} by its number in segment n_{e_r}.
  - Occupy the cell e_{ij} with index n_{e_r} in segment s_i.
  - Update n_{e_r} for the next released cell.
  - Return the address of the engaged cell e_{ij} and terminate.

The algorithm of the cell releasing operation in sc-memory segment (*release*) can be described as follows:

- Step 1: Verify the correctness of the given address of cell e_{ij}:
  - If the cell address does not exist in sc-memory, terminate with an error.
  - If the cell address exists in sc-memory, proceed to step 2.
- Step 2: Using the cell address, determine the corresponding segment e_i of the cell in sc-memory and proceed to step 3.
- Step 3: Release the cell e_{ij}:
  - Update the information about the cell e_{ij}, mark it as released.
  - Update the number of the last released cell in segment n_{e_r}.
  - Go to step 4.
- Step 4: Update the information about the released segments.
  - If the released cell was the first released cell in the segment, update the information of the last released segment in sc-memory $n_{s_{lr}}$.
  - Go to step 5.
- Step 5: Terminate.

The described algorithms may include synchronization mechanisms to ensure data integrity during multi-threaded sc-memory accesses. All synchronization operations in these algorithms can be performed using appropriate synchronization objects m_s, m_e, ${m^n}_{s_{lv}}$, ${m^n}_{s_{lr}}$. The model of the synchronization object will be discussed later.

Basically, the advantages of the described algorithms are due to the advantages of the cell engaging algorithm in sc-memory, which are as follows:

- The cell engaging algorithm in sc-memory tries to find a vacant segment before creating a new one. If there is no vacant segment, it tries to create a new one if the maximum number of segments has not been reached. This approach avoids wasting memory on creating unnecessary segments.
- By updating the number of the last released cell in a segment, the algorithm keeps track of which cells are available for reuse. This tracking ensures that the engaging process is fast and does not require searching the entire memory for a released cell.

To analyze the complexity of these algorithms, let us consider their main characteristics:

- In the algorithm of cell engaging operation in sc-memory:
  - finding a vacant segment and engaging a cell in it requires traversing many segments and checking their status, which can be accomplished in a time proportional to the number of segments;
  - creating a new segment and selecting a cell in it requires a fixed number of operations independent of the data size;
  - finding a released segment also requires traversing multiple segments.
- In the algorithm of cell releasing operation in sc-memory:
  - checking the correctness of the cell address and determining the appropriate segment can be accomplished in a time proportional to the number of segments, or faster if an efficient data structure is used to store the segments;
  - releasing a cell and updating segment information requires a fixed number of operations.

The complexity of the algorithms depends on the data structures used and how they are processed. Assuming that multiple segments are processed efficiently, for example using internal lists in released segments and released segment cells, the underlying complexity of the algorithms will be determined by the number of operations required to process each step. Thus:

- The algorithm of cell engaging operation in sc-memory can have a complexity from $O(1)$ to $O(n)$, where n is the number of segments, depending on whether a free segment is found or a new segment needs to be engaged;
- The algorithm of cell releasing operation in sc-memory basically has a complexity of $O(1)$, since most operations are performed in a fixed amount of time, except for address correctness checking, which may require $O(n)$ in the worst case.

So, an sc-element storage is a set of cells, each of which can store some sc-element (be engaged) or can be empty (free):

$$(\forall e \in E : (e \in E_e) \vee (e \in E_f)).$$

Each cell $e_{ij} \in E, e_{ij} \in s_i, s_i \in S$ has a unique internal address $a = \langle i, j \rangle \in A$. That is, the following statement always holds:

$$(\forall e_{ij} \in E, \exists! a \in A : (e_{ij} \in s_i) \wedge (s_i \in S) \wedge (a = \langle i, j \rangle)).$$

Each cell stores either an sc-node N or an sc-connector C:

$$(\forall e \in E : (e \in N) \vee (e \in EC)),$$
$$N \cup C = E, N \cap C = \varnothing.$$

It is assumed that if a cell stores some sc-element, it stores information characterizing this sc-element. Each cell in sc-memory $e \in E$ can be represented as a tuple:

$$e = \langle FI, EI, CI \rangle,$$

where

- FI — characteristics of the stored sc-element, including the type of sc-element and its states,
- EI — information about sc-elements incident with the given sc-element,
- CI — information about the number of incoming and outgoing sc-connectors for a given sc-element.

The characteristics of an sc-element FI can be represented as:

$$FI = \langle t, s \rangle,$$

where

- $t \in T$ is the syntactic type of a given sc-element (e.g., sc-node, sc-connector, ostis-system file, etc.),
- $s \in ES$ is the state of the cell (e.g., "engaged", "free", etc.).

$$T = T_n \cup T_c,$$
$$T_n = \{node,\ file\}, T_c = \{connector,\ arc,\ edge\},$$

where

- T is the set of all possible syntactic types of sc-elements;
- T_n is set of all possible syntactic types of sc-nodes, $node, file$ — actual sc-node label and ostis-system file label, respectively;

- T_c is the set of all possible syntactic types of sc-connectors, $connector, arc, edge$ — the actual sc-connector label, sc-arc label, and sc-edge label, respectively;

$$ES = \{engaged, free\},$$

where

- ES is the set of all possible cell states;
- $engaged$ — the cell is "engaged";
- $free$ — the cell is "free".

Information about incident sc-connectors EI can be represented as:

$$(\forall n \in N : (n \ni EI = \langle b_o,\ b_i \rangle)),$$

$$(\forall c \in C : (c \ni EI =$$

$$= \langle b_o,\ b_i,\ b,\ e,\ n_{bo},\ p_{bo},\ n_{bi},\ p_{bi},\ n_{eo},\ p_{eo},\ n_{ei},\ p_{ei} \rangle)),$$

where

- $b_o \in A$ is the sc-address of the initial sc-connector outgoing from the given sc element,
- $e_i \in A$ is the sc-address of the initial sc-connector incoming into the given sc-element,
- $b \in A$ is the sc-address of the initial sc-element of the sc-connector,
- $e \in A$ is the sc-address of the final sc-element of the sc-connector,
- $n_{bo} \in A$ is the sc-address of the next sc-connector outgoing from the initial sc-element,
- $p_{bo} \in A$ is the sc-address of the previous sc-connector outgoing from the initial sc-element,
- $n_{bi} \in A$ is the sc-address of the next sc-connector incoming into the initial sc-element,
- $p_{bi} \in A$ is the sc-address of the previous sc-connector incoming into the initial sc-element,
- $n_{eo} \in A$ is the sc-address of the next sc-connector outgoing from the final sc-element,
- $p_{eo} \in A$ is the sc-address of the previous sc-connector outgoing from the final sc-element,
- $n_{ei} \in A$ is the address of the next sc-connector incoming into the final sc-element,
- $p_{ei} \in A$ is the address of the previous sc-connector incoming into the final sc-element.

In this case, the following statements are true:

$$(\forall e \in E, \exists! b_o, b_i \in A : (Inc(e, b_o) \wedge Inc(e, b_i))),$$

$$(\forall\ e\ \in C,\ \exists!\ n_{bo},\ p_{bo},\ n_{bi},\ p_{bi}\ \in\ A\ :$$
$$((Inc(e,\ n_{bo}) \wedge Inc(e,\ p_{bo})) \wedge (Inc(e, n_{bi}) \wedge Inc(e,\ p_{bi})))),$$

$$(\forall\ e\ \in C,\ \exists!\ n_{eo},\ p_{eo},\ n_{ei},\ p_{ei}\ \in\ A\ :$$
$$((Inc(e,\ n_{eo}) \wedge Inc(e,\ p_{eo})) \wedge (Inc(e,\ n_{ei}) \wedge Inc(e,\ p_{ei})))),$$

where Inc is the binary relation of incidence of two sc-elements.

Information about the number of incoming and outgoing sc-connectors C can be represented as:

$$CI = \langle c_{in}, c_{out} \rangle,$$

where

- $c_{in} \in (\bar{\mathbb{N}} \cup \{0\})$ is the number of incoming sc-connectors in a given sc-element,
- $c_{out} \in (\bar{\mathbb{N}} \cup \{0\})$ is the number of outgoing sc-connectors from a given sc-element.

This information can be used to optimize the isomorphic search for sc-constructions over a given graph-template [2].

The model of storage of sc-elements in sc-memory provides:

- storage of sc-constructions, their sc-elements, characteristics and incident relations between them;
- ability to create, modify, search and delete sc-elements.

The advantages of this model are as follows:

- it provides efficient fragmentation and defragmentation of cells;
- algorithms for allocating and freeing a memory segment have asymptotic complexity from $O(1)$ to $O(n)$, where n is the number of segments that must be traversed to find a free segment.

B. Model of storage of external information constructions in sc-memory

The model of storage of external information constructions in sc-memory can be represented as

$$\begin{aligned} FS\ &= \langle CH, M_s, n_{ch_{le}}, n_{ch_{max}}, m_{ch}, tr, \\ &TSO,\ SOF,\ FSO,\ FSPI \rangle, \end{aligned}$$

where

- $CH = \langle ch_1,\ ch_2,\ ...,\ ch_i,\ ...,\ ch_n \rangle,\ i = \overline{1,\ n}$ is the sequence of dynamically allocated file segments in sc-memory of fixed size n;
- $ch_i = \{\langle\langle l_{s_{i1}}, s_{i1}\rangle, ..., \langle l_{s_{ij1}}, e_{ij}\rangle, ..., \langle l_{s_{im}}, e_{im}\rangle\rangle, n_{s_l},\ m_s\},\ j = \overline{1,\ m},$ — the i-th file segment of fixed size m, consisting of cells – pairs of string lengths $l_{s_{ij}}$ and strings themselves $s_{ij} \in STR$,
- $n_{s_l} \in (\bar{\mathbb{N}} \cup \{0\})$ — the index of the last engaged cell in the file segment ch_i,
- $m_s \in M$ — the object that synchronizes access to $n_{s_{le}}$;
- $M_s \subseteq CHS \times M$ — a dynamic oriented set of file and cell pairs and their corresponding synchronization objects;
- $n_{ch_{le}} \in (\bar{\mathbb{N}} \cup \{0\})$ — the index of the last engaged file segment in sc-memory ($n_{ch_{le}} =\ n$),

- $n_{ch_{max}} \in (\bar{\mathbb{N}} \cup \{0\})$ — the maximum number of file segments in sc-memory;
- $m_{ch} \in M$ — the object that synchronizes access to $CH, n_{ch_{le}} and n_{ch_{max}}$;
- $tr \subset TRM$ — rules (terms) for finding terms in strings;
- TSO — correspondence between string terms and file cell numbers with these sc-memory strings;
- SOF — correspondence between file cell numbers with strings in sc-memory and ostis-system files of which these strings are contents;
- FSO — correspondence between ostis-system files and file cell numbers with sc-memory strings, which are the contents of these strings;
- $FSPI = \{allocate,\ free,\ dump,\ load\}$ — internal programming interface of file storage in sc-memory.

Unlike sc-element storage, where the cell size is fixed and cells can be allocated in advance as some fixed sequence, this cannot be done in file storage, because cells can store strings of unknown length in advance. Therefore, the accounting of released cells, their fragmentation and defragmentation processes may be more complicated. In this connection, the problem of external fragmentation is not solved in this model, as it is solved in the sc-element storage model.

Each file cell $s_{ij} \in STR, s_{ij} \in ch_i, ch_i \in CH$ has some unique internal address $fa = \langle i, j \rangle \in FA \subset A$. That is, the following statement is always true:

$$(\forall s_{ij} \in STR, \exists! fa \in FA : (s_{ij} \in ch_i) \wedge (ch_i \in CH) \wedge$$
$$\wedge\ (fa = \langle i, j \rangle)).$$

The $allocate$ and $free$ operations can be defined for file segment cells. Their algorithms are quite simple, so they will not be considered.

File storage specifies operations to enable saving $dump$ and loading $load$ of all sc-memory.

This model is focused on the fact that any string can be partitioned into a set of terms, by which, using the TSO mapping, we can determine the indexes of strings that contain these terms [23], [24]. Then, by string indexes it is possible to obtain: from CH — the strings themselves, from SOF — the ostis-system files that contain these lines. Using the FSO mapping it is possible to find the string, which is contained by the given ostis-system file.

$$TSO = TRM \times FA,$$

$$SOF = FA \times FN, FSO = FN \times FA, FN \subset N.$$

The model of storage of external information constructions in sc-memory provides:

- storing of the contents of ostis-system files;
- setting the contents to a given ostis file;
- retrieving contents from a specified ostis file;
- retrieving ostis-files by their contents;
- obtaining ostis-system files by their content substring.

The advantages of this model are as follows:

- asymptotic complexity of adding new strings to the storage is $O(1)$ without taking into account the complexity of access time to file storage segments;
- asymptotic complexity of searching ostis-system files and their contents is $O(1)$ without taking into account the complexity of access time to segments and cells of the file storage and correspondences between ostis-system files and their contents.

C. Model of storage of subscriptions to events in sc-memory

The model of storage of subscription to events in sc-memory can be specified by the following tuple

$$RS = \langle V, m_v, RSPI \rangle,$$

where

- $V = \{v_1,\ v_2,\ ...,\ v_i,\ ...,\ v_n\}, i = \overline{1,\ n}$ is set of subscriptions to events in sc-memory of size n;
- $v_i = \langle e,\ t_v,\ a_v,\ m_v \rangle \in V$ is a subscription to an event in sc-memory;
- e is an sc-element (a cell) in sc-memory that is being "listened";
- $t_v \in T_v$ is a type of event in sc-memory;
- $ag_v \in AG$ is an agent subscribed to an event;
- $m_v \in M$ is an object that synchronizes access to subscription elements;
- $RSPI = \{subscribe, unsubscribe, notify\}$ — internal programming interface of storage of subscriptions to events in sc-memory.

All cells in sc-memory can E be listenable $E_l \subseteq E$ and non-listenable $E_{nl} \subseteq E$. For them, the following statements are true:

- $E_l \cup E_{nl} = E,\ E_n \cap E_{nl} = \varnothing,$
- $E_l \cap E_e = E_l,\ E_l \cap E_f = \varnothing,$
- $E_{nl} \cap E_e = E_{nl},\ E_l \cap E_f = \varnothing,$

$$T_v = \{aoc,\ aic,\ roc,\ ric,\ re,\ cc\},$$

where

- aoc is the event of adding an outgoing sc-connector from the listened sc-element;
- aic is the event of adding an incoming sc-connector to the listened sc-element;
- roc is the event of removing an outgoing sc-connector from the listened sc-element;
- ric is the event of removing an incoming sc-connector from the listened sc-element;
- re is the event of removing the listened sc-element;
- cc is the event of changing the content of the listened ostis-system file.

For the sets of listened and unlistened cells, the corresponding transitions can be defined in the form of:

- operation of creating a subscription to an event in sc-memory $subscribe : E \times T_v \times AG \rightarrow E_l$:

$$subscribe(e, t_v, ag_v) = \begin{cases} e_{ij} \in E_l, & \text{if } e_{ij} \in E_e \\ \text{Error}, & \text{if } e_{ij} \notin E_e; \end{cases}$$

- operation of removing a subscription to an event in sc-memory $unsubscribe : E_l \rightarrow E_{nl}$:

$$unsubscribe(e_{ij}) = \begin{cases} e_{ij} \in E_{nl}, & \text{if } e_{ij} \in E_e, \\ \text{Error}, & \text{if } e_{ij} \notin E_e. \end{cases}$$

In addition, the following operation can be defined for the set of listened events $notify : (E_l \times C) \cup E_l \rightarrow P_w$:

$$notify(e_l, c) = \begin{cases} p \in P_w & \text{if } e_l \in E_l, c \in C, Inc(e_l, c), \\ \text{Error}, & otherwise; \end{cases}$$

The notify operation can be used to notify (initiate) a process about a new event (creation of an outgoing arc c from sc-element e_l, deletion of element e_l, etc.).

The model of storage of subscriptions to events provides:

- storing of event subscriptions in sc-memory;
- ability to subscribe and unsubscribe to an event in sc-memory;
- ability to notify about an event in sc-memory.

D. Model of storage of processes in sc-memory

The model of storage of processes in sc-memory can be defined as

$$PS = \langle P_a, Q_{wp}, n_{map}, PS, PSF, PAG, PSPI \rangle,$$

where

- $P_a \subseteq P$ is the set of active processes in sc-memory;
- Q_{wp} is the queue of processes waiting to start in sc-memory;
- n_{map} is the the maximum possible number of active processes at a given time, $|Q_w p| \leq n_{map}$;
- PS is the mapping between active processes and sc-element storage segments;
- PSF is the mapping between active processes and file storage segments;
- PAG is the mapping between active processes and agents;
- $RSPI = \{activate, deactivate\}$ — internal programming interface of storage of sc-memory processes.

$$PS = (P_a \cup P_w) \times S,$$

$$PSF = (P_a \cup P_w) \times CH,$$

$$PAG = (P_a \cup P_w) \times AG.$$

The PS and PSF mappins are used to assign processes to segments of the sc-element storage and file storage. If there are enough free segments in the storage, each process is assigned separate segments in both storages.

All processes in sc-memory P can be waiting $P_w \subseteq P$, active $P_a \subseteq P$ or finished $P_f \subseteq P$.

In this case, each active and waiting process corresponds to an agent that executes it:

$$(\forall p \in P_a, \exists! ag \in AG : (\langle p, ag \rangle \in PAG)),$$

$$(\forall p \in P_w, \exists! ag \in AG : (\langle p, ag \rangle \in PAG)).$$

The following statements are true for all types of processes:

- $P_w \cup P_a \cup P_f = P, P_w \cap P_a \cap P_f = \varnothing$;
- $|P_a| \leq n_{map}$.

Transition between waiting and active processes can be defined as the function $activate : P_w \rightarrow P_a$:

$$activate(p_w) =$$

$$= \begin{cases} p_w \in P_a, & \begin{cases} ((\exists s_i \in S_f) \wedge (n_s \leq n_{s_{\max}})) \\ ((\exists ch_i \in CH) \wedge (n_{ch} \leq n_{ch_{\max}})), \end{cases} \\ \text{Error}, & \text{otherwise}; \end{cases}$$

Transition between active and finished processes can be defined as the function $deactivate : P_a \rightarrow P_f$:

$$deactivate(p_a) = \begin{cases} p_a \in P_f, & \text{if } , p_a \in P_a, \\ \text{Error}, & \text{otherwise}; \end{cases}$$

A process is considered to be finished if it is not active and not waiting:

$$((p \in P_f) \Leftrightarrow ((\neg p \in P_a) \wedge (\neg p \in P_w))).$$

The model of storage of sc-memory processes provides:

- efficient one-to-one allocation of writer-processes to sc-element storage and file storage segments;
- queuing new processes when the device's processing power is limited, and activating processes from the queue when some active process has finished its work.

E. Model of coordinated access (synchronization) of processes to sc-memory

Each synchronization object $m \in M$ can be represented as [25], [26], [27]:

$$m = \langle c_{ar}, f_{aw}, Q_{rw}, m_u \rangle,$$

where

- c_{ar} is the active reader count;
- f_{aw} is the flag that shows whether a reader is active at a given moment;
- Q_{rw} is the queue of readers and writers;
- m_u is the object used to synchronize access to elements of a given synchronization object.

The queue of readers and writers is a sequence of requests to acquire a particular resource:

$$Q_{rw} = \langle q_1, q_2, ..., q_j, ..., q_m \rangle.$$

Each request q_j includes a unique thread identifier id_j, a thread type (reader or writer) tt_j, and a condition variable that allows messages to be exchanged between processes (threads) cv_j:

$$q_j = \langle id_j, tt_j, cv_j \rangle.$$

This queue ensures that no thread is left hungry.

To coordinate access to data structures in sc-memory, mechanisms for acquiring and releasing resources for reader-threads P_r (hereinafter — readers) and writer-threads P_w (hereinafter — writers) are required.

$$P = P_r \cup P_w.$$

These mechanisms should include:

- a reader resource acquisition operation (*acquire_read*) that allows a reader-thread to acquire a synchronization object to start reading the resource and suspend the execution of all other writer-threads while there are readers in the reader-writer queue;
- a reader resource release operation (*release_read*) that allows a reader-thread to release the synchronization object after finishing reading the resource and notify all other writer-threads to execute if there are no active readers in the reader-writer queue after releasing the synchronization object;
- a writer resource acquisition operation (*acquire_write*) that allows a writer-thread to acquire a synchronization object to start modifying the resource and suspend execution of all other reader-threads and writer-threads while there is an active writer in the reader-writer queue;
- a writer resource release operation (*release_write*) that allows a writer-thread to release the synchronization object after the resource modification is complete and notify all other reader-threads and writer-threads to execute if there are no active readers in the reader-writer queue after the synchronization object is released;
- a reader multi-resource acquisition operation (*acquire_read_n*) that allows a reader-thread to acquire multiple synchronization objects for reading in the order necessary to prevent deadlocks;
- a reader multi-resource release operation (*release_read_n*) that allows a reader-thread to release multiple synchronization objects in the reverse order of acquisition;
- a writer multi-resource acquisition operation (*acquire_write_n*) that allows a writer-thread to acquire multiple synchronization objects to modify in the order necessary to prevent deadlocks;
- a writer multi-resource release operation (*release_write_n*) that allows a writer-thread to release multiple synchronization objects in the reverse order of acquisition.

Allocation of sc-memory to writers can be done segment-by-segment using a specialized table T_{ps}, which allows to determine whether a given vacant sc-memory segment has been acquired by another writer:

$$s : S \rightarrow S_v, \; T_{ps} \subseteq P_w \times S_v.$$

Let us recall that a free sc-memory segment can be either a segment with vacant cells or a segment with released cells. When allocating sc-memory, the first thing that is done is to search for vacant segments that are not used by other writers. If no such segments are found, new segments are allocated. If there is no available space in the sc-memory for new segments, writers can use segments from the list of engaged vacant segments.

To ensure coordinated read access to segments, each segment contains a unique synchronization object.

$$m_s : S \times M \rightarrow S_m,$$

$$m_{ch} : CH \times M \rightarrow CH_m.$$

In addition to segments, synchronization objects are also temporarily assigned to sc-memory cells and events to be registered in it. Synchronization objects of sc-memory cells can be stored in a specialized table T_{em}. These objects are used to synchronize access to the sc-element information contained in the sc-memory element:

$$T_{em} \subseteq A \times M.$$

These objects synchronize the subscription and unsubscription to events through a single table, as well as the initiation of the sc-agents themselves

$$v_i = \langle t_v^i, \; A_v^i, \; m_v^i \rangle, t_v^i \in T_v, A_v^i \subseteq A_v, m_v^i \in M,$$
$$v_m : V \times M \rightarrow V_m.$$

The model of synchronization of process access to sc-memory provides:

- parallel access to sc-memory, i.e. the possibility of parallel execution of actions in sc-memory without violating correctness of data structures in it;

- absence of deadlocks, races and hungry processes in sc-memory;
- fast parallel creation of sc-elements in sc-memory due to distribution of processes over sc-memory segments;
- fast non-blockable parallel search of sc-constructions provided that no other operations are performed on these sc-constructions.

F. Model of sc-memory programming interface

The model of sc-memory programming interface can be defined as follows

$$PI = N^T \times C^{E \times E \times T} \times F^{N \times L} \times$$
$$\times \{E^{E \times \{E \cup T\} \times E} \cup \ E^{\{E \cup T\} \times T \times E} \cup E^{E \times T \times \{E \cup T\}}\} \times$$
$$\times \, T^E \times \ F^L \times L^F \times \{\top, \bot\}^E \times \ V^{E \times T_v \times AG} \times \{\top, \bot\}^V$$

where

- N^T is the operation of creating an sc-node with the specified type;
- $C^{E \times E \times T}$ is the operation of creating an sc-connector between two given sc-elements with the specified type;
- $F^{N \times L}$ is the operation of setting the contents to an sc-node;
- $\{E^{E \times \{E \cup T\} \times E} \cup \ E^{\{E \cup T\} \times T \times E} \cup E^{E \times T \times \{E \cup T\}}\}$ are operations of searching for three-element sc-constructions by given first and/or second and/or third sc-elements;
- T^E is the operation of obtaining the type of the given sc-element;
- F^L is the operation of obtaining ostis-system files by their contents;
- L^F is the operation of obtaining the content from the ostis-system file;
- $\{\top, \bot\}^E$ is the operation of deleting the specified sc-element.
- $V^{E \times T_v \times AG}$ — operation of creating a subscription to an event in sc-memory;
- $\{\top, \bot\}^V$ — operation of removing a subscription to an event in sc-memory;

Let us consider some of the algorithms of the described operations. The algorithm of the operation of creating an sc-node with the specified type N^T can be described as follows:

- Step 1: Verify that the specified sc-element type is a subtype of sc-node.
  - If the specified sc-element type is not a subtype of sc-node, then terminate the algorithm with an error.
  - Otherwise, proceed to step 2.
- Step 2: Allocate a new sc-memory cell for the sc-node.
  - If the sc-memory is engaged, terminate the algorithm with an error.
  - Otherwise, proceed to step 3.
- Step 3: Set the type for the cell as sc-node with the specified type, go to step 4.
- Step 4: Return the resulting sc-address of the sc-node and terminate.

The algorithm for the operation of creating an sc-connector between two given sc-elements with the specified type $C^{E \times E \times T}$ can be described as follows:

- Step 1: Verify that the specified sc-element type is a subtype of sc-connector.
  - If the specified sc-element type is not a subtype of sc-node, then terminate the algorithm with an error.
  - Otherwise, proceed to step 2.
- Step 2: Check that the sc-addresses of the start and end sc-elements are valid.
  - If the sc-address is not valid, then terminate the algorithm with an error.
  - Otherwise, proceed to step 3.
- Step 3: Allocate a new sc-memory cell for the sc-connector.
  - If the sc-memory is engaged, terminate the algorithm with an error.
  - Otherwise, proceed to step 4.
- Step 4: Set the type for the cell as sc-connector with the specified type, go to step 5.
- Step 5: Add the sc-connector to the list of outgoing and incoming arcs of the start and end sc-elements, go to step 6.
- Step 6: Notify the outgoing and incoming arc addition events, go to step 7.
- Step 7: Return the resulting sc-connector address and terminate.

The algorithm of the operation of deleting a given sc-element $\{\top, \bot\}^E$ can be described as follows:

- Step 1: Attempt to acquire a cell in sc-memory by the sc-address of the sc-element.
  - If the cell is not found, terminate the algorithm with an error.
  - Otherwise, proceed to step 2.
- Step 2: Initialize the stack to remove sc-elements, go to step 3.
- Step 3: Put the sc-address of the sc-element to be deleted into the deletion stack, go to step 4.
- Step 4: Place the sc-addresses of all sc-connectors for which the given sc-element is the start or end sc-element and the sc-addresses of all connectors for which the found sc-connectors are the start or end sc-elements on the deletion stack, go to step 5.
- Step 5: While the deletion stack is not empty, set the cells for the sc-elements as released and release

all those cells.
  - For all sc-connectors to be deleted, notify outgoing and incoming sc-connector deletion events.
  - For all sc-cells to be deleted, notify sc-cell deletion events.
  - Go to step 6.
- Step 6: Destroy the stack for sc-element deletion, go to step 7.
- Step 7: Terminate.

The algorithm for the operation of setting the content to a given sc-node $F^{N \times L}$ can be described as follows:

- Step 1: Attempt to retrieve a cell in sc memory by the sc address of the sc node.
  - If the cell is not found, then terminate the algorithm with an error.
  - Otherwise, proceed to step 2.
- Step 2: Change the sc-node type to an ostis-system file, go to step 3.
- Step 3: Add the string to the file storage of sc-memory.
  - Add the string to a free segment of the file storage.
  - Assign matches between this string and the specified ostis file.
  - Notify the event of changing the content of the ostis-system file.
  - Go to step 4.
- Step 4: Terminate.

The principles of search operations were discussed in [2].

This programming interface provides all the necessary functionality for working with sc-constructions, file constructions, events and processes in the memory.

G. Conclusions

The proposed model of the shared semantic memory includes a formal description of the following (!):

- how to represent, store, and process graph and string constructions, events, and processes in the memory;
- how to ensure efficient execution of operations in this shared memory;
- how to coordinate multiple processes running at the same time on the same memory location,
- how to efficiently utilize the available computing power, etc.,

and allows to (!):

- efficiently organize joint storage of graph constructions and string content of external information constructions not represented as a graph, using graph-dynamic and event-driven models;
- efficiently manage the address space, i.e. distribute information about these constructions the in memory in the most effective way;
- efficiently allocate processes to work with these constructions in single-threaded and multi-threaded environments;
- provide coordinated (synchronized) execution of several processes in one memory;
- ensure consistency of operations at the level of representation and processing of data in the memory.

This model has many merits, but the following issues remain unresolved:

- how to ensure the security of information storage and processing, that is:
  - how to ensure access rights for constructions stored in the memory;
  - how to efficiently process and assign these access rights to processes;
  - and more;
- how to ensure consistency of operations at the knowledge representation and processing level, i.e.:
  - how to implement transactions for graphs;
  - how to ensure the integrity and atomicity of some group of operations on a subgraph;
  - how to ensure error-free execution of these transactions;
  - and more;
- how to ensure the storage and processing of information in teams of intelligent systems [28], that is:
  - how to organize storage and processing of information in distributed memory, i.e. in memory not on one device, but on multiple devices;
  - how to efficiently organize data transfer over a network between several devices;
  - and more;
- how to organize the configuration of memory components from the memory itself.
- and so on.

Nevertheless, the results of this paper are very significant for future work. These questions will be discussed in the following papers. Let us consider some obtained quantitative characteristics of the implementation of the proposed model.

IV. The software implementation of sc-memory for next-generation intelligent systems

A. Description of the sc-memory implementation

The current version of sc-machine is implemented on the Linux operating system (Ubuntu-22.04) [29] and is available on GitHub [30]. When developing sc-machine according to the described model, we used modern development environments (CLion, VSCode), containerization tools (Docker), programming languages (C, C++, CMake), as well as standard libraries and frameworks supplied together with compilers of the programming languages used. The development was based on the models and tools described in the previous section, as well as the *OSTIS Technology Standard* described in the current version of the *OSTIS-2023 monographs* [9].

The current Software implementation of sc-memory has the following features:

- The memory allocation and destruction mechanisms of the GLib library are used to manage dynamic memory.
- Prefix trees [31] and linked lists are used as data structures to store *information constructions* that do not belong to *SC-code*. The reasons for that are as follows [32]:
  - prefix structures are fairly easy to understand and minimal in their syntax;
  - prefix structures are convenient enough to store and handle "key-value" relations;
  - access to a value by a key occurs in the worst case for the length of this key [33].

  The Implementation of file memory allows storing and searching any kind of information constructs (including binary files).
- To synchronize processes in sc-memory, monitors are implemented and used [34], [27]. They provide:
  - locking mechanisms to prevent multiple processes from simultaneously accessing shared resources, eliminating the possibility of mutual exceptions, race conditions, and data access conflicts;
  - high-precision time synchronization between processes that allows them to work in a coordinated mode, which eliminates the possibility of some processes being hungry.

  The implementation of monitors uses mutexes, condition variables, and queues.
- The current *Programming interface of the Software implementation of sc-memory* allows:
  - implement platform-dependent components to a necessary and sufficient extent, almost independently of sc-memory implementation.
  - implement basic tools for designing platform-independent ostis-systems.
- The current Implementation of sc-memory is fully consistent with the current Implementation of scp-interpreter.

In general, *sc-memory* can be implemented in different ways. For example, another variant of *ostis-platform sc-memory* can be realized by a program implementation of *Neo4j DBMS*. The difference between such a possible *sc-memory* implementation and the current one is that the storage of *graph constructions* and the control of the flow of actions over them should be realized more by means provided by *Neo4j DBMS*, while the representation of *graph constructions* should be implemented in its own way, because it depends on the *SC-code syntax*. [18].

B. *Efficiency of sc-memory operations*

The current Software implementation of sc-memory in the Software platform for ostis-systems allows to store and represent *sc-constructions*, external *information constructions* not belonging to *SC-code*, as well as to control and coordinate processes in it.

The results of sc-memory operations testing, which includes the implementation of the process control model, showed that parallel execution of sc-memory operations is efficient when the number of operations is large enough (e.g., 1,000,000 operations) (Table 1).

Table I
Efficiency of using 4 physical threads to perform 1,000,000 sc-memory operations compared to 1 physical thread

Number of physical threads	1 thread	4 threads	
	Response time, ms	Response time, ms	Speedup, times
Operations of addition (modification)			
Operation of sc-node creation	958,025	369,680	2.591
Operation of sc-connector creation	1,299.740	787.001	1.652
Operation of adding content to ostis-system file	29,885.500	9,555.450	3.128
Operations of search			
Operation of searching sc-connectors outgoing from a given sc-element	642.378	203.005	3.164
Operation of searching an ostis-system file by its contents	1,608.650	928.555	1.732
Operations of deletion			
Operation of deleting an sc-element	1,850.950	1,746.270	1,060
Operation of deleting sc-connectors outgoing from a given sc-element	1,704.620	2,115.500	0.806

Testing and evaluation of the effectiveness of the ostis-systems software platform were conducted on one of its latest versions — 0.9.0. This version of the platform solved the problem of controlling processes in the shared semantic memory. During the testing we calculated the main efficiency (performance) indicators of operations over sc-memory in single-threaded and multithreaded environments: response time and throughput, and also calculated the speedup [35] obtained by using parallelism when performing a group of operations of the same class over sc-memory [36]. The computer used was an

HP ProBook Hewlett Packard laptop with a *Intel(R) Core(TM) i7-4900MQ* processor (4 cores with 2 threads) with a configured core frequency of 3.20 GHz, 16 GB RAM, and 256 GB SSD.

At the same time, parallel execution of a small number of operations over sc-memory (for example, 100 or 10,000 operations) in some cases can be worse than their sequential execution (Table 2).

This behavior is related to the peculiarities of the control mechanisms of the processes in the shared semantic memory, the classes of operations to be performed, and their specified input values in the context of the problem to be solved. For example, all sc-construction search operations with the same sc-elements, executed in parallel, do not block each other. For example, the speed of parallel execution of operations on ostis-system files depends on the size of the buffer used when reading external information constructions and writing them to disk, as well as on the length of the information constructions themselves.

Table II
Efficiency of using 4 physical threads to perform 100 sc-memory operations compared to 1 physical thread

Number of physical threads	1 thread	4 threads	
	Response time, ms	Response time, ms	Speedup, times
Operations of addition (modification)			
Operation of sc-node creation	0.099	1.306	0.076
Operation of sc-connector creation	0.150	0.422	0.356
Operation of adding content to ostis-system file	9.521	4.128	2.307
Operations of search			
Operation of searching sc-connectors outgoing from a given sc-element	0.530	0.241	2.200
Operation of searching an ostis-system file by its contents	0.339	1.453	0.233
Operations of deletion			
Operation of deleting an sc-element	0.144	1.494	0.096
Operation of deleting sc-connectors outgoing from a given sc-element	0.182	0.938	0.194

The figure 1 shows Dependence of speedup coefficient from parallel execution of a group of operations of the same class on 4 processes on the number of operations in this group, and the figure 2 shows Dependence of the execution time of a group of operations on the number of processes used.

C. Efficiency of network operations over sc-memory

Network access to sc-memory is provided by the server subsystem of the ostis-systems software platform, implemented on the basis of Websocket and JSON languages (protocols) and providing network operations (commands) over sc-memory [3].

In the process of testing the implementation, the throughput of its commands was calculated. During the load testing a test client system implemented in C++ was used. The same device was used as the device used for testing operations over sc-memory. As a result, it was found out that when sending 1000 different commands: *commands for creating sc-elements*, *commands for processing contents of ostis-system files* and *commands for deleting sc-elements* — the time spent on their processing did not exceed 0.2 seconds. At the same time, in some cases it took no more than 0.14 seconds to process 1000 *commands for creating sc-elements*, while for *commands for deleting sc-elements* it took no more than 0.12 seconds, *commands for processing the contents of ostis-system files* — no more than 0.10 seconds, *commands to search for sc-constructions isomorphic to a given five-element graph-template* — no more than 0.45 seconds.

D. Conclusions

From the test results, it is clear that the current implementation of the ostis-systems software platform is an effective means of processing distributed information using both the software interface and the network interface and communication protocols.

The current *Implementation of sc-memory* provides:

- stability in single-threaded and multi-threaded modes;
- dast speed of work in single-threaded and multi-threaded modes;
- reliability of knowledge and data storage and processing in single- and multi-threaded modes.

The proposed shared semantic memory model enables efficient tracking and synchronization of parallel data accesses. The implementation of this model demonstrates a significant increase (by 2-3 orders of magnitude) in the throughput of parallel task execution compared to previous versions of the platform. However, to ensure (causal, sequential) consistency of processes and their operations, besides the data level, it is necessary to manage the knowledge level [37] .

V. Conclusion

In this paper, a model and implementation of the shared semantic memory has been proposed and discussed in detail, including (!):

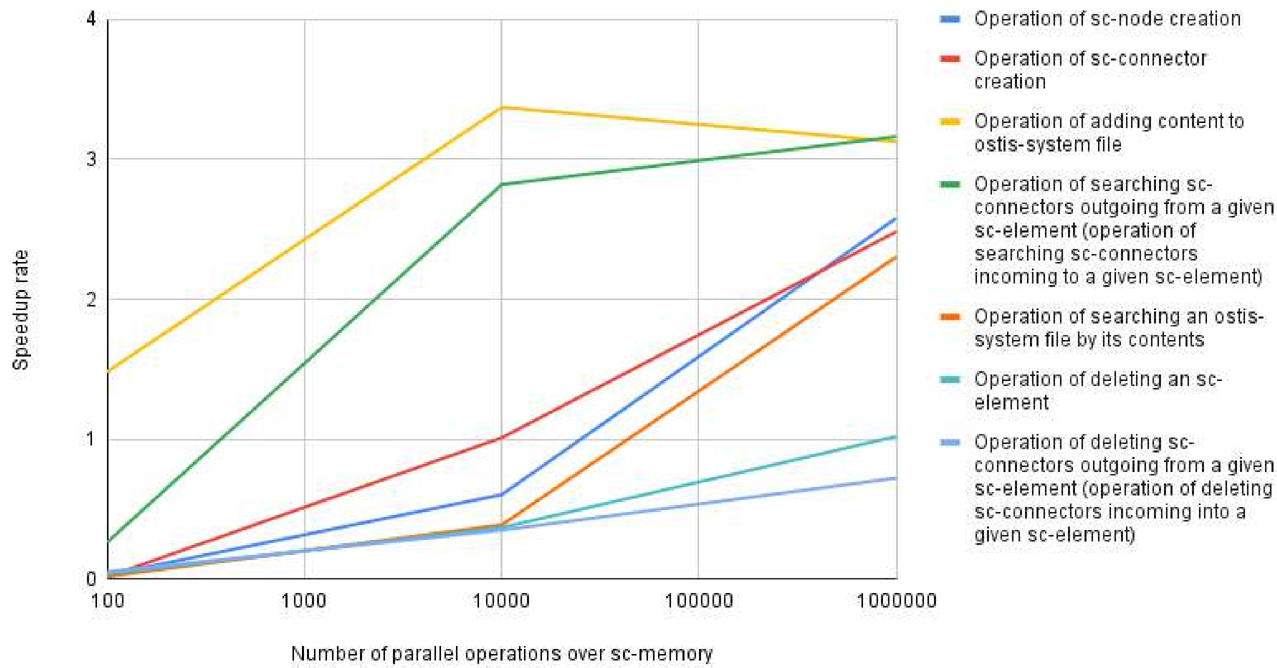


Figure 1. Dependence of speedup coefficient from parallel execution of a group of operations of the same class on 4 processes on the number of operations in this group

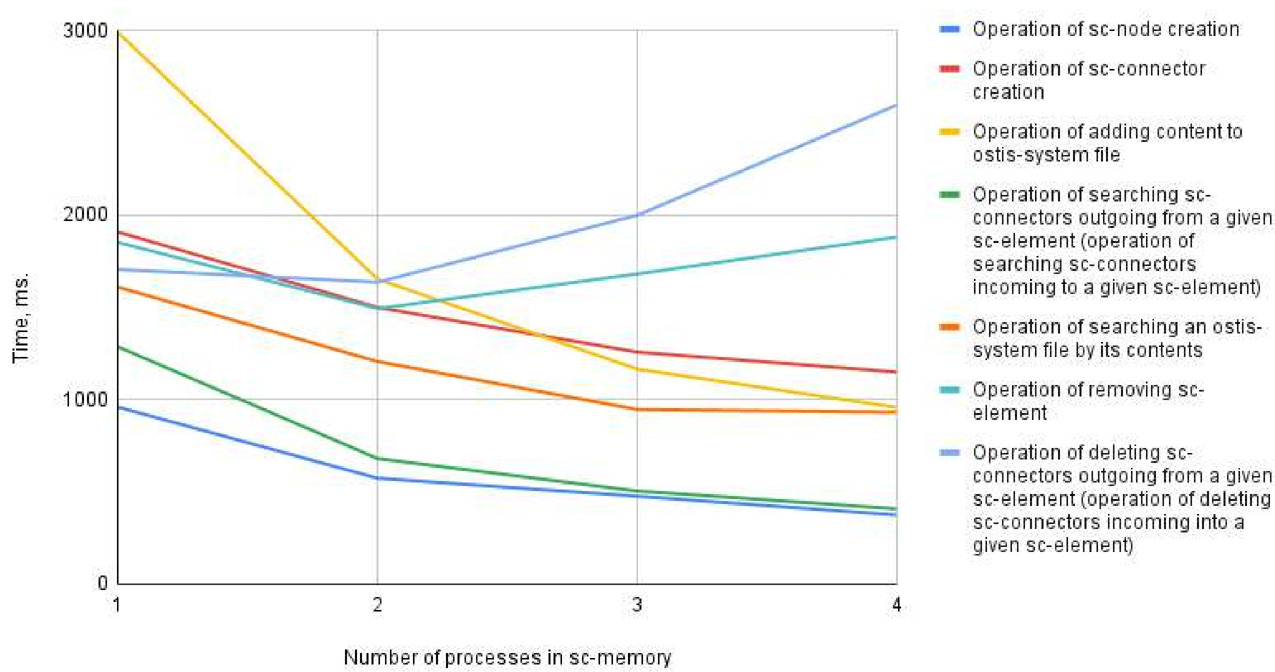


Figure 2. Dependence of the execution time of a group of operations on the number of processes used

- a storage for unified representation and processing of graph constructions;
- a storage for unified representation of string constructions used as file contents in graph constructions;
- a storage for managing events in this memory;
- a storage for managing processes running in this memory;
- a set of operations for working with this memory.

The proposed model of the shared semantic memory includes:

- models and algorithms for allocating and releasing cells in this memory, providing:
  - reusability of the released memory segments;
  - ability to utilize new vacant memory segments;
- Models and algorithms to efficiently allocate processes in this memory;
  - rapid parallel creation of elements in the memory by allocating processes over the segments of the memory;
  - fast unblockable parallel search of constructions, provided there are no other operations on these constructions.
- Models and algorithms for managing subscriptions to events in this memory;
- Models and algorithms for synchronizing the execution of processes in the shared memory sections, providing:
  - parallel access to sc-memory, i.e. possibility of parallel execution of actions in sc-memory without violating correctness of the data structures in it;
  - absence of deadlocks, races and hungry processes operating in sc-memory.

Promising directions to further this line of work are:

- development of a model for distributed unified representation and processing of information in the unified semantic memory;
- development of a model for representation and storage of platform-dependent agent programs;
- development of a consistency model to ensure correctness of agents' operation on constructions in the memory;
- development of a model of memory configuration from the memory itself.

In addition, other equally important areas of work are:

- improving the documentation of the current Implementation of sc-memory and the current Software implementation of ostis-platform;
- improvement of methodologies and tools for developing documentation of software systems;
- improvement and mass distribution of the Software implementation of ostis-platform and intelligent systems developed on its basis.

The formally described model of semantic memory is consistent with the previously described ontological model of this memory [3]. The author of this paper believes that the used approach to modeling of complex objects will help to simplify the understanding of the operation of intelligent systems developed according to the principles of the OSTIS Technology.

Acknowledgment

The author would like to thank the scientific teams of the departments of Intelligent Information Technologies of Belarusian State University of Informatics and Radioelectronics and Brest State Technical University for their help in the work and valuable comments.

References

[1] N. Zotov, "Design principles, structure, and development prospects of the software platform of ostis-systems," *Otkrytye semanticheskie tekhnologii proektirovaniya intellektual'nykh system [Open semantic technologies for intelligent systems]*, pp. 67—-76, 2023.

[2] ——, "Implementation of information retrieval subsystem in the software platform of ostis-systems," *Otkrytye semanticheskie tekhnologii proektirovaniya intellektual'nykh system [Open semantic technologies for intelligent systems]*, pp. 77–94, 2023.

[3] ——, "Software platform for next-generation intelligent computer systems," in *Otkrytye semanticheskie tekhnologii proektirovaniya intellektual'nykh system [Open semantic technologies for intelligent systems]*. BSUIR, Minsk, 2022, pp. 297—-326.

[4] V. Golenkov, N. Guliakina, and D. Shunkevich, *Otkrytaja tehnologija ontologicheskogo proektirovanija, proizvodstva i jekspluatacii semanticheski sovmestimyh gibridnyh intellektual'nyh komp'juternyh sistem [Open technology of ontological design, production and operation of semantically compatible hybrid intelligent computer systems]*, V. Golenkov, Ed. Minsk: Bestprint [Bestprint], 2021.

[5] V. Golenkov *et al.*, "Associative semantic computers for intelligent computer systems of a new generation," in *Otkrytye semanticheskie tekhnologii proektirovaniya intellektual'nykh system [Open Semantic Technologies for Intelligent Systems (OSTIS)]*, vol. 7. BSUIR, Minsk, 2023, pp. 39–60.

[6] V. Golenkov, "Ontology-based design of intelligent systems," *Otkrytye semanticheskie tekhnologii proektirovaniya intellektual'nykh system [Open semantic technologies for intelligent systems]*, pp. 37–56, 2017.

[7] D. Shunkevich, "Ontology-based design of hybrid problem solvers," in *Otkrytye semanticheskie tekhnologii proektirovaniya intellektual'nykh system [Open semantic technologies for intelligent systems]*. Cham: Springer International Publishing, 2022, pp. 101–131.

[8] S. Auer, V. Kovtun, M. Prinz, A. Kasprzik, M. Stocker, and M. E. Vidal, "Towards a knowledge graph for science," in *Proceedings of the 8th international conference on web intelligence, mining and semantics*, 2018, pp. 1–6.

[9] V. Golenkov, Ed., *Tehnologija kompleksnoj podderzhki zhiznennogo cikla semanticheski sovmestimyh intellektual'nyh komp'juternyh sistem novogo pokolenija [Technology of complex life cycle support of semantically compatible intelligent computer systems of new generation]*. Bestprint, 2023.

[10] A. Palagin and N. Petrenko, "To the question of system-ontological integration of subject area knowledge," *Mathematical Machines and Systems*, vol. 1, no. 3-4, pp. 63–75, 2007.

[11] M. Orlov, "Comprehensive library of reusable semantically compatible components of next-generation intelligent computer systems," in *Otkrytye semanticheskie tekhnologii proektirovaniya intellektual'nykh system [Open semantic technologies for intelligent systems]*. Minsk : BSUIR, 2022, pp. 261–272.

[12] V. Golenkov, N. Guliakina, V. Golovko, V. Krasnoproshin, “Methodological problems of the current state of works in the field of artificial intelligence,” in *Otkrytye semanticheskie tekhnologii proektirovaniya intellektual’nykh system [Open semantic technologies for intelligent systems]*, ser. 5, V. Golenkov, Ed. BSUIR, Minsk, 2021, pp. 17–24.
[13] V. Ryen, A. Soylu, and D. Roman, “Building semantic knowledge graphs from (semi-) structured data: a review,” *Future Internet*, vol. 14, no. 5, p. 129, 2022.
[14] P. Barnaghi, A. Sheth, and C. Henson, “From data to actionable knowledge: Big data challenges in the web of things [guest editors’ introduction],” *IEEE Intelligent Systems*, vol. 28, no. 6, pp. 6–11, 2013.
[15] C. Kellogg, “From data management to knowledge management,” *Computer*, vol. 19, no. 01, pp. 75–84, 1986.
[16] D. Shunkevich, “Universal model of interpreting logical-semantic models of intelligent computer systems of a new generation,” in *Otkrytye semanticheskie tekhnologii proektirovaniya intellektual’nykh system [Open semantic technologies for intelligent systems]*, V. Golenkov, Ed. BSUIR, Minsk, 2022, p. 285–296.
[17] Golenkov, V. V., “Graphodynamic models of parallel knowledge processing: principles of construction, implementation, and design,” in *Open semantic technologies for designing intelligent systems (OSTIS-2012): Proceedings of the II International Scientific and Technical Conference, Minsk, February 16-18, 2012*, Belarusian State University of Informatics and Radioelectronics; editorial board: V. V. Golenkov (chief editor) [et al.]. Minsk, 2012, pp. 23–52.
[18] V. Ivashenko, “General-purpose semantic representation language and semantic space,” in *Otkrytye semanticheskie tekhnologii proektirovaniya intellektual’nykh system [Open semantic technologies for intelligent systems]*, ser. Iss. 6. Minsk : BSUIR, 2022, pp. 41–64.
[19] V. V. Golenkov and N. A. Gulyakina, “Structuring the semantic space,” in *Open semantic technologies for designing intelligent systems (OSTIS-2014): Proceedings of the IV International Scientific and Technical Conference*, V. V. Golenkov, Ed. Minsk: BSUIR, 2 2014, pp. 65–78, chief editor Golenkov, V. V. [and others].
[20] ——, “Principles of building mass semantic technology for component design of intelligent systems,” in *Open semantic technologies for designing intelligent systems (OSTIS-2011): Proceedings of the international scientific and technical conference*, V. V. Golenkov, Ed. Minsk: Belarusian State University of Informatics and Radioelectronics, 2 2011, pp. 21–58, chief editor Golenkov, V. V. [et al.].
[21] D. Shunkevich, “Hybrid problem solvers of intelligent computer systems of a new generation,” *Otkrytye semanticheskie tekhnologii proektirovaniya intellektual’nykh system [Open semantic technologies for intelligent systems]*, no. 6, pp. 119–144, 2022.
[22] “Neo4j graph Database Platform | Graph Database Management System [Electronic resource],” April 2024. [Online]. Available: https://neo4j.com/
[23] T. Kahveci and A. K. Singh, “An efficient index structure for string databases,” in *VLDB*, vol. 1, 2001, pp. 351–360.
[24] M. Barsky, U. Stege, and A. Thomo, “Structures for indexing substrings,” in *Full-Text (Substring) Indexes in External Memory*. Springer, 2012, pp. 1–15.
[25] N. V. Zotov, “Model of process management in shared semantic memory of intelligent systems,” in *Information Technologies and Systems 2023 (ITS 2023)*, L. Y. Shilin, Ed. Minsk: Belarusian State University of Informatics and Radioelectronics, 11 2023, pp. 53–54, proceedings of the International Scientific Conference.
[26] J. L. W. Kessels, “An alternative to event queues for synchronization in monitors,” *Communications of the ACM*, vol. 20, no. 7, pp. 500–503, 1977.
[27] C. A. R. Hoare, “Monitors: An operating system structuring concept,” *Communications of the ACM*, vol. 17, no. 10, pp. 549–557, 1974.
[28] A. Zagorskiy, “Principles for implementing the ecosystem of next-generation intelligent computer systems,” in *Otkrytye semanticheskie tekhnologii proektirovaniya intellektual’nykh system [Open semantic technologies for intelligent systems]*. BSUIR, Minsk, 2022, p. 347–356.
[29] R. Love, *Linux system programming: talking directly to the kernel and C library*. O’Reilly Media, Inc., 2013.
[30] “Software implementation of semantic networks processing storage [Electronic resource],” 2024, mode of access: https://github.com/ostis-ai/sc-machine. — Date of access: 30.03.2024.
[31] R. Bayer, “Prefix b-trees,” *ACM Transactions on Database Systems (TODS)*, vol. 2, no. 1, pp. 11–26, 1977.
[32] P. Ferragina and R. Grossi, “The string b-tree: A new data structure for string search in external memory and its applications,” *Journal of the ACM (JACM)*, vol. 46, no. 2, pp. 236–280, 1999.
[33] D. Belazzougui, “Fast prefix search in little space, with applications,” in *European Symposium on Algorithms*, 2010, pp. 427–438.
[34] M. I. Cole, *Algorithmic skeletons: structured management of parallel computation*. Pitman London, 1989.
[35] L. Gonnord, L. Henrio, L. Morel, and G. Radanne, “A survey on parallelism and determinism,” *ACM Computing Surveys*, vol. 55, no. 10, pp. 1–28, 2023.
[36] N. V. Zotov, “Quantitative indicators of operations efficiency over shared semantic memory of intelligent systems,” in *Information Technologies and Systems 2023 (ITS 2023)*, L. Y. Shilin, Ed. Minsk: Belarusian State University of Informatics and Radioelectronics, 11 2023, pp. 51–52, proceedings of the International Scientific Conference.
[37] L. P. Miret, “Consistency models in modern distributed systems. an approach to eventual consistency,” *Master. MA thesis. Universitat Politecnica de Valencia, Spain*, 2014.

ФОРМАЛЬНАЯ МОДЕЛЬ ОБЩЕЙ СЕМАНТИЧЕСКОЙ ПАМЯТИ ДЛЯ ИНТЕЛЛЕКТУАЛЬНЫХ СИСТЕМ НОВОГО ПОКОЛЕНИЯ

Зотов Н.В.

В работе подробно рассматривается формальная модель семантической памяти для интеллектуальных систем, структура, её элементы, соответствия между ними, правила и алгоритмы. Описывается реализация на основе данной модели, приводятся количественные показатели её эффективности.

Received 15.03.2024

Adaptive User Interfaces for Intelligent Systems: Unlocking the Potential of Human-System Interaction

Mikhail Sadouski, Pavel Nasevich, Maksim Orlov, Alexandra Zhmyrko
Belarusian State University of
Informatics and Radioelectronics
Minsk, Belarus
Email: sadovski@bsuir.by, fallenchromium@gmail.com, orlovmaksimkonst@gmail.com, aleksashazh@gmail.com

***Abstract*—The paper analyzes the capabilities of computer systems and the level of development of tools for interacting with them (user interfaces). Based on the analysis, an approach to the design of adaptive user interfaces of intelligent systems based on the OSTIS Technology is proposed. The semantic model of such interfaces, proposed earlier, has been clarified and extended, the proposed architecture of such systems is given. Adaptive user interfaces designed on the basis of the proposed approach will provide new scenarios of user interaction with computer systems.**

***Keywords*—adaptive user interface, intelligent systems, user interface of ostis-systems**

I. Introduction

In the modern world, people use computer systems of various purposes daily. A key component of such systems that directly influences their efficiency is the user interface, which in a broad sense is a set of tools that provide interaction between a person and the system.

A large part of the cost of developing computer systems is in the design, testing, and development of the user interface [1].

A poorly designed user interface limits the potential of the system by increasing the threshold of entry and reducing the efficiency of interaction with users or making some interaction scenarios impossible [2].

With the development of information society, the need of users for computer systems capable of solving various classes of problems, including tasks that are difficult to formalize, has led to an increased pace of development of computer technologies, the creation of a large number of models, methods and tools for the design and development of computer systems, including intelligent computer systems with increased requirements for interoperability, component compatibility and flexibility of scenarios of interaction with the user because of their ability to self-learning and solving complex problems, as well as problems, in the initial data and algorithms of solution of which there is an influence of non-factors.

However, even computer systems that we use on a daily basis are severely limited in their functionality due to the limited means of interaction with these systems. There is a mismatch between the current level of development of user interfaces and the problem-solving capabilities of computer systems.

Each user has unique needs and the application of adaptive user interfaces for intelligent systems becomes essential. The ease and flexibility of dynamically changing user interfaces based on user tasks that are not predetermined in the design of the system becomes key and allows for greater potential for interaction with intelligent computer systems.

This article analyzes the capabilities of computer systems and the level of development of tools for interacting with them (user interfaces).

The user interface is considered in this context as the language of communication between the system and the user, together with the means for such communication, emphasizing the content of the interaction more than the specific technical aspects of the implementation of the interaction. This analysis is based on a description of the historical development of computer systems (and, as a consequence, the development of the many classes of problems that computer systems solve).

Based on the analysis, an approach to the design of adaptive user interfaces of intelligent systems based on the OSTIS Technology is proposed. The proposed [3] semantic model of such interfaces is clarified and extended, and the proposed architecture of such systems is given. Adaptive user interfaces designed on the basis of the proposed approach will provide new scenarios of user interaction with computer systems.

II. Analysis of existing approaches to solving the problem

A. Challenges of modern user interfaces

The challenges of modern user interfaces that cause users to fail to take advantage of the full potential of computer system problem solvers can be categorized as follows:

- A mismatch between the user's skills and means of interacting with the system and the actual means of interaction provided by the system. An example is an over-complicated user interface for inexperienced users, or a user interface that does not take into account that the user is fully informed about the algorithm for solving a problem and takes time and attention away from unnecessary explanations. This leads to the fact that the user cannot predict the navigation through the user interface — his cognitive load increases, the time required to solve the problem increases [4].
- Mismatch between the task and environment for which the user interface was designed with the actual task and environment of the system. This refers to the case where the system's problem solver is capable of solving a broader and more general class of problems than the system's user interface allows, since only in such a case can the user interface alone be said to be the limiting factor in the applicability of the system — [5].
- Lack of user's ability to make changes to the interface for integration with other user interfaces and/or systems. The impossibility of building an integrated working environment (i.e. a new user interface, qualitatively different from the multitude of interfaces designed to solve each of the sub-tasks of the complex task) makes it impossible to build scenarios for automatic integration of computer system subsystems and severely limits the user's ability to integrate computer systems, because without the ability to change the user interface, the program behavior can only be changed programmatically [6], [7].

There are a lot of examples of functionalities limited on the user interface side, which are nevertheless technically realizable, but we will limit ourselves to a few significant ones:

- Lack of integration of system components: your word processor does not prompt you to open the last quarter's financial report, even though you were emailed it yesterday and are scheduled to check it on your electronic calendar today.
- Lack of automatic adaptation to the environment: the online store does not offer you to make a pickup to the branch you are closest to at the moment.
- Lack of ability to personalize the user interface: your fitness app does not allow you to fully customize the start screen with health metrics that are irrelevant to you and with workout suggestions that do not match your goal and access to fitness equipment.
- Solving complex tasks: for an average user the interface for automated solution of a multi-step task is not standard and is always available. For example, to process a call or a letter from a customer, enter relevant information from the letter into the project accounting system, assign a person with the necessary competence to be responsible for the project and send the customer a letter with the contact of the person responsible for the project.

It is worth noting that for each of these examples it is possible to find a counterexample: a computer system that did take into account the described use case, but it is worth recognizing that in general the described limitations exist in the vast majority of computer systems.

B. Overview of the evolution of computer systems and their user interfaces

Let us consider the development paths of computer systems and their user interfaces - the major milestones, and what they were driven by (discoveries or user needs).

Over the last 60 years, a large number of approaches to user interface design have been developed. The approaches focused on different directions, such as "which side of the interaction is primary" (the emotional state of the user, the purpose of the system, feedback from the user, the design of the program that implements the algorithm for solving the problem) and "how to convey the meaning of the objects of the user interface" — confrontation of metaphorical and idiomatic approach [8], [9].

At the same time, the paradigms used to implement the user interface have evolved. For example, the transition from a model where each interaction has an underlying function call to perform the process and provide feedback to an object-oriented interface where the UI elements correspond to the properties of the system entities (reactive approach).

Currently, the approach when the user interface is described by a set of states and transitions in the algorithm of problem solving is popular (i.e. the paradigm of so-called "wizards", which leads the user step-by-step from the need to its resolution). Each of the approaches has a place, but a special place in the design of user interfaces for intelligent computer systems, based on the dynamic nature of the tasks that can be set before it, takes the paradigm of interface design based on the problem being solved and the algorithm for its solution. To date, there are no comprehensive means of building adaptive user interfaces (taking into account the environment, the characteristics of the user and his device), while dynamically taking into account the information about the algorithm for solving the problem.

C. Conclusions

Based on the representation of the user interface described in I, we can conclude that the problems described in II-A are the result of a methodological error in the design of interfaces. At the moment, the factors affecting how the user interface should look like are taken into account by the developer at the time of creation of the computer system and are laid in its program code in the form of ready-made structures created under the influence of the model of communication with the user, the model of the environment, the model of the user and the proposed algorithm for solving the problem. Often this modeling process does not lead to an optimal result. In addition, computer systems tend to receive new functions within their life cycle, which entails updating the user interface, which in turn leads to repeated manual design of the user interface taking into account all the factors described above.

User interfaces of computer systems are suboptimal largely because the toolkit for creating user interfaces with tools to solve the problems described above is not sufficiently refined, versatile, or widespread. In order to create a complete interface development toolkit, it is necessary to describe the mechanisms of system-human communication, in particular, to define the classes of transmitted messages in order to build a model of human communication based on the task facing the system. It is also necessary to describe what components of the user interface implement certain messages between the system and the user; describe relevant properties of the user and the environment, and automate the rules by which the user interface should change in the presence of certain properties of the environment or the user. Thus, formalize the experience of experts in building user interfaces and turn it into an automated algorithm. Such a tool would have the ability to generate an interface for systems with dynamic problem formulation, adapt and personalize the interface to the user and the environment.

III. Proposed approach

A. General principles of building adaptive user interfaces of intelligent systems

An ontological approach based on the semantic model proposed in [3] is suggested to build adaptive user interfaces of intelligent systems. This approach assumes:

- "lexical" interface description — a description of the components from which the interface is formed;
- "syntactic" description of an interface — rules for forming a correct interface from its components;
- semantic description. Knowledge about what entity the displayed component is a sign of. The semantic description also includes the purpose, scope of application of the interface components, description of the user's interface activity, etc. The semantic description also includes the purpose, scope of application of the interface components, description of the user's interface activity, etc.

The following ontologies need to be implemented within the approach:

- ontology of problems and algorithms for their solution;
- ontology of user interfaces of intelligent computer systems;
- ontology of components of user interfaces of intelligent computer systems;
- ontology of the context of use of user interfaces of intelligent computer systems;
  - ontology of users of intelligent computer systems;
  - ontology of intelligent computer systems users' actions;
  - ontology of the environment of intelligent computer systems;
  - ontology of devices of intelligent computer systems;
- ontology of external information constructs;
- ontology of incoming and outgoing messages of intelligent computer systems.

In addition to the proposed ontologies, it is necessary to develop tools for automatic interface generation and editing of its model to enable interface modification during operation and tools to support the design of user interfaces.

The above means are also proposed to be realized with the help of the above ontologies, which allow to fully describe in the knowledge base of the intelligent system both the interface itself and the principles of its interaction with the user, as well as the mechanisms of adaptation of the interface depending on the user and the environment. It is important to note that algorithms for solving user tasks are also proposed to be formed dynamically in the knowledge base of the system.

The application of this approach requires a basis in the form of intelligent systems design technology, which allows solving complex problems and whose components integrate with each other and have a synergistic effect. In turn, the application of the ontological method of building adaptive user interfaces within the framework of such technology will make intelligent systems practically applicable for a large number of classes of tasks.

The technology that meets these requirements is the OSTIS Technology. Intelligent systems developed according to the OSTIS Technology are called ostis-systems [10].

The advantages of the OSTIS Technology are as follows:

- unification of different types of knowledge with the help of SC-code;
- ontological approach to knowledge structuring;
- availability of a variant of realization of the platform of interpretation of semantic models of intellectual systems (sc-models);
- problem solver is based on a multi-agent approach in which agents interact with each other solely by specifying the actions they perform in a shared semantic memory;

- library of reusable knowledge base components and problem solvers;
- the possibility of semantic representation of logical formulas and statements;
- availability of a variant of implementation of the interpreter of logical models of problem solving.

The OSTIS Technology allows to implement the user interface in the simplest and most efficient way due to the use of automation tools for the design of user interface components, as well as due to the library of reusable components of ostis-systems, which significantly reduces the time of interface development [11].

The main ostis-system within the OSTIS Ecosystem is the OSTIS Metasystem [12]. The OSTIS Metasystem allows to automate the development of ostis-systems and their components, including user interfaces [13].

The internal universal language of knowledge representation is SC-code [14]. With the help of SC-code the interface of ostis-systems itself and the programs and rules that are used when using the interface are recorded in a unified way. That is, using SC-code, it is possible to write the interface, and also the program which can not only implement some business logic of system, but also the program which will change the interface itself which is used at interaction with the user. Through the use of SC-code and the principles of the OSTIS Technology it is possible to realize the adaptability of user interfaces in the simplest and most flexible way.

To implement the user interface as a subsystem with its own knowledge base and problem solver, it is necessary to develop the listed family of user interface ontologies, as well as a collective of agents for the functioning of the user interface:

- non-atomic sc-agent of interpreting sc-model of user interfaces of ostis-systems;
- non-atomic sc-agent of generation of sc-model of user interface of ostis-systems;
- non-atomic sc-agent of adaptation of sc-model of user interface of ostis-systems to a particular user and external environment;
- non-atomic sc-agent of interpretation of user actions; user interfaces of ostis-systems.

B. User interface module for ostis-systems

The implementation of the user interface of ostis-systems is based on:

- extensibility without the need to change the logic of the module (for personalization and adaptation on the part of the user and for styling and fine-tuning on the part of the system developer);
- maximum simplification for developers to add a new user interface platform (other than the current web-oriented);
- ensuring the simplest possible integration of the UI module into existing systems;
- ensuring that the final interface is comparable in speed to user interfaces developed by traditional means;
- support for a high level of abstraction for operating user interface objects, thus not limiting developers using any means to communicate with users (including virtual reality helmets or interaction with real world objects).

In this regard, the following architectural decisions have been made:

- to implement the functionality of translating the sc-model of the user interface into a specific platform-dependent markup on the side of the ostis-systems problem solver;
- to implement the functionality of user actions interpretation on the side of ostis-systems problem solver;
- to implement a mechanism to track updates to the user interface model and pass only the changed part of the markup to the user interface implementation platform;
- to use a web-oriented platform as the first supported platform for user interface implementation;
- not to use web primitives to describe interfaces, but to limit ourselves to higher-level concepts and relations, such as size, algorithm of mutual arrangement, decomposition of a user interface element (an example of the description of the "Button" component is presented in Figure 1);
- to provide the possibility of applying the rules of user interface adaptation defined declaratively in the knowledge base of ostis-systems.

The general architecture of the module is presented in Figure 2.

The order of interaction can be described as follows:

- the user interacts with the user interface of the ostis-system to solve the task he/she needs;
- information about the interaction is transferred to the knowledge base of the system;
- initiation of a team of agents to solve the user's task is performed;
- results of the user's task solution are stored in the system's knowledge base;
- the user interface agent collective is initiated;
- user interface agents modify the interface model in the knowledge base based on the received information about the result of the performed task as well as the current usage context;
- as a result, adaptation is performed - the user interface model in the knowledge base of the system is changed;
- changing the user interface model in the knowledge base of the system leads to an updated visualization of the user interface.

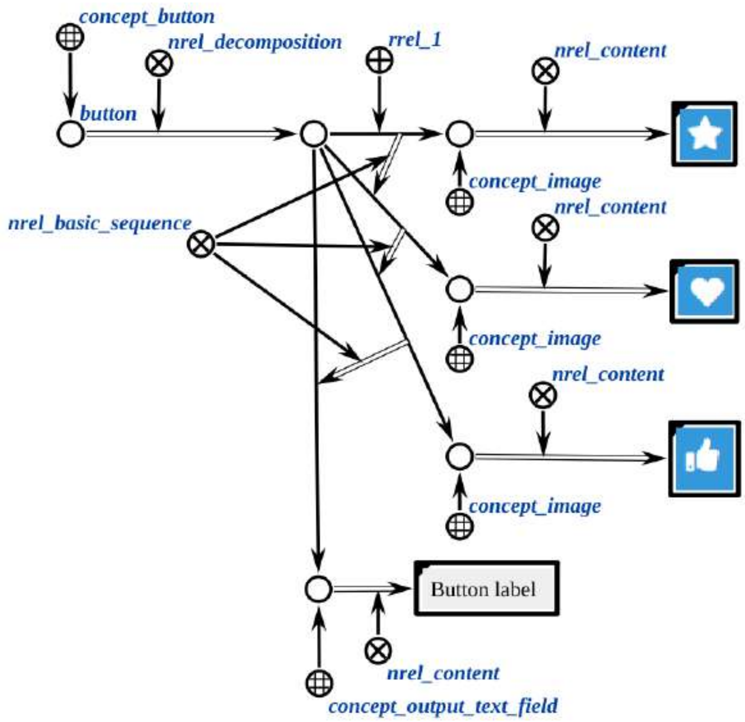


Figure 1. Example of button decomposition

Description of user's action takes place in the knowledge base of the system. An example of formalization of a single button press is shown in Figure 3. The sequence of click-based UI updates is shown in Figure 4.

The algorithm for updating the UI after a user action involves the following steps:

1) The user performs the action of pressing the button.
2) The user interface captures information about the user's action in the system's knowledge base.
3) When this information appears, the user action interpretation agent (UIActionAgent) is initiated.
4) This agent searches the knowledge base for an internal system action that should be executed as a result of pressing a button. The search is based on information about the class of the user action and the UI component for which it was initiated.
5) Having received a template for creating an internal action in the system's knowledge base, UiActionAgent initiates its execution.
6) An internal system action is performed. If necessary, the user interface model in the knowledge base is changed.
7) An UpdateModelAgent call is executed to update the user interface model.
8) For each modified component of the user interface, the agent of forming a visualization of this component for the used platform is called.
9) The modified part of the user interface is redrawn.

The key components of the module are:

- User interface sc-model generator. It is responsible for compiling requirements for the current state of the user interface based on the solution step of the current task. Based on the solution algorithm, the generator finds a set of input information received from the user and a set of elements on which it is necessary to give feedback. Further, the generator specifies the types of received information and feedback, searches for the most general classes of user interface components that solve the problem of input and output of information for this particular task and creates/modifies the sc-model of the user interface containing all the necessary elements.
- SC-model adaptation agent. It is responsible for adding the properties necessary for interpretation to the user interface components (e.g. size, color of elements, their location, etc.) based on the system settings, knowledge about the environment and rules of adaptation to the environment, knowledge about the user and rules of adaptation to certain user characteristics.
- Interpreter. It converts a specific user interface sc-model into an external language compatible with the user interface implementation platform, so that the user interface can be recreated on the platform or the state of an existing interface can be updated.
- User action interpreter. It obtains information about user interactions with the interface and triggers the system to perform the action described in the user

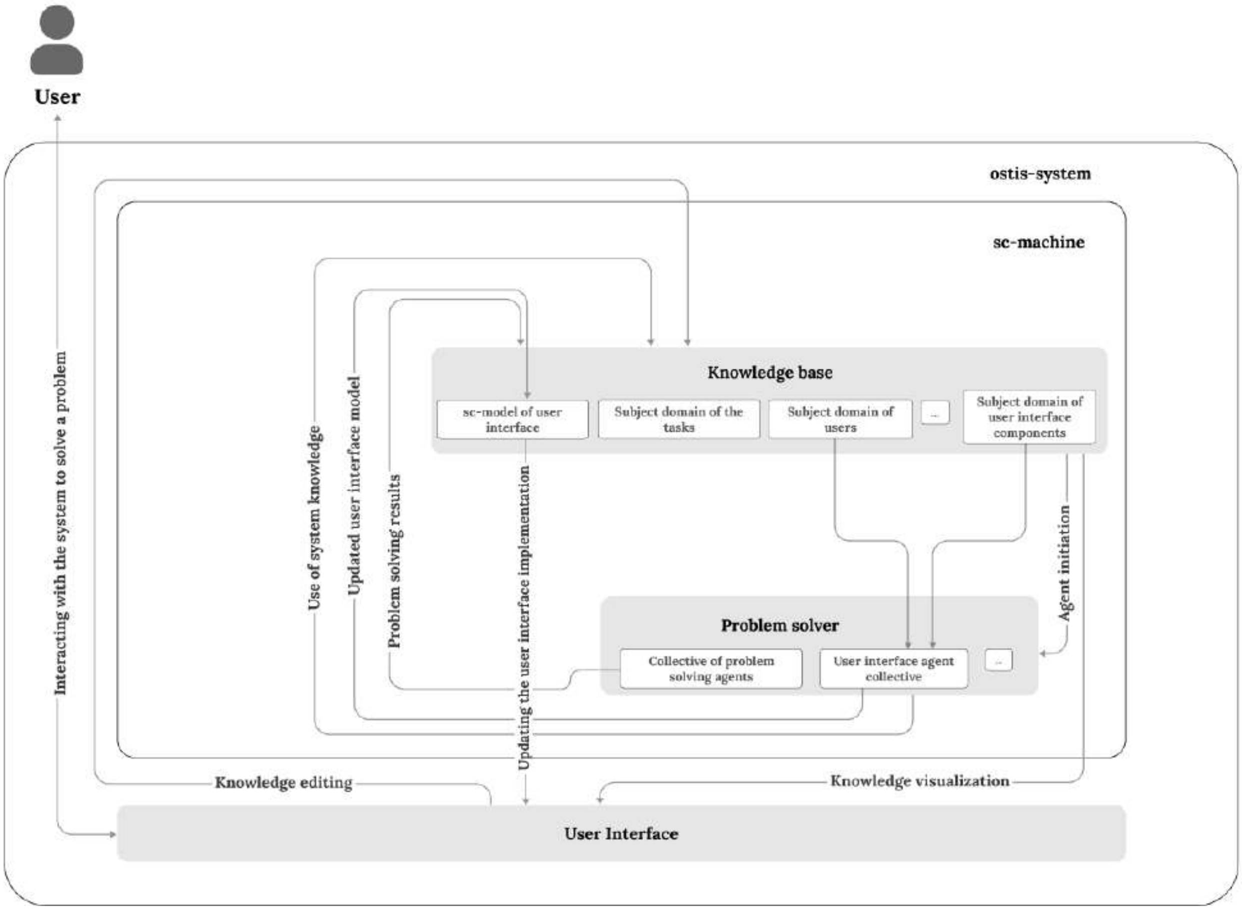


Figure 2. User interface architecture

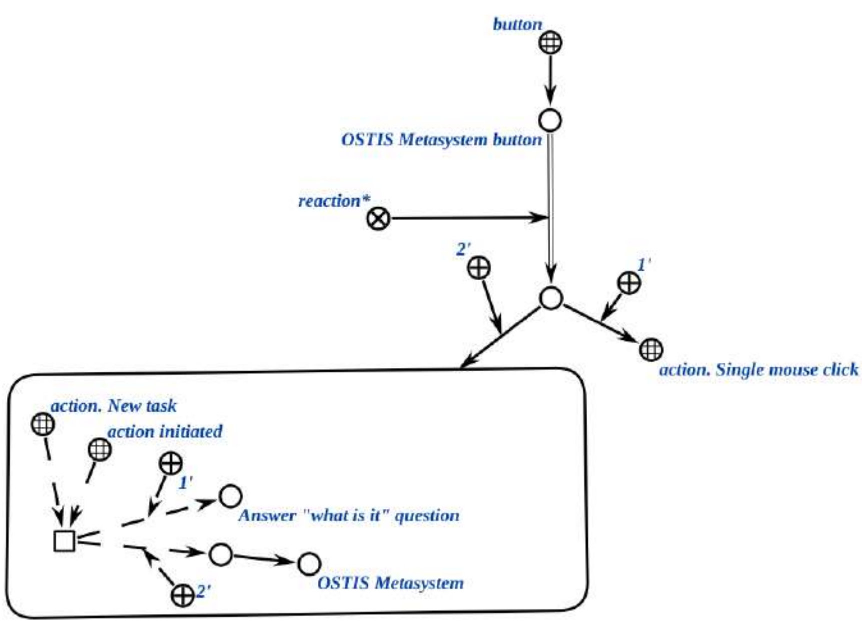


Figure 3. User action example

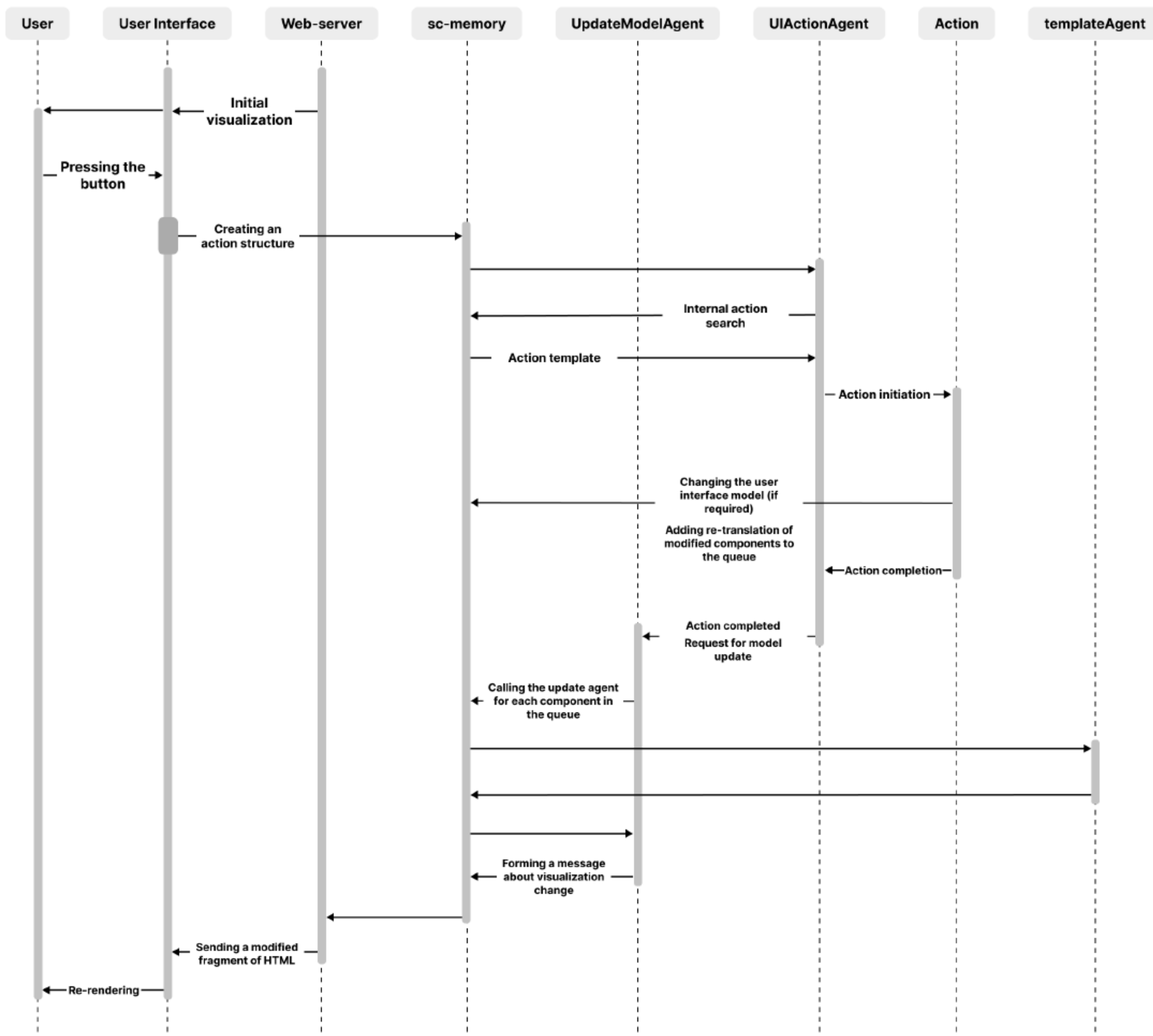


Figure 4. Sequence diagram for updating the user interface based on user interaction

interface component model as a reaction to one or another type of interaction.

The user interface implemented according to the specified architecture will be generated on the basis of its model in the knowledge base, adapted to the needs of users, and dynamically changed depending on the tasks to be solved.

IV. Conclusion

The paper analyzes the capabilities of computer systems and the level of development of tools for interaction with them (user interfaces). Based on the analysis, an approach to the design of adaptive user interfaces of intelligent systems is proposed to provide new scenarios of user interaction with computer systems.

To apply this approach, the OSTIS Technology is used, which allows solving complex problems and whose components integrate with each other and have a synergistic effect. In turn, the application of ontological approach based on the semantic model of building adaptive user interfaces within the framework of the OSTIS Technology allows making intelligent systems practically applicable for a large number of classes of tasks. The paper refines and extends the previously proposed semantic model of adaptive user interfaces of intelligent systems and presents their proposed architecture.

Acknowledgment

The authors would like to thank the scientific collectives of the departments of Intelligent Information Technologies of the Belarusian State University of Informatics and Radioelectronics and Brest State Technical University for their help and valuable comments.

References

[1] (2024, February) Data-driven ui: unlimited power. [Online]. Available: https://mobius-piter.ru/en/2018/spb/talks/v96lokugwe8cwggio8ois/

[2] B. A. Myers and M. B. Rosson, "Survey on user interface programming," ser. CHI '92. New York, NY, USA: Association for Computing Machinery, 1992, p. 195–202. [Online]. Available: https://doi.org/10.1145/142750.142789

[3] M. E. Sadouski, "Semantic models and tools for designing adaptive user interfaces of intelligent systems," *Informatics*, vol. 20, no. 3, p. 74–89, Sep. 2023. [Online]. Available: http://dx.doi.org/10.37661/1816-0301-2023-20-3-74-89

[4] S. Abrahão, E. Insfran, A. Sluÿters, and J. Vanderdonckt, "Model-based intelligent user interface adaptation: challenges and future directions," *Software and Systems Modeling*, vol. 20, no. 5, p. 1335–1349, Jul. 2021. [Online]. Available: http://dx.doi.org/10.1007/s10270-021-00909-7

[5] A. Wolff, P. Forbrig, A. Dittmar, and D. Reichart, "Linking gui elements to tasks: Supporting an evolutionary design process," vol. 127, pp. 27–34, 01 2005.

[6] D. Shunkevich, "Agentno-orientirovannye reshateli zadach intellektual'nyh sistem [Agent-oriented models, method and tools of compatible problem solvers development for intelligent systems]," in *Otkrytye semanticheskie tekhnologii proektirovaniya intellektual'nykh system [Open semantic technologies for intelligent systems]*, V. Golenkov, Ed. BSUIR, Minsk, 2018, pp. 119–132.

[7] N. Zotov, "Semantic theory of programs in next-generation intelligent computer systems," in *Open semantic technologies for intelligent systems*. BSUIR, Minsk, 2022, pp. 297—-326.

[8] V. N. Lukin, A. L. Dzyubenko, and Y. B. Chechikov, "Approaches to user interface development," *Program. Comput. Softw.*, vol. 46, no. 5, p. 316–323, sep 2020. [Online]. Available: https://doi.org/10.1134/S0361768820050059

[9] M. G. Shishaev and V. V. Dikovickij, "Tekhnologiya sinteza adaptivnyh pol'zovatel'skih interfejsov dlya mul'tipredmetnyh informacionnyh sistem," *Trudy Kol'skogo nauchnogo centra RAN*, no. 5 (24), pp. 101–108, 2014.

[10] V. Golenkov, Ed., *Tehnologija kompleksnoj podderzhki zhiznennogo cikla semanticheski sovmestimyh intellektual'nyh komp'juternyh sistem novogo pokolenija [Technology of complex life cycle support of semantically compatible intelligent computer systems of new generation]*. Bestprint, 2023.

[11] M. Orlov, "Control tools for reusable components of intelligent computer systems of a new generation," *Open semantic technologies for intelligent systems*, no. 7, pp. 191–206, 2023.

[12] K. Bantsevich, "Metasystem of the ostis technology and the standard of the ostis technology," in *Open semantic technologies for intelligent systems*, ser. Iss. 6, V. Golenkov, Ed. BSUIR, Minsk, 2022, pp. 357–368.

[13] A. Zagorskiy, "Principles for implementing the ecosystem of next-generation intelligent computer systems," *Open semantic technologies for intelligent systems*, no. 6, pp. 347–356, 2022.

[14] V. Golenkov, N. Guliakina, and D. Shunkevich, *Otkrytaja tehnologija ontologicheskogo proektirovanija, proizvodstva i jekspluatacii semanticheski sovmestimyh gibridnyh intellektual'nyh komp'juternyh sistem [Open technology of ontological design, production and operation of semantically compatible hybrid intelligent computer systems]*, V. Golenkov, Ed. Minsk: Bestprint [Bestprint], 2021.

АДАПТИВНЫЕ ПОЛЬЗОВАТЕЛЬСКИЕ ИНТЕРФЕЙСЫ ИНТЕЛЛЕКТУАЛЬНЫХ СИСТЕМ: РАСКРЫТИЕ ПОТЕНЦИАЛА ВЗАИМОДЕЙСТВИЯ "ЧЕЛОВЕК-СИСТЕМА"

Садовский М. Е., Насевич П. Е., Орлов М. К., Жмырко А. В.

В работе проведен анализ возможностей компьютерных систем и уровня развития инструментов взаимодействия с ними (пользовательских интерфейсов). На основе анализа предложен подход к проектированию адаптивных пользовательских интерфейсов интеллектуальных систем на основе Технологии OSTIS. Уточнена и расширена предложенная ранее семантическая модель таких интерфейсов, приведена предлагаемая архитектура таких систем. Проектируемые на основе предложенного подхода адаптивные пользовательские интерфейсы обеспечат новые сценарии взаимодействия пользователей с компьютерными системами.

Received 01.04.2024

Natural Language Text Generation from Knowledge Bases of ostis-systems

Artem Goylo
Intelligent Semantic Systems
Minsk, Belarus
Email: artemgoylo@gmail.com

Sergei Nikiforov
Belarusian State University of Informatics and Radioelectronics
Minsk, Belarus
Email: nikiforov.sergei.al@gmail.com

Olga Golovko
Brest State Technical University
Brest, Belarus
Email: olechkavgolovko@gmail.com

***Abstract*—The article discusses an approach towards generating coherent natural language texts from fragments of ostis-system knowledge bases. The architecture of the abstract sc-agent of translating a fragment of the knowledge base into a natural language is described. The task of generating a natural language text is subdivided into three sub-tasks: structure filtering, knowledge base fragment decomposition, and generating an equivalent natural language text. Two ways of linearizing a knowledge base fragment are proposed: one based on a predefined order specification and an algorithmic one. Generation of an equivalent natural language text is proposed to be done in two stages. Firstly, an intermediate rough natural language representation is generated using rule-based mappings between sc-text constructions and their corresponding natural language verbalizations. Secondly, the intermediate representation is converted into a coherent natural language text with the help of a large language model. Finally, three possible applications of the proposed approach are described.**

***Keywords*—Natural Language Generation, Natural Language Processing, semantic network, Open Semantic Technology for Intelligent Systems (OSTIS), SC-code (Semantic Computer Code), knowledge base, discourse structure**

I. Introduction

During the last couple of years we have seen a sharp improvement of AI-related technologies. With the rise of Large Language Models (LLMs), automatically generated natural language texts have reached a never-before-seen level of adequacy and coherence. However, as it has been well-established, LLMs are prone to hallucinations and factual distortions when generating responses [1]. One of the potential solutions to this problem is using a reliable source of information to provide verified information as context for an LLM (for example, see [2]). Knowledge bases can serve as such a reliable source of information.

The OSTIS Technology (Open Semantic Technology for Intelligent Systems) [3] is a technology of complex support of the next-generation intelligent computer systems development life cycle. The technology is particularly focused on using knowledge bases as the core of intelligent computer systems, which are called ostis-systems in the context of the technology.

The foundation of the OSTIS Technology is a universal means of semantic representation (coding) of information in the memory of intelligent computer systems, called sc-code. Texts in sc-code (sc-texts and sc-constructions) are unified semantic networks that have a basic set-theoretic interpretation. The elements of such semantic networks are called sc-elements (sc-nodes and sc-connectors, which, depending on whether they are directed or not, are called sc-arcs or sc-edges). The universality and unified nature of sc-code allow to use it in order to describe all kinds of knowledge and problem-solving methods, which, in turn, considerably reduces the difficulty of integrating methods and knowledge within a system as well as within a collective of such systems.

As sc-texts are, essentially, semantic networks, it follows that they are non-linear in their nature. This fact posits a particular challenge for the natural language generation task, since an sc-text needs to be linearized before being translated into a natural language text.

The goal of this work is to outline a module for generating natural language texts based on both complete and self-contained fragments of a knowledge base, as well as arbitrarily chosen fragments. This module is envisioned as part of the natural language interface of an ostis-system, described in greater detail in our earlier work [4]. To achieve this, we will have to address the following issues:

- Filtering irrelevant parts of an sc-text stored in the knowledge base;
- Linearization of a non-linear text that is a certain graph structure within the knowledge base;
- Translation of a linearized sc-text into a natural language.

II. State of the art

Automatic natural language text generation is a well-researched problem, with many different approaches having been proposed for solving it.

Traditional approaches usually based their natural language generation systems on the rules which were developed using the various theories of discourse structure, in particular, Rhetorical Structure Theory [5], which establishes 25 relations that bind sentences within a text segment. Such approaches usually formalize texts as trees whose nodes are specific text segments that are partitioned into smaller segments connected via a particular relation, with leaves of such trees being particular clauses. An example of an RST-based approach is [6]. Tree-based representation of texts is in line with such foundational discourse theories as [7].

The issue with the formalisms used by the traditional approaches is that both discourse macrostructure, as well as relations between individual sentences, are quite flexible and allow for a certain degree of individual variation. Besides, the subject domain of rhetorical relations at the micro-level is not completely formalized, and such relations can be numerous [6, p. 17], which complicates the development of natural language generation systems. On the other hand, some theories emphasize the impossibility of a complete account of relations that hold between text segments and propose a limited set of basic relations that are usually structural, rather than semantic or rhetorical [8].

Of the traditional approaches to natural language generation, our approach, which will be described below, is most similar to [9]. This approach utilizes a formal model of discourse strategies to generate coherent natural language texts. Prior to text generation, knowledge is extracted from a knowledge base, after which its relevancy is established and irrelevant parts are removed. The approach focuses on generating a coherent text as a response to a certain question, which is why it needs to establish relevancy of knowledge found in the knowledge base. The authors define three types of questions: questions about information available in the knowledge base, questions about definitions, and questions about differences between entities in the knowledge base [9, p. 7]. Text is linearized according to the chosen discourse strategy, and there is no predefined discourse structure available in the knowledge base [9, p. 8]. Such an approach allows to generate variable surface structures describing the same knowledge representation due to the focus on discourse strategies [9, p. 9]. Discourse structure is formalized as specific patterns of usage of rhetorical predicates (overall 16 such predicates are defined, e.g. attributive, equivalent, specification, explanation, evidence, analogy, etc.)

Modern approaches focus on using the latest achievements in the field of AI, in particular, neural networks. Among them, one popular way of generating natural language texts is by using some intermediate semantic representation: for example, formalisms of Discourse Representation Theory [10], [11]. An approach that is similar to ours is the variant of translating RDF-triples into a natural language text, proposed in [12].

However, a significant number of works within such approaches focuses on sentence-level generation, rather than document-level generation. At the same time, approaches that emphasize document-level generation [10], [11], though acknowledging the task of text linearization (i. e. ordering of text segments), focus more on solving particular problems at the sentence-level connected with the chosen semantic formalism, such as variable naming.

Still other neural network-based approaches do not use intermediate semantic representation in natural language generation, see, for example, [13]. However, pure neural approaches sometimes have to contend with the issue that high-level semantic relations, which are important to capture when generating a coherent natural language text, present a challenge for neural networks [13, p. 1]

III. Suggested approach

We propose to address the issue of generating a coherent natural language text by adopting a multi-agent approach to the design of the respective module of natural language generation based on the OSTIS Technology.

The foundation of a knowledge base developed using the technology is a hierarchical system of semantic models of subject domains and ontologies. An ostis-system problem solver is represented by a collective of agents (sc-agents) that interact with each other by specifying the information processes in the semantic memory that they execute [14].

An abstract sc-agent is a certain class of functionally equivalent sc-agents, different instances (i. e. representatives) of which can be implemented in different ways. [14]

Below we will describe our suggested approach by providing the specifications of agents required for solving the problem of translating a fragment of the knowledge base into a natural language text.

The proposed implementation of the agent of translating a fragment of the knowledge base into a natural language has the following decomposition:

Abstract sc-agent of translating a fragment of the knowledge base into a natural language
⇒ *abstract sc-agent decomposition**:
  {• *Abstract sc-agent of structure filtering*
  • *Abstract sc-agent of fragment decomposition*
  • *Abstract sc-agent of generating an equivalent natural language text*
    ⇒ *abstract sc-agent decomposition**:
      {• *Abstract sc-agent of generating a rough version of a natural language text*
      • *Abstract sc-agent of converting the rough version into a correct natural language text*

}
}

The agents that are part of this abstract sc-agent will be discussed below. We will also discuss the agent of structure filtering that may be used in order to remove specified parts of a fragment before translating it into a natural language. The introduction of such an agent is explained by the fact that the structures that are to be translated (e. g., the semantic neighborhood of a certain concept) could be quite large and/or include irrelevant information, which would lead to a bloated natural language text that would be more difficult to comprehend.

A. Abstract sc-agent of structure filtering

The input of this agent is the structure to be filtered, and the output of it is a certain subset of the structure.

The filtering is done by using templates. Two kinds of template sets are proposed: the set of exclusive templates, i.e. the templates that are used to exclude a part of the structure before translation, and the set of inclusive templates, i. e. the ones that specify what part of the structure needs to be kept.

If a set of inclusive templates is passed to the filtering agent then only the corresponding part of the structure will be outputted. If exclusive templates are used then only the part that does not fit the template will be translated.

It is also necessary to introduce additional logic — the agent needs to check for connections between the elements of the part to be excluded and the part that is to be kept.

For example, let us consider the situation in figure 1.

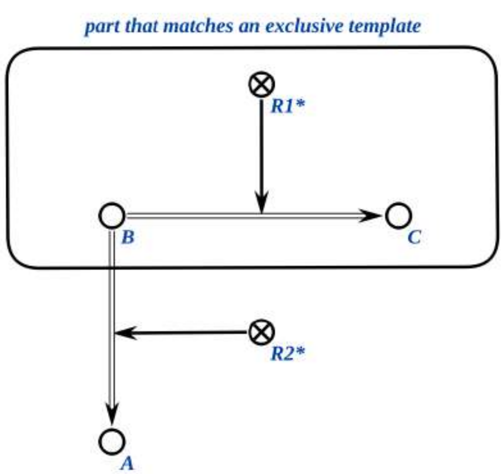


Figure 1. An example of a structure prior to filtering.

Here, it is expected that what will be left after filtering is the fragment of the structure in figure 2: even though the node *B* was part of the pattern found using an exclusive template, it still has a connection with the part that is to be included after filtering.

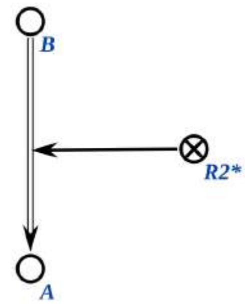


Figure 2. An example of the result of filtering.

B. Abstract sc-agent of fragment decomposition

The goal of this agent is decomposing a fragment of the knowledge base into an ordered set of substructures. The order of the substructures indicates the ordering of the content of each structure during after translation.

Classic works on discourse structure note that the structure of specific discourses depends on their genre (e. g., a story and a scientific paper will have different structures) [7]. Thus, different fragments of the knowledge base can have their own respective canonical structures. Hence it is important to discuss, which fragments of ostis-system knowledge bases are most likely to be translated into a natural language, and to formalize their structure.

The structure of a text depends on the class of the structure to be translated and the purpose of its translation, i. e. to what kind of message it can be used as a response.

The input of the agent is a structure to be translated, while the output is an ordered set of substructures that is the decomposition of the original structure.

This set of substructures is obtained in two stages:

- firstly, a predefined text structure specification is used;
- secondly, the ordering of elements within the substructures is derived algorithmically.

At the first stage, we propose to use a predefined specification of ordering of substructures stored in the knowledge base. The substructures are assumed to be the partition of the original structure (fragment) in the knowledge base. Such specifications can be defined for sc-texts of subject domains and other frequently used fragments of the knowledge base.

For example, we propose that a subject domain has a specification that includes the next elements in the following order: classes of objects of research (first, maximal classes and then non-maximal ones), explored relations, didactic relations (e.g. explanations and annotations), key signs, the list of segments that comprise the subject domain. After those elements comes the text of the segments themselves. The order of the segments is specified by the author of the subject domain specification, which eliminates the need to determine their order automatically.

Having such a specification will improve the quality of the resulting natural language text. However, in order to allow for translation of an arbitrarily defined fragment of the knowledge base, as well as to define the order of elements within the aforementioned substructures, an algorithmic way of deriving the order is needed.

However, the elements listed above (substructures of the original fragment of the knowledge base (in this case, the sc-text of a subject domain)) contain elements of their own, which also need to be linearized. Therefore, at the second stage, we derive the order of elements within such substructures based on the concepts contained in them. This process includes:

- Derivation of the order of concepts in the structure to be translated, i. e., the order in which their semantic neighborhoods should be translated;
- Derivation of the order of elements within such semantic neighborhood.

At the first step we propose to build a tree (graph) of dependencies between concepts according to the relations between them, and then to use this tree to derive the order of elements. For example, if a fragment has multiple classes of objects, then the first to be translated should be the supersets, followed by the subsets; sets should be translated before their elements, and so on.

At the second step (derivation of the order of elements within the semantic neighborhood of each concept) we propose to use a predefined order of relations and parameters. This will allow us to specify, for example, that when translating the semantic neighborhood of a concept the first elements to be translated should be the concept's identifiers, then its definition, then its membership in different sets followed by all the subsets of the given concept. There can be multiple potential variants of such a specification, depending on the class of the fragment.

We should note that this agent, and the structure specifications used by it, can be utilized not only for translating fragments of the knowledge base into a natural language, but also for translating them into other variants of linear representation of sc-code, for example, SCn.

C. Abstract sc-agent of generating an equivalent natural language text

As mentioned above, this agent in turn includes the following agents:

- Abstract sc-agent of generating a rough version of a natural language text
- Abstract sc-agent of converting the rough version into a correct natural language text

The input of the first agent is a structure to be translated, while the output is an ostis-system file with the resulting text. The text obtained as a result of this agent's execution may not fully correspond to the grammar of a particular natural language (in our case, English).

This approach explicitly sidesteps a much more complex task of microplanning (i.e. mapping of certain information in the semantic representation to the verbalization of this information) [12]. Instead of solving the problem of generating referring expressions, lexicalization, and so on, at this stage we propose to use a straightforward approach of using a finite set of specific rules of translating sc-code expressions into a natural language.

Currently, the agent has a simplified variant of implementation that is reduced to implementing a number of translators, each of which is dedicated for processing certain sc-code constructions (e. g., parameters of elements, their relations, etc.) Every construction has a corresponding natural language verbalization that is used during the translation. An example of a construction to be translated can be seen in figure 3.

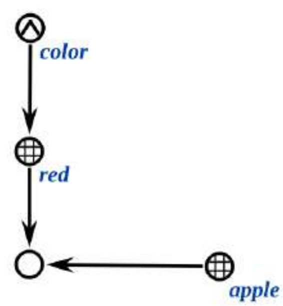


Figure 3. An example of a construction to be translated.

The membership arc corresponds to the English *is*, while the parameter in this case is translated as a pair (parameter, value). Therefore, the construction above will be translated by this agent into the following rough natural language text: *is apple, color red.*

This implementation has been chosen because it is relatively uncomplicated. In the future, we plan to elaborate it in way that the agent would arrive at a rough natural language verbalization algorithmically using formalization of natural language syntax proposed in our earlier work [15].

Complete rule-based algorithmic translation of knowledge base fragments into an adequate and coherent natural language text appears to be an infeasible task due the complexity of decision-making at each stage of the process. This is exemplified by the fact that the practice of designing fully rule-based intelligent systems has been largely supplanted by application of neural network-based solutions, which, in the case of large language models, outperform all currently available methods of automatic natural language text generation.

It is for this reason that we propose to introduce the *Abstract sc-agent of converting the rough version into a correct natural language text*. This agent is implemented using a large language model. Its input and output arguments are ostis-system files. The input file contains the original rough natural language text that needs to be transformed, and the output file contains the text generated by a large language model.

Using a large language model is a convenient alternative to rule-based generation because it allows to sidestep certain sub-tasks of generating a coherent natural language text, such as choosing particular means of cohesion and coherence, which allows us to reduce the task to forming an ordered set of substructures that sets the order of segments of a coherent text, while individual verbalization choices are made by the large language model, which they in general excel at [16].

Using an intermediate representation for now also increases the likelihood of obtaining acceptable results without language model hallucinations [1], since the model in this case is not utilized in a zero-shot scenario and is provided with an extensive context that has a formal nature.

Thus, to generate the resulting natural language text we propose to use and intermediate representation (the output of the agent of generating a rough version of a natural language text), which is necessary at this stage because existing neural network solutions cannot be directly integrated with knowledge bases of ostis-systems. In the future, the OSTIS Technology will have support for "native" representation of neural networks as well as the means of preprocessing the input for traditional neural networks in such a way as to enable them to handle sc-code constructions [17]. This will eliminate the need for translating fragments of the knowledge base into intermediate variants of representation, and will enable us to use the actual sc-text of a knowledge base fragment as input for a large language model.

IV. Potential applications

Finally, we would like to discuss potential applications of the natural language generation module described above. These are three main ways in which it can be used:

- Exporting an arbitrary fragment of the knowledge base in a natural language;
- Navigating the knowledge base in a natural language;
- Dialog with an ostis-system using a natural language.

A. Exporting an arbitrary fragment of the knowledge base in a natural language

In this scenario, the fragment to be exported is specified by the user manually. For this application, the corresponding ostis-system can support various existing natural language text formatting styles.

One potential benefit of translating arbitrary fragments of the knowledge base into a natural language is that it makes it possible to use knowledge bases appropriately as the primary means of storing knowledge. Whereas, traditionally, knowledge has been stored mostly in natural language texts of various kinds, having a system that allows to translate formalized representations of knowledge into natural language texts on demand will significantly help with complex automation of various types of human activity [18].

B. Navigating the knowledge base in a natural language

The main way to navigate he current OSTIS Metasystem interface [19] is by navigating semantic neighborhoods of elements and/or other constructions. The external languages of sc-code representation used for outputting the content of the Metasystem's knowledge base are SCn and SCg [3].

It is possible to introduce a new way of navigating knowledge bases of ostis-systems whereby the fragments are translated into a natural language, which makes it possible to interact with ostis-systems effectively for users who are unfamiliar with the languages of external representation of sc-code.

This application would require additional work on the translation module in order to allow for hyperlinks within the natural language text markup, which will make the navigation easier.

C. Dialog with an ostis-system using a natural language

We plan to provide for the ability to communicate with an ostis-system using a natural language by implementing a question-answering support subsystem for users of the OSTIS Metasystem [19]. This subsystem should allow the user to ask questions about any knowledge stored in the Metasystem's knowledge base and get a response in a natural language.

The pipeline of this subsystem can be decomposed into the following stages:

- Message classification and question argument identification;
- Response generation;
- Translation of the response into a natural language using the means described above.

During the ongoing implementation of the prototype of this subsystem we have decided to use one of the existing neural network-based classifiers for the task of message classification and question argument identification: Rasa [20], Wit.AI [21], and others. We consider

Rasa to be the preferable option due to the possibility of local deployment and its open-source nature.

This approach has been chosen in order to obtain quickly a working prototype of the system. In the future, neural network-based classifiers can be replaced with an sc-agent of natural language understanding based on the approach discussed in [15].

The input of the response generation agent is a message that has been classified, while the output is a structure from the knowledge base that is an appropriate response to the message. An example of message classification received by the agent is available in figure 4.

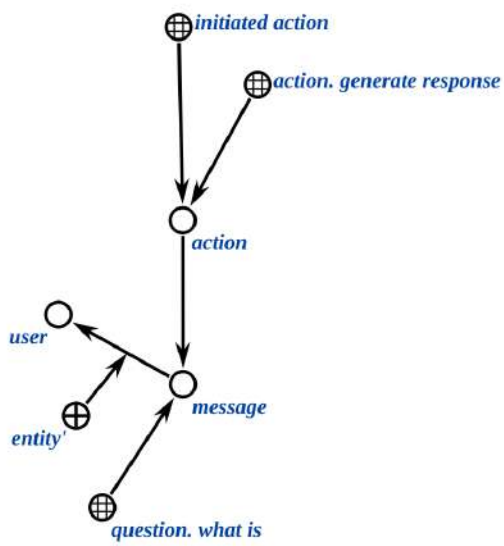


Figure 4. An example of message classification.

The agent operates in two steps:

- firstly, it tries to formulate a response using search templates;
- secondly, it tries to formulate a response by executing an appropriate action, in case the first step was unsuccessful.

The first step is introduced because responding to certain user questions can be reduced to searching for a relatively simple construction in the knowledge base, which can be implemented by mapping the corresponding classes of questions to certain search templates, as well as mapping message arguments to variables contained in such search templates. The response is a structure that contains the result of search by a template that corresponds to a certain class of questions after variables have been replaced with the corresponding question arguments.

Extending the set of questions that can be answered using search templates is an uncomplicated task that does not require modifying the problem solver. This task is reduced to introducing a new search template and specifying its connection with a class of questions and its arguments.

However, such constructions may be difficult to describe using one search template, or the answer may not be reducible to simple search and may require detailed transformations. For this reason, the second step is necessary.

If formulating a response using search templates is impossible, then the system searches for classes of actions connected to the corresponding class of questions by the relation *response action**. An instance of such action is then created with the corresponding argument received in the question.

Let us illustrate this using the question *What is X?* as an example. The response to such questions is a description of a certain element in the knowledge base, i. e. its semantic neighborhood. An example of the connection between a class of actions and a class of questions described above can be seen in figure 5.

In order to handle questions with two or more arguments that have different roles, the roles in the message can be mapped to respective arguments of actions in the knowledge base.

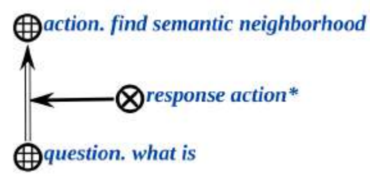


Figure 5. An example of the mapping between a class of actions and a class of questions.

Thus, when the agent receives the message illustrated in figure 4, an instance of the action of finding a semantic neighborhood is created. Then the problem solver waits until the agent is executed, and the agent's response is then connected with the message. An example of a construction obtained in this way can be seen in figure 6.

V. Conclusion

We have provided a sketch of the architecture of a module for translating fragments of ostis-system knowledge bases into coherent natural language texts. Our proposed approach subdivides the task of generating a natural language text into three sub-tasks: structure filtering, knowledge base fragment decomposition, and generating an equivalent natural language text.

The most important sub-task is knowledge base fragment decomposition because it ensures cohesion and coherence of the resulting natural language text. We have proposed two preliminary ways of solving this task: specification of element ordering within a fragment of the knowledge base as a sort of schema of the overall structure of the resulting natural language text, and

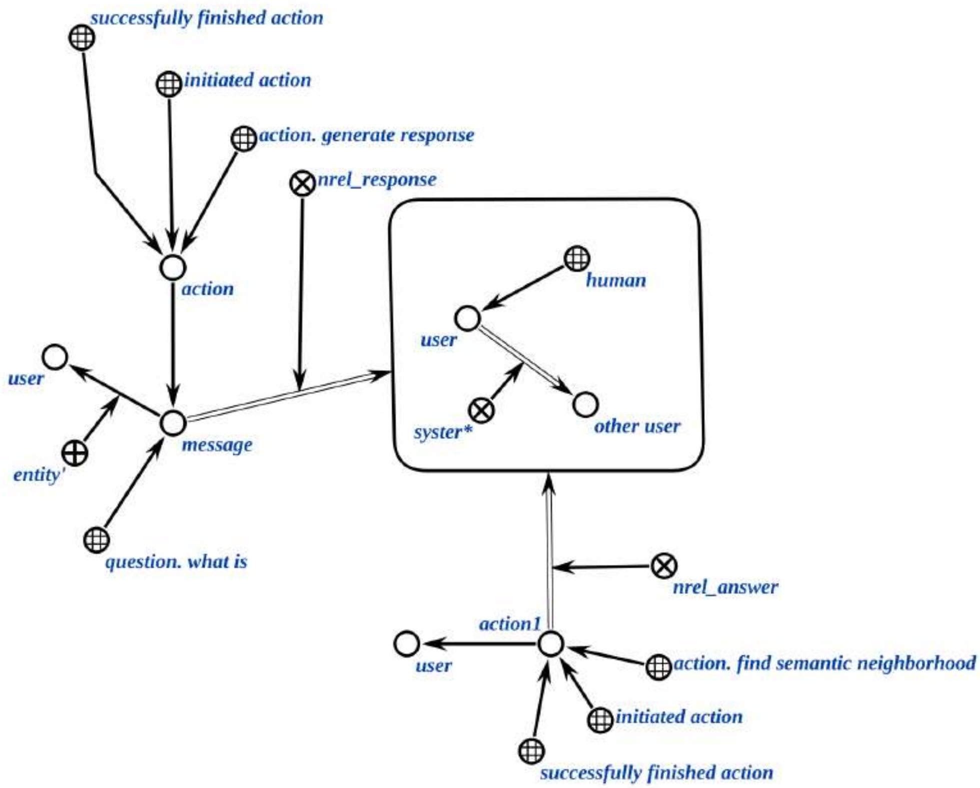


Figure 6. An example of the output of the response generation agent.

algorithmic approach that builds a tree of dependencies between certain relations.

The actual generation of a natural language text is proposed as a two-step process, whereby a large language model generates the resulting text from an intermediate representation.

Due to the preliminary character of our work, there are certain limitations. Our approach does not discuss linearization of graph-based formal texts in greater detail apart from positing relatively straightforward schemata to be used during translation. In fact, given that the ultimate application of the module discussed here is natural language dialog between humans and intelligent systems, our proposed approach can be improved in three different ways:

- Firstly, understanding natural language questions to the system can be done using not a simple classifier but rather a combination of syntactic and semantic analysis modules that use a formalization of natural language syntax and semantics.
- Secondly, the linearization task can be solved in a much more elaborate manner. This would require formalization of a discourse structure model within the knowledge base of an ostis-system. The model can then be used to intelligently derive macro- and microstructures of sc-texts to be translated into a natural language.
- Finally, the actual translation of linearized sc-texts into a natural language needs further elaboration. An obvious improvement is to eliminate specific rules (mappings) of translating sc-constructions into certain predefined natural language verbalizations, which would require designing a proper natural language synthesis module.

All of the above can be considered as potential directions for future research.

Acknowledgment

The authors would like to thank the research teams of the Department of Intellectual Information Technologies of Belarusian State University of Informatics and Radioelectronics and the Department of Linguistics Disciplines and Cross-Cultural Communication of Brest State Technical University for their help and valuable comments.

References

[1] T. Yu, S. Zhang, and Y. Feng, “Truth-aware context selection: Mitigating the hallucinations of large language models being misled by untruthful contexts,” 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:268364338

[2] Y. Wu, N. Hu, S. Bi, G. Qi, J. Ren, A. Xie, and W. Song, “Retrieve-rewrite-answer: A kg-to-text enhanced llms framework for knowledge graph question answering,” 2023.

[3] V. Golenkov, Ed., *Tehnologija kompleksnoj podderzhki zhiznennogo cikla semanticheski sovmestimyh intellektual'nyh komp'juternyh sistem novogo pokolenija [Technology of complex life cycle support of semantically compatible intelligent computer systems of new generation].* Bestprint, 2023.
[4] A. Goylo and S. Nikiforov, "Natural language interfaces of next-generation intelligent computer systems," *Open semantic technologies for intelligent systems*, no. 6, pp. 209–216, 2022.
[5] W. Mann and S. Thompson, "Rethorical structure theory: Toward a functional theory of text organization," *Text*, vol. 8, pp. 243–281, 01 1988.
[6] E. H. Hovy, "Automated discourse generation using discourse structure relations," *Artificial Intelligence*, vol. 63, no. 1, pp. 341–385, 1993. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0004370293900213
[7] W. Kintsch and T. A. van Dijk, "Toward a model of text comprehension and production." *Psychological Review*, vol. 85, pp. 363–394, 1978. [Online]. Available: https://api.semanticscholar.org/CorpusID:1825457
[8] B. J. Grosz and C. L. Sidner, "Attention, intentions, and the structure of discourse," *Computational Linguistics*, vol. 12, no. 3, pp. 175–204, 1986. [Online]. Available: https://aclanthology.org/J86-3001
[9] K. R. McKeown, "Discourse strategies for generating natural-language text," *Artificial Intelligence*, vol. 27, no. 1, pp. 1–41, 1985. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0004370285900827
[10] J. Liu, S. Cohen, and M. Lapata, "Text generation from discourse representation structures," 01 2021, pp. 397–415.
[11] C. Wang, R. van Noord, A. Bisazza, and J. Bos, "Evaluating text generation from discourse representation structures," in *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, A. Bosselut, E. Durmus, V. P. Gangal, S. Gehrmann, Y. Jernite, L. Perez-Beltrachini, S. Shaikh, and W. Xu, Eds. Online: Association for Computational Linguistics, Aug. 2021, pp. 73–83. [Online]. Available: https://aclanthology.org/2021.gem-1.8
[12] C. Gardent, A. Shimorina, S. Narayan, and L. Perez-Beltrachini, "The WebNLG challenge: Generating text from RDF data," in *Proceedings of the 10th International Conference on Natural Language Generation*, J. M. Alonso, A. Bugarín, and E. Reiter, Eds. Santiago de Compostela, Spain: Association for Computational Linguistics, Sep. 2017, pp. 124–133. [Online]. Available: https://aclanthology.org/W17-3518
[13] J. Guan, X. Mao, C. Fan, Z. Liu, W. Ding, and M. Huang, "Long text generation by modeling sentence-level and discourse-level coherence," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, C. Zong, F. Xia, W. Li, and R. Navigli, Eds. Online: Association for Computational Linguistics, Aug. 2021, pp. 6379–6393. [Online]. Available: https://aclanthology.org/2021.acl-long.499
[14] D. Shunkevich, "Hybrid problem solvers of intelligent computer systems of a new generation," *Open semantic technologies for intelligent systems*, no. 6, pp. 119–144, 2022.
[15] A. Goylo and S. Nikiforov, "Means of formal description of syntax and denotational semantics of various languages in next-generation intelligent computer systems," *Open semantic technologies for intelligent systems*, no. 6, pp. 99–118, 2022.
[16] L. Shu, L. Luo, J. Hoskere, Y. Zhu, Y. Liu, S. Tong, J. Chen, and L. Meng, "Rewritelm: An instruction-tuned large language model for text rewriting," 2023.
[17] M. Kovalev, A. Kroshchanka, and V. Golovko, "Convergence and integration of artificial neural networks with knowledge bases in next-generation intelligent computer systems," *Open semantic technologies for intelligent systems*, no. 6, pp. 173–186, 2022.
[18] V. V. Golenkov, N. A. Gulyakina, and D. V. Shunkevich, "Klyuchevye problemy i strategicheskie tseli rabot v oblasti iskusstvennogo intellekta [key problems and strategic goals of research in the field of artificial intelligence]," 2023, pp. 9–13.
[19] "ostis-metasystem repository," Available at: https://github.com/ostis-ai/ostis-metasystem, (accessed 2024, March).
[20] "Rasa," Available at: https://rasa.com/, (accessed 2024, March).
[21] "Wit AI," Available at: https://wit.ai/, (accessed 2024, March).

ГЕНЕРАЦИЯ ЕСТЕСТВЕННО-ЯЗЫКОВЫХ ТЕКСТОВ ИЗ БАЗ ЗНАНИЙ OSTIS-СИСТЕМ

Гойло А. А. Никифоров С. А. Головко О. В.

В статье описывается подход к генерации связных текстов на естественном языке из фрагментов баз знаний ostis-систем. Описана архитектура абстрактного sc-агента трансляции фрагмента базы знаний в естественный язык. Задача генерации естественно-языкового текста подрязделяется на три подзадачи: фильтрация структуры, декомпозиция фрагмента базы знаний и генерация эквивалентного естественно-языкового текста. Предлагаются два способа линеаризации фрагментов базы знаний: использование заданной спецификации порядка элементов фрагмента и алгоритмический. Предлагается выполнять генерацию эквивалентного естественно-языкового текста в два этапа. На первом этапе формируется черновое естественно-языковое представление на основе правил сопоставления конструкций sc-текстов с соответствующими им естественно-языковыми формулировками. На втором этапе такое промежуточное представление транслируется в связный естественно-языковой текст с использованием большой языковой модели. Описываются три возможных применения предлагаемого подхода.

Received 18.03.2024

Principles of Building Intelligent Robotic Systems

Aliaksandr Kroshchanka
Brest State Technical University
Brest, Belarus
kroschenko@gmail.com

Mikhail Kovalev
Belarusian State University of Informatics and Radioelectronics
Minsk, Belarus
michail.kovalev7@gmail.com

***Abstract*—The paper proposes a concept for the building of collaborative robotic systems using OSTIS technology. The developed concept is based on the integration of robotic, symbolic and neural network components. The main provisions of the approach are illustrated by the project of a robotic system for sorting objects with specified characteristics. Recommendations are given on the application of the proposed concept for the construction of collaborative robotic systems in the context of the development of new generation intelligent computer systems based on the use of OSTIS technology.**

***Keywords*—OSTIS, collaborative robotics, hybrid intelligent systems, object detection, manipulators**

I. Introduction

In modern collaborative robotic systems, robots follow a set algorithm of actions, including the performance of some predefined operations (e. g., grasping and moving an object, positioning at a certain point, performing a certain operation on an object). Technically, the realization of such actions does not cause difficulties in case of an ideal workflow. However, there may be situations when there are deviations from the established algorithm of actions, for example, absence of an object in a given point by the beginning of the operation, wrong type of object or impossibility to perform the operation due to blocking of moving parts, appearance of unauthorized persons in the area of manipulator operation, etc. These problems can be solved by using machine learning methods, in particular, neural networks. For example, a detector network allows you to determine the type of object moving along the conveyor, another model calculates the position of the manipulator at the next moment of time depending on the technical operation being performed, etc. However, the use of only specialized auxiliary models, for example, a computer vision model for detecting missing objects, will not be able to help in the correct identification of the place where objects of a given type can be located and where, in case of absence of the object on the conveyor, it will be necessary to deploy the manipulator to grab the part. Information about important aspects of the manufacturing process, which is sufficiently variable, needs to be systematically stored because the alternative — the need for constant direct code editing in the context of changes occurring to the robot — is unacceptable and, moreover, can often lead to errors. In addition, the experience that such systems may acquire during their operation is also clearly important and needs to be properly represented for reuse in other contexts and processes.

Thus, the development of principles and recommendations for the construction of intelligent robotic systems is an urgent topic of research, because it allows to streamline the process of developing such systems. The use of modern tools for designing intelligent systems, which undoubtedly includes OSTIS technology, allows the representation and operation of knowledge, which is a valuable resource in robotic systems.

The subsequent sections of the paper are organized as follows: section II sets out the problem of building knowledge-based robotics systems, in the same section an overview of existing solutions is given; section III describes the proposed concept for building intelligent robotics systems; section IV describes the developed prototype of an intelligent robotics system for sorting objects of a given type; finally, section V summarizes the main conclusions of the proposed concept and describes the main conclusions of the intelligent robotics system for sorting objects of a given type.

II. Problem formulation

The idea of developing robotic systems based on the use of knowledge bases has been widely investigated in various works. For example, in [1], the authors give an overview of knowledge bases used in robotic systems to find missing tools. Emphasis is placed on finding those objects without which further workflow continuation will not be possible. However, other applications of knowledge bases are not indicated, in particular, the possibility of using them for robot self-diagnosis, which is an important function of such systems. Other studies use knowledge bases (e. g., Cyc [2] or SUMO [3]) that are not specific to robotics, which makes it difficult to use such solutions in practice. One of the most promising solutions at the moment is the KnowRob KB ([4], [5]), which is characterized by a developed ontological basis for robotics.

In the context of the above-mentioned works, we have formulated the task of developing a concept for the construction of intelligent robotic systems based on the use of knowledge, as well as outlined the basic requirements for such systems:

- support of heterogeneous components of the robot system, i. e. support of a single open interface of interaction;
- the possibility of transferring the knowledge accumulated by the system during its operation to other robotic systems with minimal changes;
- adaptive design, i. e. the ability to change the composition of the system components without having to change the interaction logic;
- support for self-modification of the system;
- the ability of the system to expand and/or improve its set of sensors and effectors;
- the ability of the system to analyze the quality of its physical and software components.

These requirements correspond to the properties of cybernetic systems given in the OSTIS technology standard [6]. This technology is a reasonable choice for designing an intelligent robotic system, as it ensures the achievability of these properties.

In the process of building the concept of developing systems of this type, it is necessary to form a list of recommendations and general rules for designing intelligent robotic systems, and to realize an applied intelligent robotic system on the basis of the outlined theoretical provisions.

III. Proposed concept

The proposed robotic system design concept is developed using OSTIS technology.

The fundamental possibility of integrating a physical robotic system and OSTIS in the context of controlling this system is based on the developed ontology that includes a description of the basic physical components of such a system (i. e., manipulators, transporters, etc.), as well as the classes of actions that can be performed by such components.

As a formal basis for knowledge representation within the framework of OSTIS Technology, a unified semantic network with a set-theoretic interpretation is used. This representation model is called SC-code (Semantic Computer code). The elements of the semantic network are called sc-nodes and sc-connectors (sc-arc, sc-edges) [6].

Systems built on the basis of OSTIS Technology are called ostis-systems. Any ostis-system consists of a knowledge base, a problem solver and a user interface. The basis of the knowledge base is a hierarchical system of subject domains (SDs) and their corresponding ontologies. Ontologies contain descriptions of concepts necessary for formalization of knowledge within SD. Any knowledge describing some problem, its context and specification of solution methods can be represented in the form of SC-code constructs. Thus, unification of representation and consistency of different types of knowledge describing problems, their context and solution methods is ensured.

Benefits that can be achieved by using OSTIS as a design tool for intelligent robotic systems include:

- the possibility of isolating extracted knowledge, independent of the manipulator types used, and reapplying it to other robotic systems under development, eliminating the need to code single-type operations;
- convenient means of visualization, for example, with the use of SCg, which allows to determine the working conditions of the components of robotic systems;
- use of open interaction interfaces that allow adding other physical components on the fly;
- explainability of the system operation modes, which allows tracking the occurrence of emergency situations with the formation of a detailed, human-understandable report.

IV. Robotics system design

According to the proposed concept, we have carried out the design and development of a collaborative robotic intelligent system.

The main purpose of the developed system is to sort objects of a certain type while maintaining the ability to flexibly modify the filtering condition of objects. This type of robotic systems is popular and widely used in production conditions to select objects with certain properties (for example, to reject manufactured products or to organize them for subsequent packaging of only the same type of goods). Such an operation, if automated, significantly simplifies manual labor in production, reducing the amount of monotonous work performed.

The physical part of the system consists of the following components:

- manipulators (2 units);
- transporter;
- single-board computer;
- storage devices (general and for target objects);
- tube;
- camera;
- ultrasonic sensor;
- power supply;
- indicator lamp;
- auxiliary components such as conductors, relays, voltage converters, and so on.

The scheme of the main physical components of the system is shown in Fig. 1.

The main components of the software part of the proposed robotic system, as well as of any ostis-system, are **knowledge base** and **problem solvers**. The user

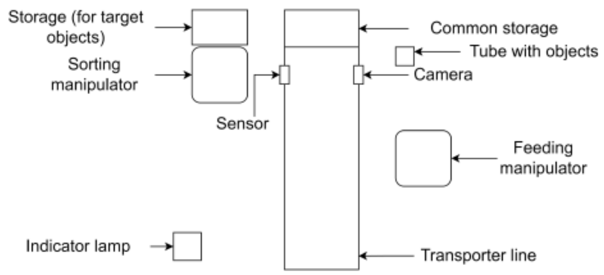


Figure 1. The scheme of the main physical components of the system

interface is the standard OSTIS technology tools for viewing and editing knowledge bases. The computer vision module, including a neural network model for object detection, is also a program component.

The functioning scenario of the developed system is reduced to the following main actions:

1) picking up the object from the tube by the feeding manipulator and moving it to the beginning of the conveyor;
2) switching on the conveyor and moving the object until the sensor is triggered;
3) disconnection of the conveyor at the moment of sensor actuation;
4) recognition of the object (type and color) by means of the installed camera;
5) moving the object from the conveyor belt to the target object storage by means of the sorting arm, if the recognized type and color match the set type and color;
6) switching on the conveyor belt and moving the object to the general storage;
7) switching off the transporter after the fulfillment of item 6.
8) switching on the green signal of the indicator when there are objects in the tube;
9) switching on the red indicator signal when there are no objects in the tube.

Let's describe the physical components of the system in more detail.

A. Physical components

Manipulators are used for gripping and moving objects. In this project, we used manipulators with different types of grippers that are widely available on the market — mechanical (pincer) grippers (Fig. 2) and vacuum grippers (Fig. 3).

Transporter is intended for moving objects to the specified point of technological operation (Fig. 4).

Single Board Computer — a specialized computer on which the OSTIS platform is deployed and the logic for controlling the system operation is implemented. In our implementation, the SBC Raspberry PI 5 (Fig. 5) [7] was used for this purpose.

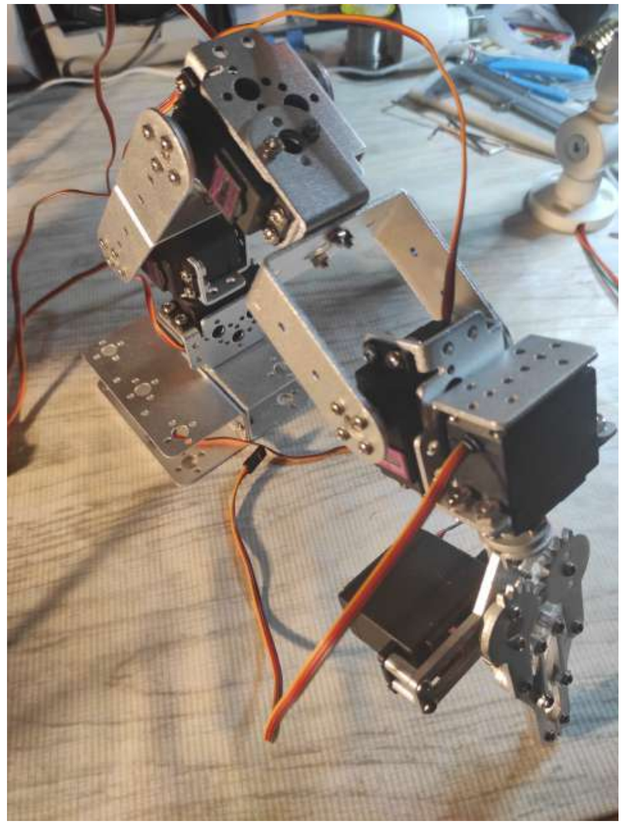


Figure 2. Manipulator with mechanical gripper

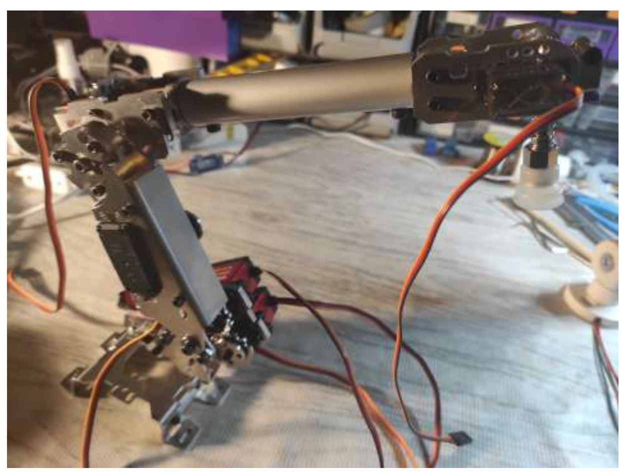


Figure 3. Manipulator with vacuum gripper

Storages — special containers used to store objects of a certain type (Fig. 6).

Tube – a container for storing objects that is shaped to be gripped by a manipulator (Fig. 7).

Camera is used to detect objects in the field of view and determine their characteristics. We used a 2 megapixel FullHD backlit camera ZONE 51 LENS.

Ultrasonic sensor is used to identify situations in which the object is in a given point of the conveyor. For

Figure 4. Transporter part

Figure 5. SBC Raspberry PI 5 with heat sink installed

our project we used the HC-SR04 sensor (Fig. 8) [8].

Power supply is required to power all physical devices in the system. We used a 360-watt, 24-volt, 15-amp power supply.

Indicator lamp is intended for light indication of the system status. In our project we used a TD-50 lamp with two color options (red and green) (Fig. 9).

B. Program components: knowledge base

The ontological approach is used for knowledge structuring, the essence of which is to represent the knowledge base as a hierarchy of subject domains and their corresponding ontologies. OSTIS technology provides a basic set of ontologies on the basis of which ontologies of applied ostis-systems are built.

Figure 6. Storage for objects with placed objects

Figure 7. Fragment of the tube with placed objects

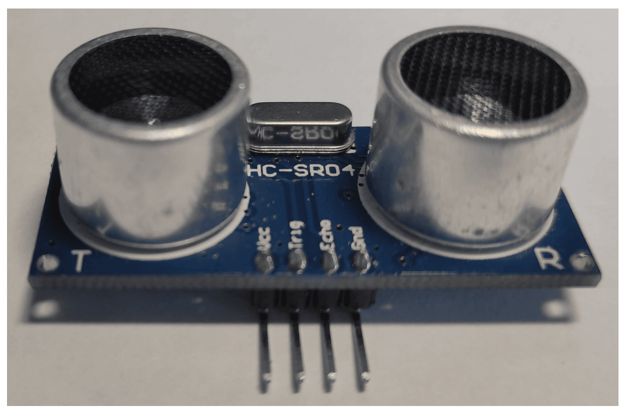


Figure 8. Ultrasonic sensor

Figure 9. TD-50 indicator lamp

The following subject domains are identified for the considered intelligent robotic system:

- Subject domain and its corresponding robotic device ontology;

- Subject domain and its corresponding ontology of physical objects — is a child of the Material Entity SD.

The robotic device SD describes the specification of the components of the physical part of the system listed above, and the specification of the actions that the system can perform with or through them.

The SD of physical objects specifies properties of the objects recognized by the system, such as color, shape, volume, weight, etc.

As the system scales up to automate more and more complex processes (line operation, shop floor operation, operation of the entire enterprise), these subject domains will be broken down into subsidiary SDs, e.g. computer vision SDs, robotic arm SDs, etc. Such development of subject domains in the field of production automation has already been done within the framework of OSTIS [9] technology.

Figure 10 shows the formalization of the system's knowledge of its own physical part in SCg-code. This formalization is necessary for the system to perform the sorting operation.

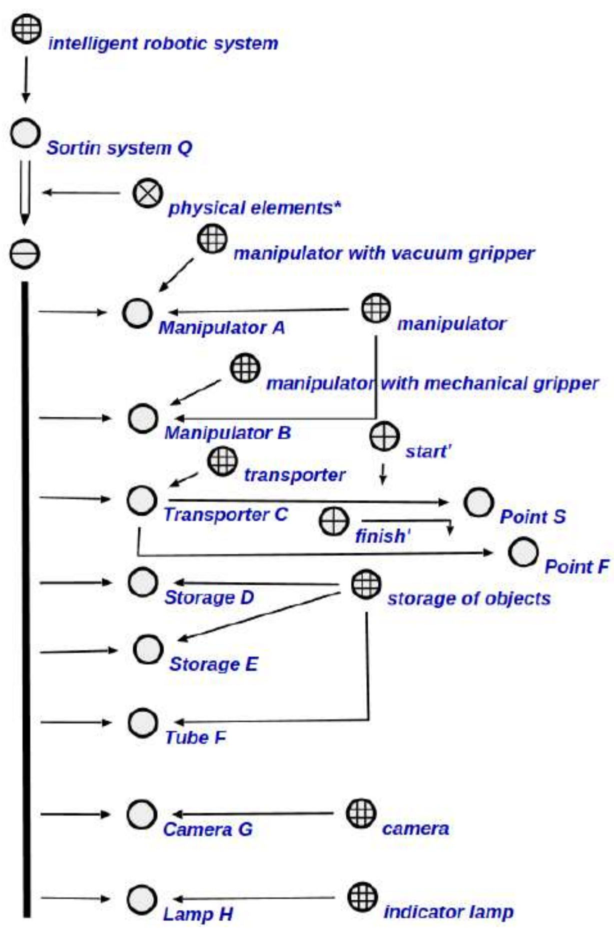


Figure 10. Formalization of the system's knowledge about its own physical part in SCg-code

According to the entity naming convention adopted in the OSTIS [6] technology standard, specific elements of the system are named with a capital letter. For convenience, each specific element is named with a letter:

1) Manipulator A is a sorting arm with vacuum gripper.
2) Manipulator B is a feeder arm with mechanical gripper.
3) Transporter C — A transporter having a starting point S as a plurality of objects placed on the transporter by the feeding arm and a point F as a plurality of objects moved by the transporter until the sensor is triggered.
4) The Storage D is a storage of target objects represented by a plurality of target objects in the storage.
5) The Storage E is a storage of objects that are not in Storage D.
6) Tuba F — an object storage represented by a set of unsorted objects.
7) and other.

C. Program components: problem solver

The problem solver deals with processing of knowledge base fragments, which is reduced to adding, searching, editing and deleting sc-nodes and sc-connectors of the knowledge base. At the semantic level, such operations are actions performed in the memory of the subject of the action, where, in general, the subject is the system itself, and the knowledge base is its memory.

The specification of an action in the knowledge base describes what should be done, with what, for what, by whom, etc., but the interpretation of actions according to this description is performed by agents. The problem solver of each ostis-system is based on a multi-agent system whose agents interact with each other only through a common knowledge base [10]. Each agent expects some event to occur in the knowledge base. For example, the appearance of a new specification of the action it should perform. When the event occurs, the agent performs the action and places its result in the same knowledge base.

The ostis-system problem solver has methods and tools to divide problems into subtasks and is able to explain its solutions at the level of describing the wording of subtasks and the sequence of their solution. When solving a problem, the problem solver breaks it into subtasks with an explicit description of the wording of each subtask, searches for a method of its solution and applies it.

Partitioning into subtasks begins with analyzing the goal of solving the problem set for the system. Formally, the goal of the developed system can be formulated as follows: for $\forall x$ such that $x \in$ the set of physical objects and $x \in$ the set of objects in Tube F, infer either $x \in$ the set of objects of Accumulator D and $x \in$ the set of target objects, or $x \in$ the set of objects of Accumulator

E and $x \notin$ the set of target objects. Figure 11 shows the formalization of the goal of the object sorting problem in SCg-code.

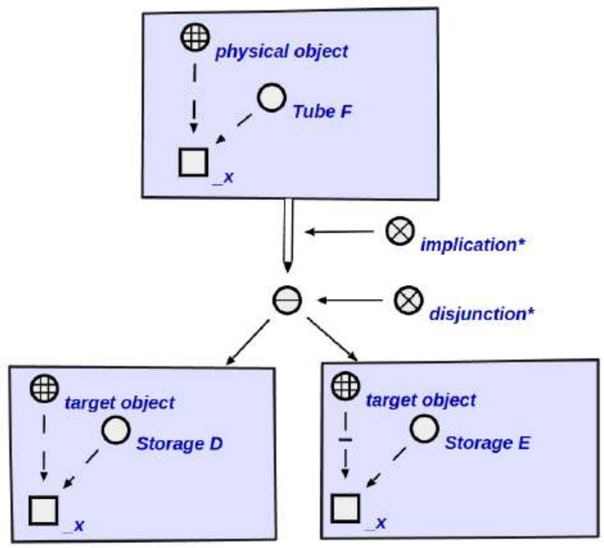


Figure 11. Formalizing the goal of solving the problem of sorting objects in SCg-code

The process of achieving the goal, depending on the system settings, can go in two directions: directly, from the initial situation, and backwards, from the target situation.

Let's consider the case of solution search from the final goal. The task solution search agent breaks the target situation into elementary sc-constructions, so-called triples and fives, and searches the system for actions or logical statements, the application of which can form the necessary part of the target situation in the knowledge base. For example, the specification of **action of moving an object to storage D** states that after its execution the object can get into storage D. However, the initial situation for applying this action is the presence of this object at the point F of the transporter and its belonging to the class of target objects. This initial situation becomes a new target of the solution search agent and the process is repeated for this target. New actions and logical statements are searched for.

Figure 12 shows the specification of the initial and final situation of the action of moving an object to storage D in SCg-code.

In the developed system, the following actions are used to solve the problem of sorting objects (given in the order in which they are found by the problem solving search agent):

1) The action of moving an object to Storage D. An object belonging to the target class and belonging to the set of objects at Point F is moved to the D storage unit. The agent interpreting this action controls Manipulator A.

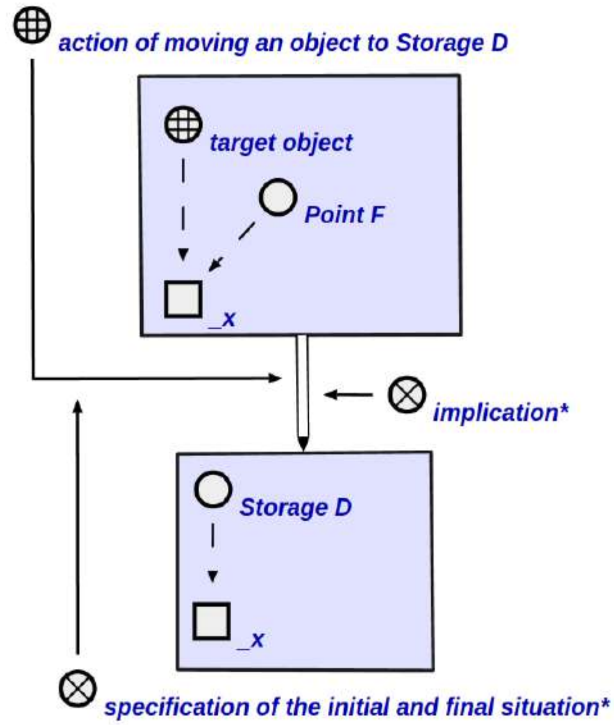


Figure 12. Specification of the start and end situation of the action of moving an object to storage D in SCg-code

2) The action of moving an object to Storage E. An object that does not belong to the target class and belongs to the set of objects at Point F is moved to Storage E. The agent interpreting this action turns on the transporter Storage E. The agent interpreting this action turns on the transporter for a predetermined amount of time.
3) The action of delivering an object from Point S to Point F. The object belonging to the set of objects at Point S ceases to belong to it and begins to belong to the set of objects at Point F. The agent interpreting this action turns on the transporter until the sensor is triggered.
4) The object classification action. The object belonging to the set of objects at Point F starts to belong or not to the set of objects of the target class. The agent interpreting this action finds a specification of the target class in the knowledge base (see figure 13) and sets this specification as the goal of a new problem solved by the agent according to the principle described above. During the search, the agent applies actions 5), 6) and 7).
5) The action of classifying an object in the image. The class of the object for which its image is specified is determined. The agent interpreting this action uses the methods described in the computer vision module to recognize the object in the image.
6) The action of determining the color of the object in

the image. The agent interpreting this action uses the methods described in the computer vision module.
7) The action of searching for objects in the image. The coordinates in the image obtained from the camera at the moment of sensor triggering are selected for the object. The agent interpreting this action uses the methods described in the computer vision module.
8) The action of moving the object onto the Point S. An object belonging to the set of objects in Tube F is moved to Point S. The agent interpreting this action controls Manipulator B.

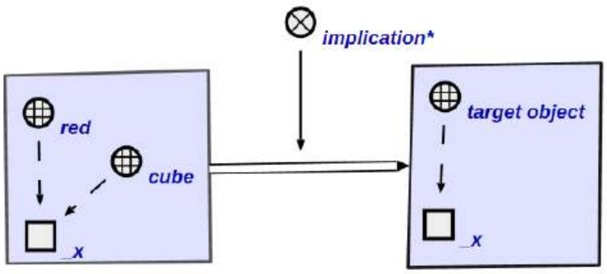


Figure 13. Specification of target class definition in SCg-code

At this point, the agents that interpret these actions use the program interfaces of the physical elements. For example, the agent for moving the object to the accumulator D implements an algorithm that uses the program interface of the manipulator A to transfer to it the rotation angles of its servos for moving the object from the conveyor to the accumulator in rigidly fixed locations. The system development plans include formalization in the knowledge base of the manipulator itself so that the rotation angles of the manipulator servo drives and the order of actions to change them are set depending on the current situation in the working area, rather than according to a predetermined algorithm.

The applied approach allows intelligent robotic systems to solve problems in a declarative way, when the order of application of methods of problem solving is formed by the system independently on the basis of the problem condition. Thus, adding new stages of object processing or methods for determining new characteristics of objects (weight, presence of certain digits, etc.) requires formalization of specifications of initial and final situations of actions and implementation of agents interpreting them, but does not require additional code for integration of these agents. Thus, the design of robotic systems is reduced to the description of the physical part and the tasks to be solved by this system without the need to program the solution of each task.

D. Program components: computer vision module

To solve the subproblem of detection of the given objects, the authors have formed a dataset of images of geometric bodies obtained by 3D printing. This sample includes photographs of three geometric bodies (cube, cylinder, cone) taken from different angles and with different camera angles. The total dataset size was 200 images, which were manually labeled. Examples of photos from the dataset, as well as photos from different angles of the same body, are shown in Figures 14 and 15.

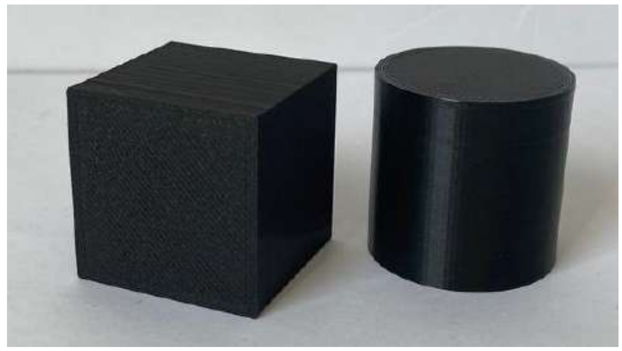


Figure 14. Objects from a dataset of figure images

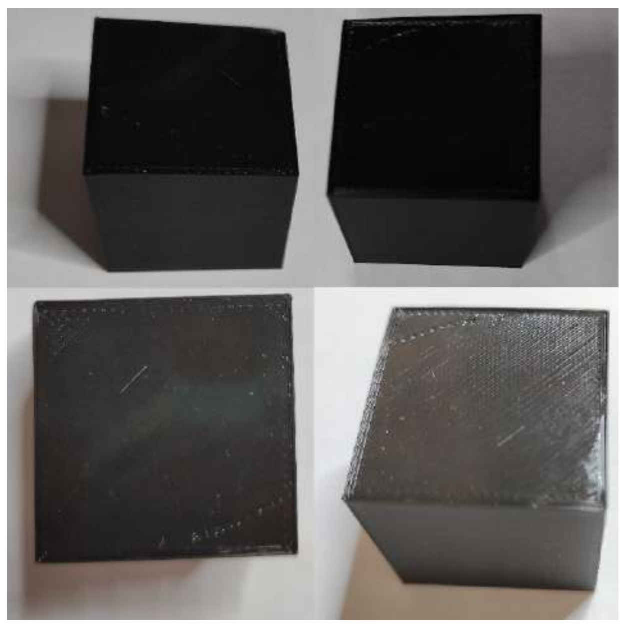


Figure 15. Images of one body taken from different angles

We used a pre-trained YOLO version 7 [11] as a neural network-detector to solve the subproblem of detecting objects of a given type. The whole dataset was split into training and test subdatasets in the ratio of 4:1. After 100 epochs of model training we got $mAP = \mathbf{97.4}\%$.

After training, the model acquired a good level of object recognition, and in some cases was even able to detect objects that were different in color from those in the training dataset (Fig. 16).

Due to the invariance of the model to the color of objects, a simple comparison using the Euclidean metric with a given reference color value was used to determine it.

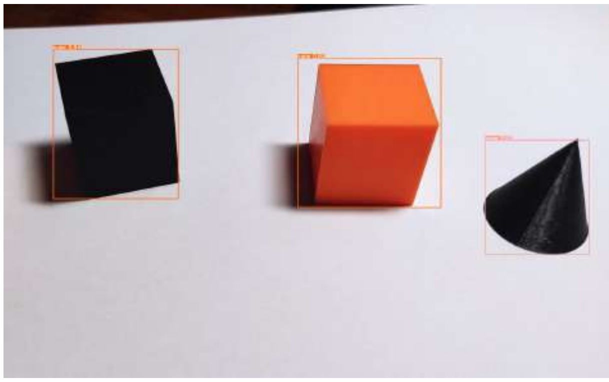


Figure 16. Objects detected after model training

V. Conclusion

This paper proposes a concept for the development of robotic ostis-systems. The proposed approach is based on the integration of robotic, symbolic and neural network components in a single system. The application of OSTIS technology for building intelligent robotic systems is substantiated, and the formalization of this subject domain is performed. Practical advice on the development of robotic ostis-systems is given.

Areas for future work include:

- increasing the versatility of the proposed concept by expanding the range of described types of robots and other auxiliary devices;
- finalization of the existing prototype system for implementation in production processes;
- implementation of diagnostic agents for the system components;
- finalization of the agents for calculation of the manipulator trajectory for arbitrary object position.

References

[1] M. Thosar, S. Zug, A. M. Skaria, and A. Jain, "A review of knowledge bases for service robots in household environments," in *International Workshop on Artificial Intelligence and Cognition*, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:198963468

[2] D. B. Lenat, "Cyc: A large-scale investment in knowledge infrastructure," *Communications of the ACM*, vol. 38, no. 11, pp. 33–38, 1995.

[3] I. Niles and A. Pease, "Towards a standard upper ontology," in *Proceedings of the International Conference on Formal Ontology in Information Systems - Volume 2001*, ser. FOIS '01. New York, NY, USA: Association for Computing Machinery, 2001, p. 2–9. [Online]. Available: https://doi.org/10.1145/505168.505170

[4] M. Tenorth, D. Jain, and M. Beetz, "Knowledge processing for cognitive robots," *KI*, vol. 24, pp. 233–240, 09 2010.

[5] M. Tenorth and M. Beetz, "Representations for robot knowledge in the knowrob framework," *Artificial Intelligence*, vol. 247, no. C, pp. 151–169, 2017.

[6] V. Golenkov, N. Guliakina, and D. Shunkevich, *Otkrytaja tehnologija ontologicheskogo proektirovanija, proizvodstva i jekspluatacii semanticheski sovmestimyh gibridnyh intellektual'nyh komp'juternyh sistem [Open technology of ontological design, production and operation of semantically compatible hybrid intelligent computer systems]*, V. Golenkov, Ed. Minsk: Bestprint [Bestprint], 2021, in Russ.

[7] "Raspberry Pi Datasheets," mode of access : https://datasheets.raspberrypi.com. – Date of access : 01.04.2024.

[8] "HC-SR04-Ultrasonic.pdf," mode of access : https://www.handsontec.com/dataspecs/HC-SR04-Ultrasonic.pdf. – Date of access : 01.04.2024.

[9] V. Golenkov, Ed., *Tehnologija kompleksnoj podderzhki zhiznennogo cikla semanticheski sovmestimyh intellektual'nyh komp'juternyh sistem novogo pokolenija [Technology of complex life cycle support of semantically compatible intelligent computer systems of new generation].* Bestprint, 2023, in Russ.

[10] D. Shunkevich, "Agent-oriented models, method and tools of compatible problem solvers development for intelligent systems," *Otkrytye semanticheskie tehnologii proektirovanija intellektual'nyh sistem [Open semantic technologies for intelligent systems]*, pp. 119–132, 2018.

[11] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," 2022.

ПРИНЦИПЫ ПОСТРОЕНИЯ ИНТЕЛЛЕКТУАЛЬНЫХ РОБОТОТЕХНИЧЕСКИХ СИСТЕМ

Крощенко А. А., Ковалёв М. В.

В статье предлагается концепция к построению робототехнических систем коллаборативного типа с использованием технологии OSTIS. Разработанная концепция базируется на интеграции робототехнического, символического и нейросетевого компонентов. Основные положения подхода проиллюстрированы проектом робототехнической системы для сортировки предметов с заданными характеристиками. Даются рекомендации о применении предлагаемой концепции для построения робототехнических систем коллаборативного типа в контексте разработки интеллектуальных компьютерных систем нового поколения, основывающихся на использовании технологии OSTIS.

Received 07.04.2024

Methodology of Machine Learning Model Development for Solving Applied Computer Vision Problems

Marina Lukashevich
Information Management Systems Department
Belarusian State University
Minsk, Belarus
LukashevichMM@bsu.by

***Abstract*—The methodology of machine learning model development for solving applied computer vision problems is presented. The article discusses the tasks of computer vision, the main components of building application systems and the challenges and limitations of the existing technological level.**

***Keywords*—Machine learning, computer vision, machine learning model, image, video**

I. Introduction

Machine learning (ML) is the creation and application of models internalized from data. In the case of traditional programming, rules are expressed in a programming language. They act on data, and computer programs provide answers. In the case of machine learning, the answers (typically called labels) are provided along with the data, and the machine infers the rules that determine the relationship between the answers and the data. Machine learning involves algorithms that learn from patterns of data and then apply them to decision making (Figure 1).

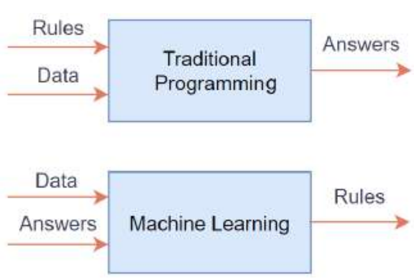


Figure 1. Machine learning vs. traditional programming

Machine learning can also be defined as the process of solving a practical problem by:

- collecting a dataset;
- algorithmically training a statistical model on that dataset.

Machine learning does not have a clear sequence of steps because it is necessary to work with different types of data (tabular data, images, video, signals, text, speech, etc.) and in different domains (data analysis, computer vision, natural language processing, robotics, etc.). Each case has its own specifics, algorithmic techniques, and tools. The main goal of this paper is to summarize theoretical background and practical experience, formalize a methodology for machine learning model building for solving computer vision problems, and formulate some practical recommendations.

II. Computer vision analysis

A. Computer vision

Computer vision (CV) is defined as the automatic extraction of information from images or videos. Some tasks require computer vision to simulate human vision. In other cases, it is necessary to perform statistical data processing, geometric transformations, etc. In practice, computer vision is a fusion of artificial intelligence, pattern recognition, digital signal and image processing, math, and physics (Figure 2). It depends on the specific problem being solved.

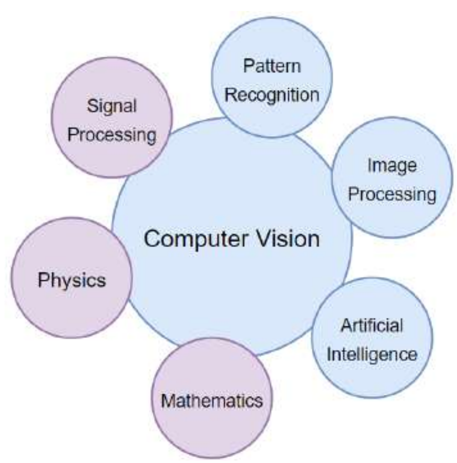


Figure 2. Interdisciplinarity of computer vision

B. Object and scene level tasks

The focus of object-level tasks is on objects in a visual scenario, and they require the analysis and understanding of various entities or instances that are associated with them. These tasks involve recognizing objects, detecting them, tracking them, detecting changes, detecting anomalies, and segmenting them.

The process of object recognition involves identifying and categorizing objects into predetermined classes or categories. In object detection, the procedure of recognizing and precisely locating an object within an image or video is performed by creating a bounding box around it.

The process of tracking the movement of objects across multiple frames of aerial video or image sequences is referred to as object tracking. The process of identifying alterations in imagery over different time instances is called change detection.

The process of systematic identification of abnormal patterns or objects within visual data and contrasting them with established norms is referred to as anomaly detection.

In semantic segmentation, semantic labels or classes are assigned to each pixel in an image. In the case of instance segmentation, semantic labels, or class labels, are assigned to each pixel in an image with distinction between individual instances of the same class. The complexity and list of computer vision tasks for object level are presented in Figure 3.

Scene-level tasks focus on a specific scene or an entire image or video scenario in visual data and involve in-depth analysis of context, composition, or environmental features. Tasks such as image registration, 3D reconstruction and terrain modeling, and localization and mapping are some of the scene-level tasks. The complexity and list of computer vision tasks for scene level are presented in Figure 4.

C. Revolution in computer vision

In 2006, Nvidia released CUDA, a programming language that allowed GPUs to be used as general-purpose supercomputers. In 2009, artificial intelligence researchers at Stanford introduced Imagenet, a collection of labeled images used to train computer vision algorithms. A revolution in computer vision occurred when neural networks aimed at working with images began to be used. These are called convolutional neural networks. In 2012, convolutional neural networks (AlexNet) significantly reduced the error in classification and approached the result, which shows in image recognition by a human about 5% of errors. And in 2015, neural networks overtook humans in recognition accuracy and showed a result of 3.6%. CNN, the ImageNet dataset, and graphics processors were the magic combination that launched a powerful advance in computer vision. In the last few years, neural networks based on transformer architecture with Attention mechanism have shown excellent efficiency in computer vision.

The most effective solutions in the field of computer vision are based on neural networks in general and on deep neural networks in particular. A large number of effective architectures of deep neural networks have been proposed by researchers. These architectures are implemented in modern deep learning frameworks and are widely used in practice. The classical theory of pattern recognition and digital image processing began to take a back seat. The classical scheme, including image preprocessing, feature computation, and decision-making, has become less effective. The use of neural networks involves feature computation and decision making by the neural network itself [1]–[8].

III. Components for solving computer vision problems

Highlight the main components necessary for realizing practical solutions for machine learning tasks.

- Datasets.
- Frameworks and libraries.
- Model architecture.
- Hardware resources.

Machine learning models are built on the basis of data. Two types of datasets can be identified:

- large datasets used for model pre-training and implementation of transfer learning technology (for example, ImageNet, COCO, etc.);
- custom datasets are collected and labeled for a specific task.

It is a good solution for scientific and educational tasks to use public datasets from well-known platforms (Kaggle, Roboflow, etc.). In addition, it should be mentioned that synthetic data can be used to train models in some cases. Synthetic data can be obtained in the following ways:

- using generative neural networks;
- building 3D models of objects, creating synthetic images and videos with different backgrounds, extraneous objects, etc. on their basis (using Blender, NVidia Omniverse, etc.).

There are two leading frameworks for deep neural network development: PyTorch and TensorFlow. Both are powerful frameworks with unique strengths. PyTorch is favored for research and dynamic projects, while TensorFlow excels in large-scale and production environments. Industry experts may recommend TensorFlow, while hardcore ML engineers may prefer PyTorch. However, there has been a general trend of increasing usage and preference for Pytorch.

In addition to frameworks, a number of libraries and IDEs are used in building computer vision solutions, as well as in experiments and development processes. Some of them are presented in Table I.

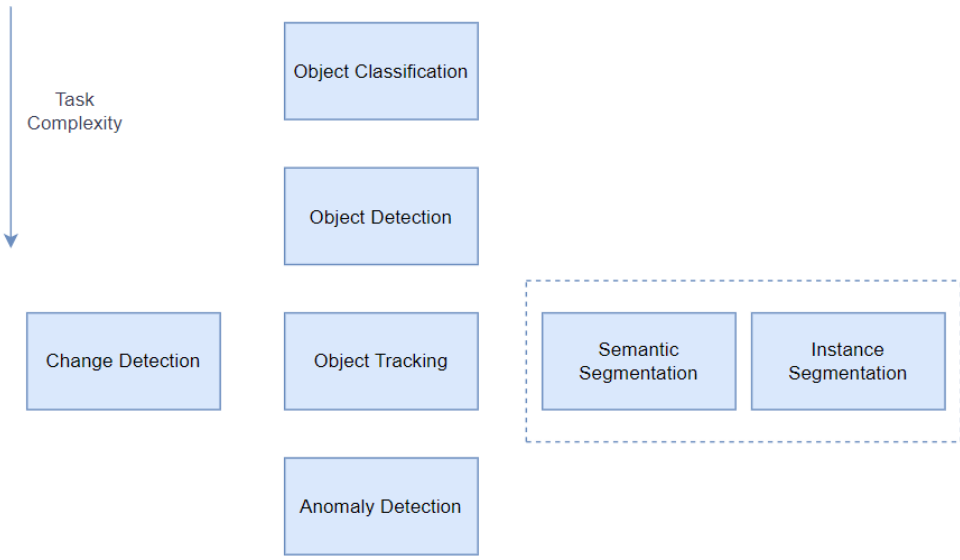


Figure 3. Complexity of computer vision tasks. Object level tasks

Table I
Some ML/CV frameworks and libraries

Data	Training/Evaluation	Libraries for ML/CV
Labeling CVTA, LabelMe, COCO-Annotator, COCO-ui	**Frameworks and Distributed Training** TensorFlow, Keras, TensorFlow Lite, PyTorch, PyTorch Lightning	**Machine Learning** Numpy, Scipy, Scikit-learn, Pandas, Matplotlib, Seaborn, h5py
Exploration Pandas, Seaborn	**Software Engineering** Python, PyCharm, VS Code, git	**Computer Vision, Digital Image Processing** Scikit-image, OpenCV, Pillow, Matplotlib
	Experiment Management TensorBoard, Weights and Biases	
	Hyperparameter Tuning Weights and Biases	

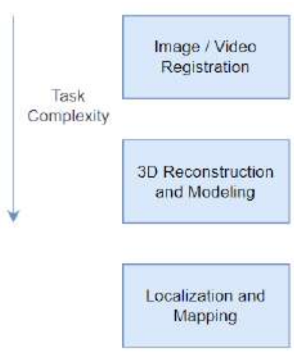


Figure 4. Complexity of computer vision tasks. Scene level tasks

When selecting a neural network architecture to solve a given problem, there are several options. As a rule, while working on projects, researchers and developers apply to each of them in the sequence in which they will be listed:

- state-of-the-art (SOTA) architecture for specific computer vision problems (classification, detection, segmentation, etc.);
- modified state-of-the-art architecture for specific computer vision problems (classification, detection, segmentation, etc.);
- costume state-of-the-art architecture for specific computer vision problems (classification, detection, segmentation, etc.).

Some state-of-the-art architectures for computer vision tasks are presented in Table II. This is by no means a complete list, but it will provide insight into the variety of neural network architectures for solving computer vision problems.

The architecture selection process should also be guided by the model size and its inference time. There may be situations when it is necessary to develop a lightweight model, for example, for cutting-edge devices, and to provide a high speed of inference.

Another necessary component for the development of computer vision solutions is computational resources. It is possible to use both GPUs on workstations and cloud computing resources (for example, Amazon Web Services, Google Cloud Platform, Microsoft Azure, etc.).

Table II
Some State-of-the-art NN architectures for computer vision

Task	SOTA Architectures
Classification	Xception, VGG16, VGG19, ResNet50, ResNet101, ResNet152, InceptionV3, MobileNet, MobileNetV2, DenseNet121, DenseNet169, DenseNet201, NASNetMobile, NASNetLarge, EfficientNetB0,EfficientNetB1, EfficientNetB2, EfficientNetB3, EfficientNetB4, EfficientNetB5, EfficientNetB6, EfficientNetB7, ConvNeXTTiny, ConvNeXTSmall, ConvNeXTBase, ConvNeXTLarge
Detection	SSD, YOLO5, YOLO7, YOLO8, YOLO9, YOLO NAS
Segmentation	Unet, FPN, Linknet, PSPNet, SegFormer

NVIDIA, the market leader, offers deep-learning GPUs.

IV. Data Importance

The main blockers in machine learning projects are data unavailability, an insignificant number of data, and low variability. Often, the choice of neural network architecture is not as critical as an underpowered dataset. Also, data collection and labeling for model building can take a significant amount of time. Especially recently, much attention has also been paid to personal data, the impossibility of using it, and data ethics in general.

Formulate some rules about data for ML model development:

- at a bare minimum, collect around 1000 examples.
- for most "average" problems, collect 10,000 - 100,000 examples.
- for “hard” problems like machine translation, high dimensional data generation, or anything requiring deep learning, collect 100,000 — 1,000,000 examples.

Generally, the more dimensions your data has, the more data you need. It is necessary to have roughly 10 times the amount of data in your examples. The more complex the problem, the more data you need.

A. Multiplicity of computer vision tasks and interdisciplinarity

We would like to emphasize that there is often a need for special, interdisciplinary knowledge when solving computer vision problems. This is evident when working with a class of medical images. In such cases, it is necessary to involve medical experts both for data labeling and for result interpretation. A good rule of thumb is to engage several experts at the same time and average their labels and estimates.

On the other hand, different computer vision problems are possible for the same class of images. It is very important to formulate the problem in terms of machine learning and computer vision (image classification, object detection, image segmentation, etc.) at the very beginning of the project. This will immediately give an understanding of how to label images (polygon, class label for a whole image, using bounding boxes, etc.) and with what architectures and algorithms to perform experiments (Table II).

As an example, here are images and possible computer vision tasks. This is a class of medical images. Figure 5 shows examples of retina images and their labels for classifying the stages of diabetic retinopathy. Figure 6 shows examples of optical disk detection. And Figure 7 shows the results of vessel segmentation in retinal images [9]–[11].

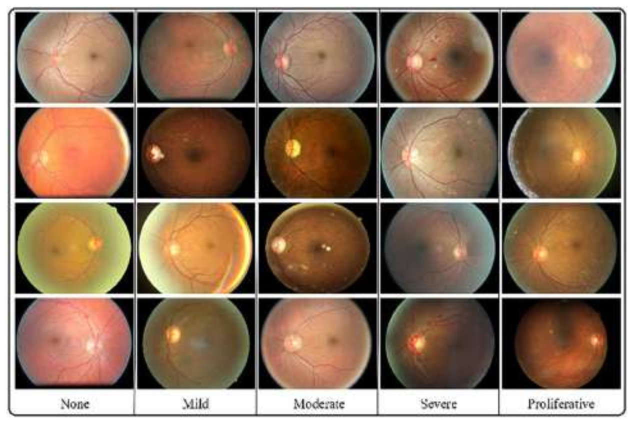


Figure 5. Diabetic retinopathy classification

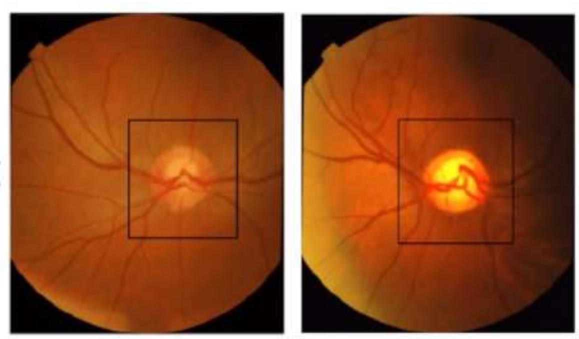


Figure 6. Optical disk detection

V. Methodology of machine learning model development for Computer Vision

Summarize the above information and formulate a methodology for machine learning model development for solving computer vision tasks. There are three logical levels to this methodology:

- coding.

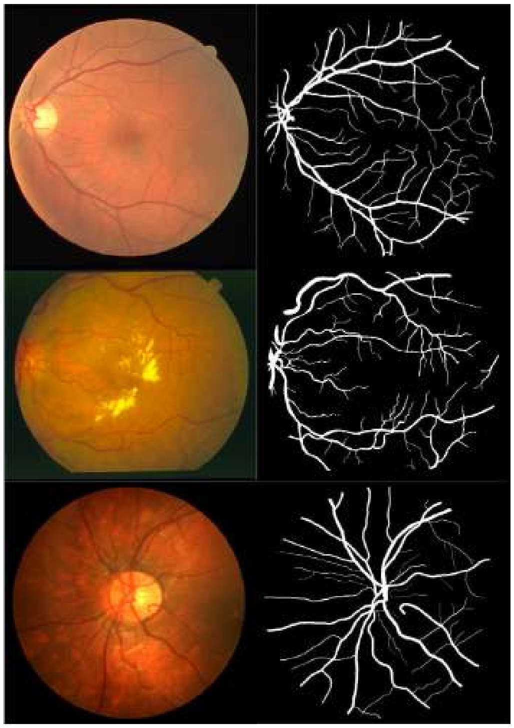


Figure 7. Retinal vessels segmentation

- model development.
- working with data.

The methodology includes next stages:

- *Model Building.* The computer vision task is formulated (classification, detection, segmentation, etc.), the stack of technologies used is determined, data is collected and labeled, promising model architectures are determined, and models are built. A common practice is to split the data set into a training set, a validation set, and a test set. The training dataset is used for model building. A validation dataset is necessary to improve the training process.
- *Model Evaluation and Experimentation.* The accuracy of the models is assessed on the test dataset based on metrics for specific tasks. The best model is chosen.
- *Productionize Model.* The model is saved using the selected format (ONNX, h5, etc.).
- *Testing.* Code and model are tested using training dataset.
- *Deployment.* The deployment of the model in the product environment is determined and implemented. There are several possible options:
  - batch deployment;
  - real time;
  - streaming deployment;
  - edge deployment.
- *Monitoring and Observability.* Continuously monitoring and testing the model's effectiveness post-deployment pose ongoing challenges. Regular monitoring is necessary to ensure accurate results, identify potential issues, and drive performance enhancements.

VI. Challenges of building intelligent systems for computer vision tasks

Despite the great theoretical and technical progress in the field of computer vision, there are a number of limitations that need to be overcome in the future. Let's name some of them.

- The diversity of visual representation, such as illumination, perspective, or occlusion in objects, is a major challenge. These variations must be overcome to eliminate any visual inconsistencies.
- With each image consisting of millions of pixels, dimensional complexity becomes another barrier to overcome. This could be done by using different techniques and methodologies.
- Real-time processing can be challenging. This comes into play when making decisions for autonomous navigation or interactive augmented realities, which require optimal performance of computational frameworks and algorithms for fast and accurate analysis.
- Ethical considerations are paramount in artificial intelligence, and computer vision is no different. This could be bias in deep learning models or any discriminatory results. This emphasizes the need for a proper approach to dataset curation or algorithm development.

Outline promising directions in the field of computer vision in the next few years: zero-shot learning, few-shot learning, and one-shot learning. It makes sense to develop scientific research in this direction. Zero-shot learning, few-shot learning, and one-shot learning are all techniques that allow a machine learning model to make predictions for new classes with limited labeled data. The choice of technique depends on the specific problem and the amount of labeled data available for new categories or labels (classes) [12].

If we move away from technical blockers and think about high-level analysis of computer vision challenges, then it is obvious that the component design of hybrid and intelligent systems will be promising in the future. It is necessary to solve the problem of the compatibility of scientific research results in the field of artificial intelligence. This problem is currently the key one preventing the active development of artificial intelligence.

a significant reduction in the effectiveness of using the component method. designing computer systems based on reusable libraries components.

Insufficiently high degree of learning about modern computer systems during their operation, which results

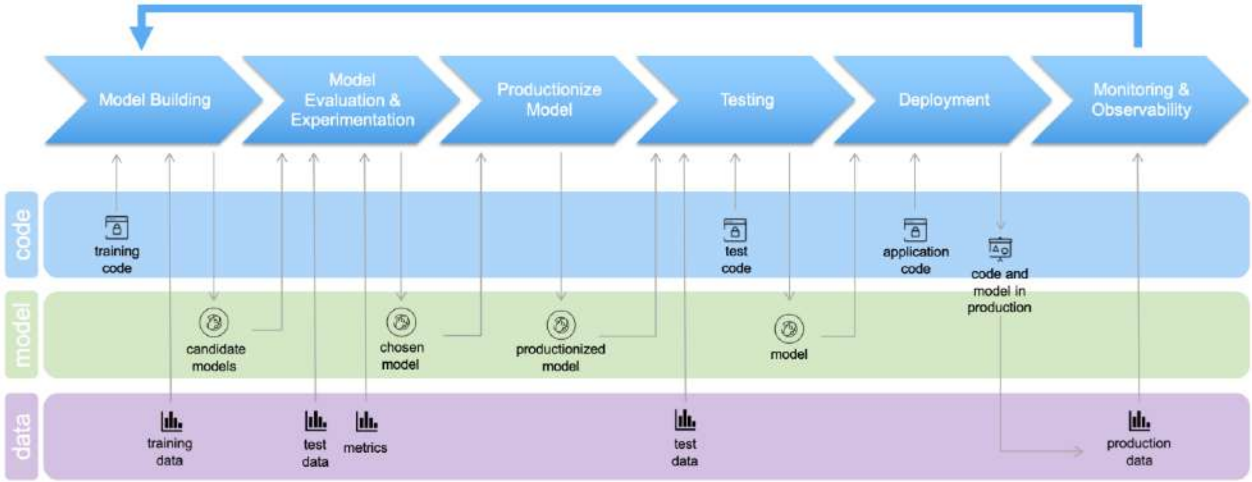


Figure 8. Methodology of machine learning model development for computer vision

in the high complexity of their maintenance and improvement, as well as their insufficiently long life cycle.

The problems of the unification of the principles for constructing various components of computer systems are solved in the OSTIS project. The OSTIS The project aims to create an open semantic technology for designing knowledge-driven systems in general and computer vision systems in particular [13].

VII. Conclusion

The methodology for developing a machine learning model for solving applied computer vision problems is presented. The article discusses the tasks of computer vision, the main components of building application systems, and the and the challenges and limitations of the existing technological level. demonstrated the need to develop areas related to the compatibility of scientific research results in the field of artificial intelligence.

References

[1] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In ICLR, 2015.

[2] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. Communications of the ACM, 60(6):84–90, 2017.

[3] Sebastian Raschka; Yuxi (Hayden) Liu; Vahid Mirjalili; Dmytro Dzhulgakov, Machine Learning with PyTorch and Scikit-Learn: Develop machine learning and deep learning models with Python , Packt Publishing, 2022.

[4] Fermüller, Cornelia and Michael Maynord. “Advanced Methods and Deep Learning in Computer Vision.” (2022).

[5] Zhengxia Zou, Keyan Chen, Zhenwei Shi, Yuhong Guo, and Jieping Ye. Object detection in 20 years: A survey. Proceedings of the IEEE, 2023.

[6] Tianfei Zhou, Fatih Porikli, David J Crandall, Luc Van Gool, and Wenguan Wang. A survey on deep learning technique for video segmentation. TPAMI, 2023.

[7] Jingjing Xu, Wangchunshu Zhou, Zhiyi Fu, Hao Zhou, and Lei Li. A survey on green deep learning. arXiv preprint arXiv:2111.05193, 2021.

[8] Jianyuan Guo, Kai Han, Han Wu, Yehui Tang, Xinghao Chen, Yunhe Wang, and Chang Xu. Cmt: Convolutional neural networks meet vision transformers. In CVPR, pages 12175–12185, 2022.

[9] R. K. Kumar and K. Arunabhaskar, "A Hybrid Machine Learning Strategy Assisted Diabetic Retinopathy Detection based on Retinal Images," 2021 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES), 2021, pp. 1-6, doi: 10.1109/ICSES52305.2021.9633875.

[10] Dagliati, A, Marini, S, Sacchi, L, Cogni, G, Teliti, M, Tibollo, V, De Cata, P, Chiovato, L & Bellazzi, R 2018, ’Machine Learning Methods to Predict Diabetes Complications’, Journal of Diabetes Science and Technology, vol. 12, no. 2, pp. 295-302. https://doi.org/10.1177/1932296817706375

[11] Lukashevich M.M. A neural network classifier for detecting diabetic retinopathy from retinal images. «System analysis and applied information science». 2023;(1):25-34. (In Russ.) ttps://doi.org/10.21122/2309-4923-2023-1-25-34

[12] Labrak, Y., Rouvier, M., Dufour, R. (2023). A zero-shot and few-shot study of instruction-finetuned large language models applied to clinical and biomedical tasks. arXiv preprint arXiv:2307.12114.

[13] V. Golenkov, N. Guliakina, V. Golovko, V. Krasnoproshin, “Methodological problems of the current state of works in the field of artificial intelligence,” Otkrytye semanticheskie tekhnologii proektirovaniya intellektual’nykh system [Open semantic technologies for intelligent systems], pp. 17–24, 2021.

МЕТОДОЛОГИЯ РАЗРАБОТКИ МОДЕЛЕЙ МАШИННОГО ОБУЧЕНИЯ ДЛЯ РЕШЕНИЯ ПРИКЛАДНЫХ ЗАДАЧ КОМПЬЮТЕРНОГО ЗРЕНИЯ

Лукашевич М. М.

Представлена методология разработки моделей машинного обучения для решения прикладных задач компьютерного зрения. В статье рассматриваются задачи компьютерного зрения, основные компоненты построения прикладных систем, проблемы и ограничения существующего технологического уровня.

Received 13.03.2024

Principles and Experience of Intelligent Decision Support and Recommender Systems Engineering

Boris Zhalezka
Department of Marketing
Belarusian National Technical University
Minsk, Belarus
boriszh@yandex.ru

Volha Siniauskaya
Department of Industrial Marketing and Communications
Belarusian State Economic University
Minsk, Belarus
volha.siniauskaya@outlook.com

***Аннотация*—In the article main principles of engineering of intelligent decision support and recommender systems are considered. Definition and concept of a generalized object are formulated. Technologies of recommender systems engineering and construction are analyzed. Classification of decision support systems is suggested. Experience of decision support and recommender systems engineering is presented.**

***Keywords*—decision support system, recommender system, generalized object, classification, neural networks, algorithmic trading, multi-agent technology**

I. Introduction

Despite existing developments in the area of decision support systems (further — DSS) engineering, these technologies presuppose the adaptation of only individual components of the DSS and do not provide the adaptation of the subject area model. This leads to the use of irrelevant and inaccurate data in the DSS, which negatively affects the efficiency of decision-making in a quickly changing environment. These problems can be solved via adapting subject area models to the conditions of decision-making tasks and well-timed updating of the models with the data, knowledge and precedent (subject) collections necessary for this model. The concept of a generalized object became the base of different DSS with combined intellect engineering and construction. The target of this research is to generalize theoretical and practical experience in the sphere of intelligent decision support and recommender systems construction.

The target of this research is to generalize theoretical and practical experience in the sphere of intelligent decision support and recommender systems construction.

II. Brief literature review

First theoretical research on decision support systems was made in USA at the Carnegie Institute of Technology in the late 1950s — early 1960s. The first main works on DSS were published in 1978-1980 by P.Keen, M. Scott Morton [1], [2], R. Sprague [3].

Investigations in the sphere of intellectualizing of DSS information technologies are wavy carried out in the world. Moreover, the increase in research activity in time coincides with the periods of development on computing and financial resources (the emergence of personal computers, evolution of the processors, memory capacity, the emergence of a user-friendly interface, the emergence of mobile Internet technologies, etc., with the simultaneous reducing of its prices).

Among the modern works on DSS there are books “Intelligent Decision Support System for IoT-Enabling Technologies: Opportunities, Challenges and Applications” (2024) by ed. S. Sahana [4]; “Intelligent Decision Support Systems for Smart City Applications” (2022) by L. Gaur, V. Agarwal and P. Chatterjee [5]; “Intelligent Decision Support Systems” (2022) by M. Sànchez-Marrè [6]; “Understanding Semantics-Based Decision Support” (2021) by S. Jain [7]; “Intelligent Decision Support Systems: A Complete Guide — 2020 Edition” by G. Blokdyk [8].

The technology of recommender systems gained popularity only in the middle of 1990s. The concept of a recommender system was first used in 1992 in the scientific publication of Xerox, in the same year in the article “Using collaborative filtering to weave an information tapestry” the term “collaborative filtering” was introduced. Subsequently, fundamental works systematizing knowledge on recommender systems were devoted to this area. One of them is “Recommender Systems: The Textbook” [9]: in this book, description, comparison, assessment of the accuracy of basic algorithms for developing recommendations to the user were made, in addition, the field of practical use of such systems is affected. The work “Recommender Systems: The Handbook” [10] deserves special attention. In it the existing variety of methods and concepts of recommender systems was systematized. This source shows how recommender systems can help users in decision-making, planning and procurement processes, illustrates the experience of using these systems in big corporations such as Amazon, Google, Microsoft.

Nowadays the transition to the next-generation computer systems takes place. Intelligent DSS and recommender systems belong to this class of information systems. Such systems should “independently evolve and interact effectively with each other in the collective solution of complex problems” [11]. One of the relevant problems of next-generation computer systems design and development is generalization of formal theory and methodology of their functioning.

III. Definition and the concept of a generalized object

Usually an object O can be represented in the view $O =< Data, Met, Mes >$, where: $Data$ — a set of internal information of the object (data); Met — a set of its own procedures for manipulating the data (methods); Mes — an external interface for interacting with other objects in the subject area (such as a permissible set of event messages outside and inside of the subject area).

However, the need to take into account the development of DSS requires a more general and flexible mechanism for describing and modeling them. Such a mechanism can be built on the base of further generalization of the object-oriented approach and, in particular, of the term “object” in the conceptual model of the subject area.

The term “generalized object” is suggested: $GO =< Data, Met, Model, Knowl, Mes, Link >$, in which models, knowledge and links with the other domain objects are encapsulated in addition to data, methods and messages.

Such model of a generalized subject area can be considered as a type of multi-object neural network, if add to the usual method of object interaction by messages the possibility of activating objects (transmission of excitation) through links with weights (priorities) indicating the value of the response threshold level for each object. The value of the response threshold level can change in accordance with the system development stage, solved problems, accumulated statistical experience of problems solving, etc. Interaction between generalized objects can be carried out through messages or changes in the links structure of the generalized object system. This system is separated from their functional part. It can be and represented by a dynamic links list, by changes in the links weights (filtering) and actuation threshold values.

Each generalized object can have several states and go from one state to another depending on the incoming messages that are the result of the activities of other generalized objects. At the same time, the generalized object changes its state when its excitation value exceeds some non-zero threshold of actuation. In general, such subject area model can be considered as a hierarchy of abstractions represented by classes of generalized system, problem and user objects. The status of the subject area actually depends on the status of each generalized object and the message queues at the input and output of these generalized objects. The latter can be considered as a database of facts about events on the base of which it is possible to determine an output machine for interpreting of existing and generating new facts in the process of simulating of modelling system functioning.

Requirements to the multi-agent DSS can be formulated by means of object-oriented classification (enumerations of the object classes involved in solving problems, their properties, relationships and behavior). The main principle of classification is the specification of object classes based on the set of internal properties inherent in class objects. After that, the requirements are sequentially detailed until the multi-agent DSS project is fully described in terms of the basic objects of the used tools.

IV. Recommender systems

In today’s world, we can often face the problem of recommending goods and services to users of any site, information system. The modern economic formation involves intense competition in various market niches, which leads to a struggle for each potential buyer among companies. In order to improve their experience of interacting with the company’s services, it is beneficial to create personalized collections that will have client response. Previously, for such recommendations list, a set of current actions and the most popular goods was sufficient, but the current situation does not allow such low-cost methods to act. A relatively new technology of recommender systems can be used for buyers’ attraction and sales increasing.

The essence of recommender systems approach presupposes the dynamic formation of recommendations personally for each specific client, which, unlike static information, significantly increases probability of coincidence with the real needs of people. Recommender systems take into account all kinds of parameters (purchase history, time and date of registration, region, purchased products, etc.), which allows prediction of the user’s wishes as accurately as possible.

Recommender systems have already found their place in many areas: in addition to e-commerce, this kind of technology has been introduced for finding books, films, music, and social media contacts.

The relevance of recommender systems is growing every year. Recommender systems are the programs aimed for the prediction of the user’s interest in certain objects and giving them the recommendation for purchase or using the items, which the user probably likes. Such recommendations are personalized and are formed for each user depending on their preferences. Although the technology itself appeared quite recently, its use is considered mandatory for all promising companies.

One of the first to become interested in introducing recommender systems was the largest American e-commerce platform Amazon. Already in the late 1990s, the best minds of the company developed their own algorithms for the so-called collaborative (joint) filtering, which offered recommendations to each client based on the history of their purchases, views, and liked goods. The success of the marketplace forced competitive companies to pay attention to the recommender systems technology. In the rapidly growing segment of social networks business-oriented network LinkedIn also uses recommender system. Its system offers a member of the community the closest to the interests, specialization and experience of the community, company, specialists. To build recommender systems, there are four main types of data filtering [12]:

- collaborative filtering;
- content filtering;
- knowledge-based filtering;
- hybrid filtering.

Often the recommender system technology is based on the principles of collaborative filtering, which analyses the actions of the most similar users with a similar profile. However, other types of filtering are also used in practice.

A. *Collaborative filtering*

The main principle of the functioning of filtering programs is the assumption that users with the same interests will subsequently have similar preferences. For the effective functioning of the model, not only the previous behavior of the client, its query history, but also the corresponding parameters of the additional cluster of similar users are taken into account. The target of collaborative filtering is to identify a certain number of customers operating with the closest patterns of behavior, and to recommend in further the goods or services liked by this group.

The next method, which in turn is already based on comparing the similarity of objects, is called item-based. Its principle can be formulated as follows: if users who rated two products liked both, then the following users who tried only one product can be offered another.

B. *Content filtering*

Content filtering is based on the assumption of the constancy of user interests. In other words, based on the client's past activity, it can be argued that in the future he will be interested in similar objects. Content filtering uses the following input data: both a set of users and a set of categories that correspond to users' queries and to the objects, which users like.

The purpose of this type of recommender systems is to build such a variety of items that are closest to the favorite categories of the current user. The main task of this methodology is to search for objects potentially close to the interests of users among a set of objects not yet viewed by the client. This search is based on finding the similarity of objects with the user's interests known to the system. The absence of the need to revealing large user groups to ensure the functionality of the method is one of the main advantages of content filtering. This method also avoids the problem of "cold start", because each object has attributes that will be analyzed in future. This type of filtering often make combination with collaborative filtering.

C. *Knowledge-based filtering*

The most resource-intensive approach is to develop knowledge-based recommender systems. The main source of information is not the user assessment of objects or their metadata, but the rules and conditions developed by experts and expert systems for forming recommendations. Some researchers consider content filtering as a special case of the knowledge-based filtering, however, due to the wide prevalence, most prefer to classify it as a separate type. For the functioning of this kind of recommender systems, it is necessary to distinguish many expert rules, similarity metrics and user interest objects. For practical application of a given rules set, it is necessary to define user's interests and preferences in terms of the subject area.

The knowledge-based approach requires deep understanding of the technical features of the product, creation of user scenarios, inclusive restrictions and rules. Undoubtedly, the use of knowledge-based systems improves the quality of the recommendations being formed, since user requests find the most accurate response from recommendation algorithms among all methodologies. In addition, this method will be indispensable in those areas of commerce where the number of regular customers is relatively small. Of course, the development of such systems is extremely time-consuming in terms of time and resources. To improve the accuracy of functioning, it is necessary to involve relevant specialists in the field of data collection and processing, building the necessary models and user behavior. Also, such systems require additional interaction from the client with the system, which can lead to the outflow of some part of the target audience, moreover, the collected data cannot always be correctly interpreted by software.

D. *Hybrid filtering*

The last type of recommender systems, hybrid, as the name suggests, is a synthesis of two or three above-mentioned methodologies. The use of hybrid recommender systems increases the efficiency, performance and accuracy of algorithms, and compensate their lacks. For example, the most used combinations are:

- combining collaborative and content filtering approaches (with different weights);
- using some content-based filtering properties in collaborative filtering algorithms;
- partial using of knowledge-based filtering rules in recommender systems based on content filtering;
- building a separate model corresponding to business needs and subject area terms, combining rules of all three types.

There is no unified algorithm for the functioning of hybrid systems, which allows researchers to apply a wide range of modern methodologies to create unique models. It was the hybrid type that became the basis of recommender systems in large companies to ensure better personalized interaction of users with their services.

Let's take a closer look at two of the most popular methods of recommender systems – collaborative filtering and content filtering. Algorithms of these types of filtering can be classified into three main categories:

- anamnestic methods, or methods based on the analysis of available estimates (memory-based filtering), are a family of algorithms that are based on statistical methods, the purpose of which is to search for the nearest group of users to the analyzed user; this approach is similar to the closest neighbors method, and recommendations are formed as a result of calculating a similarity measure based on a matrix of estimates of the users in the database; the main representative of memory-based algorithms is the used-based and item-based weighting of estimations;
- model-based filtering, in which a descriptive model of user and object assessments is preliminary formed, and priority relationships between them are distinguished; the main complexity of this method is its preliminary stage, where resource-intensive training of the model takes place; different approaches can be used to building such a model: cluster analysis methods, Markov decision process (MDP), singular value decomposition (SVD), latent semantic analysis (LSA), principle component analysis (PCA), etc.;
- hybrid methods, which suppose synthesis of several approaches to achieve a better result; for example, collaborative filtering systems can take advantage of a relatively easy-to-interpret anamnestic method with the efficiency and performance of model-based methods, the purpose of which is to increase the speed of recommender system work.

Problems of development of recommender systems may include following situations:

- sparseness of data, which means that due to a huge number of data the matrix "object-user" in system's database becomes difficult to processing that complicates overall algorithm's work;
- scalability, which means that traditional data processing algorithms may not cope with the growing flow of new customers and the goods they evaluate; for example, it can be extremely difficult to perform operations on matrices illustrating information about tens of millions of users and hundreds of thousands of objects, especially because the requirement for modern recommender system is an instant response to customer requests from all over the world;
- "cold start", which arises in the case of new customers and goods emergence, because the absence of the information about the previous user interaction; this problem can be partially solved by the use of content/knowledge analysis or so-called "average"user;
- the lack of unified names of analyzed objects (especially in users' queries) may have negative influence on the efficiency of joint filtering methods; recommender systems do not have the ability to define a hidden speech association, which can lead to the including of the same objects to different classes; for example, an algorithm will not be able to find the coincidence of the "toys for children" and "children's toys" queries;
- fraud, for example, companies interested in the profitable sale of their own products can artificially underestimate the goods of their competitors and wind up a positive rating of their products, that will lead to the recommendation of the products of firms using such frauds;
- market diversity, which allows users to explore the vast expanses of marketplaces in better products search, and such consumer boom does not always correlate with the work of some methods of collaborative filtering, for example, purely based on the rating and success of sales of goods that do not take into account the possibility of promoting little-known, new goods, which can adversely affect the diversity of the market and will lead to the survival of only the largest market players to whom the main user attention has already been focused;
- presence of the clients at the market, whose opinion is strikingly different from the majority; for such users, algorithms may not find unique like-minded people suitable for users, it can reduce the quality of their personalized recommendations.

V. Experience of DSS and recommender systems engineering

There are several main results in DSS engineering have been received by the authors:

- methods of efficiency assessment of the rule and model bases in DSS have been elaborated, which include the following coefficients: rule base certainty, rule base coverage, rating class efficiency and rating efficiency; formulae for calculation of these coefficients are deducted with the using of the rough sets theory [13];
- theoretical and practical approaches of DSS engineering for stock markets have been developed, which include the technology of the liquidity evaluation [14], the technology of algorithmic trading by means of 5-component oscillator [15], the technology of securities prices prediction with the use of the neural network [16];
- the concept of algorithmic marketing and machine learning in relation to the marketing activity, which

is based on decision trees and ABC-analysis and allows reducing time necessary for market big data processing [17];

- the methodology of multi-agent DSS design and developing, which includes approach to the modelling of representation of knowledge about the subject area on the base of the concept of generalized objects [18];
- adaptation of the technology of DSS engineering to recommender systems design, which will be considered below.

Example of the recommender system was realized for the purpose of the choice of musical tracks for different sport training. The initial data for analysis are the most popular audio sets for a specific training session (playlists of reputable sports publications, well-known fitness instructors and trainers, popular music editions).

For the study 4 types of sports training were used:

- yoga, aimed at achieving internal harmony and tranquility;
- cardio training, including a set of intensive exercises that increases the heart rate;
- running, aimed at increasing endurance;
- power exercises that contribute to an increase in muscle strength.

Client-server representational state transfer architecture (REST) is used for the integration of the recommender system with the mobile applications [19]. General scheme of interaction between Android client and database in REST architecture is represented in Fig. 1. The main requirements to this client-server architecture are the following: the server cannot store client information between requests, that’s why interface should be unified.

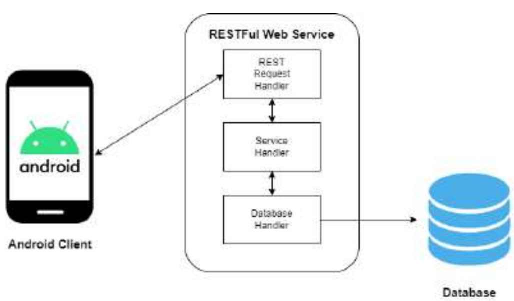


Рис. 1. General scheme of interaction between Android client and database in REST Architecture.

Due to creation of music track pattern for each sport training it is necessary to create a big data set on the base of expert data. A large number of listeners (at least 15,000 people) will serve as an indicator of quality. As a result, a set of 20 playlists was formed for each class, totaling more than 1000 unique tracks, on the base of which the desired average was calculated. Then, unique playlist identifiers are read from the generated files in order to eventually process each composition from the audio selection and sequentially extract all the properties of the tracks from it. To form the desired patterns, the function of calculating the average value of the set of components was used.

As a result, 10 average indices represented in Fig. 2 have been calculated: acousticness, valence, danceability, energy, instrumentalness, key, liveness, loudness, speechiness, tempo.

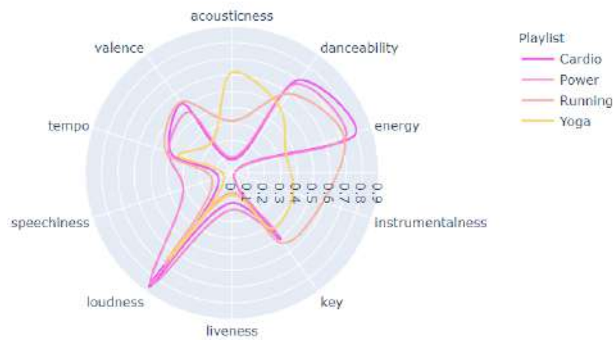


Рис. 2. Example of classification of the musical tracks by recommender system.

After forming a pattern sample for each of the classes, it is necessary to select a set of tracks that are most similar to the pattern (200 songs are selected in the application). The search for songs is carried out from a large external data set, the storage of which is organized in a .csv file.

For example, the program uses a file with 32880 unique compositions, where, in addition to a unique identifier, genre, the name and listing of the artists participating in the recording of the song, the extracted audio properties are presented. The playlist downloaded as a result of the operation of the HTTP library is displayed as interactive buttons, clicking on which leads to the playback of the corresponding song (Fig. 3).

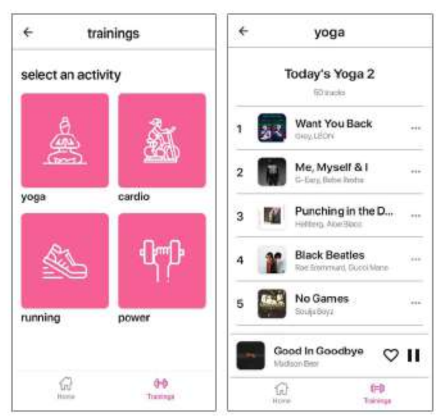


Рис. 3. Interface of the mobile recommender system for choosing musical tracks for different sport trainings.

The recommender system also has own media player, where it is possible to pause/continue playing a song, turn on the next/previous composition; and mix and loop modes are available.

VI. Classification of decision support systems

DSS can be classified into 9 different classes in accordance with the used information data, models and knowledge. In table 1 structure formulae for each class of DSS are represented [20].

Таблица I
DSS classification

DSS structure formula	Information for decision making	Used data, models and knowledge
=objects + information + information collection tools	all which exists	all factual data about subject area
=alternatives + data + links	all which can be useful	actual data
=alternatives + criteria + criteria's values	all which is necessary (from that which exist)	relevant (selected) data
=data + models	all which can be formalized (modelled)	formalized data (actual models)
=models + rules criteria's estimations	all which must be modelled	relevant models
=rules of alternatives' estimation + alternatives' rating	any variants	results of decision variant modelling (actual knowledge)
=rules of alternatives' choice + set of acceptable alternatives	all the best (from that which exist)	relevant (generalized) knowledge
=problem situations + set of their decisions examples in the form of subject collections	all useful (about what formalized information exists)	formalized knowledge about existing experience (precedent base)
=inference system + best alternative	best (possible) variant	decision on the base of digital intellectualizing

This classification allows creating a more efficient enterprise business model. This is achieved mainly due to the rational management of automation systems for physical operations of production and related business processes, integrated into united information space in accordance with the key subsystems of the Industry 4.0 concept (Product Lifecycle Management, Big Data, SMART Factory, cyber-physical systems, Internet of Things, interoperability).

VII. Conclusion

In this research the experience of intelligent decision support and recommender systems construction is systematized. The concept of generalized object is formulated for DSS engineering. Principles of recommender systems are systematized, and include possible types of data filtering. Their advantages and disadvantages are analyzed. An example of recommender systems for musical tracks choice for the different sport trainings is presented. Classification of DSS and their structure formulae in accordance with the used information data, models and knowledge are suggested. The use of these results made possible to reduce the time for creating models by an order of magnitude. Also suggested principles of DSS and recommender systems design make the systems more transparent, and the results are more justified and explainable. Received experience can be used in development of next-generation intelligent DSS.

Список литературы

[1] P. G. W. Keen, and M. S. Scott Morton "Decision support systems: an organizational perspective", Reading, Mass., Addison-Wesley Pub. Co., 1978. 264 p.

[2] P. G. W. Keen "Decision support systems: a research perspective", Cambridge, Massachusetts: Center for Information Systems Research, Alfred P. Sloan School of Management, 1980. 50 p.

[3] Sprague, R;(1980). "A Framework for the Development of Decision Support Systems". MIS Quarterly, 1980. Vol. 4, No. 4, pp. 1–25.

[4] "Intelligent Decision Support System for IoT-Enabling Technologies: Opportunities, Challenges and Applications" by ed. S. Sahana. Nova Science Publishers, Inc.; 1st edition, 2024. 334 p.

[5] L. Gaur, V. Agarwal, and P. Chatterjee "Intelligent Decision Support Systems for Smart City Applications", Wiley-Scrivener; 1st edition, 2022. 242 p.

[6] M. Sànchez-Marrè "Intelligent Decision Support Systems", Springer, 2022. 1317 p.

[7] S. Jain "Understanding Semantics-Based Decision Support", Chapman and Hall/CRC; 1st edition, 2021. 162 p.э

[8] G. Blokdyk "Intelligent Decision Support Systems: A Complete Guide – 2020 Edition", 5STARCooks, 2021. 311 p.

[9] C. C. Aggarwal "Recommender Systems: The Textbook", Springer, 2016. 498 p.

[10] F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor "Recommender Systems: The Handbook", Springer, 2011. 842 p.

[11] V. Golenkov, and N. Gulyakina "The Main Directions, Problems and Prospects of the Development of the Next-Generation Intelligent Computer Systems and the Corresponding Technology", Open Semantic Technologies for Intelligent Systems, 2023, Issue 7, pp. 15 – 26.

[12] S. A. Nikolenko "Recommender systems", Saint-Petersburg, Center of Speech Technologies Publishing, 2012. 53 p. (In Russian).

[13] B. A. Zhelezko, and O. A. Siniavskaya "Efficiency assessment of the rule and model bases in decision support systems", in Computer Data Analysis and Modeling: Complex Stochastic Data and Systems: Proc. of the Ninth Intern. Conf., Minsk, Sept. 7-11, 2010. In 2 vol. Vol. 2. Minsk: Publ. center of BSU, 2010, pp. 254 – 256.

[14] B. Zhelezko, O. Siniavskaya "Modeling technology of the liquidity evaluation on stock markets", Journal of Computational Optimization in Economics and Finance. 2012. Volume 4. Issue 2-3, pp. 187 – 197.

[15] B. A. Zhalezka, A. O. Stadnik, V. A. Siniauskaya "Complex 5-component oscillator as a tool of technical analysis in algorithmic trading", in Computer Data Analysis and Modeling: Stochastics and Data Science: Proc. of the XIII Intern. Conf., Minsk, Sept. 6-10, 2022, BSU; eds.: Yu. Kharin [et al.]. Minsk: BSU, 2022, pp. 228 – 232.

[16] B. A. Zhalezka, V. V. Kachanovich, V. A. Siniauskaya "Using of the neural network for securities' prices prediction", in Russian science: relevant research and engineering, proc. of the VIII All-Russian scientific and practical conference, 10 October 2019, in 2 volumes [ed. S.I. Ashmarina, A.V. Pavlova et al.]. Volume 1. Samara: Samara State Economic University, 2019, pp. 34 – 38. (In Russian).

[17] V. Siniauskaya, N. Viarbitsky, B. Zhalezka "Using of the concept of algorithmic marketing and machine learning in the marketing data processing", in Pattern Recognition and Information Processing PRIP'2019): Proceedings of the 14th International Conference, 21-23 May 2019, BSUIR, Minsk. Minsk: "Bestprint", 2019, pp. 380 – 383.

[18] B. Zhalezka, V. Siniauskaya "Theory and Practice of Multi-Agent Systems Construction", in Pattern Recognition and Information Processing (PRIP'2021): Proceedings of the 15th International Conference, 21–24 Sept. 2021, Minsk, Belarus. Minsk: UIIP NASB, 2021, pp. 106 – 110.

[19] REST architecture [Online]. Available: https://habr.com/ru/post/38730/ (In Russian).

[20] B. A. Zhalezka "Methodical and instrumental equipment of the management of info-communication structure development of industrial enterprises under conditions of regional integration and digitalization", in Proc. of the forum "Prospects for Eurasian Economic Integration", dedicated to 10th anniversary of the Eurasian Economic Commission within the framework of the 18th International Scientific Seminar "World Economy and Business Administration", XX International Scientific and Technical Conference "Science to education, production, economics". Minsk, March 16-17, 2022, BNTU. Minsk: Four quarters, 2022, pp. 115-117. (In Russian).

ПРИНЦИПЫ И ОПЫТ ПРОЕКТИРОВАНИЯ ИНТЕЛЛЕКТУАЛЬНЫХ СИСТЕМ ПОДДЕРЖКИ ПРИНЯТИЯ РЕШЕНИЙ И РЕКОМЕНДАТЕЛЬНЫХ СИСТЕМ

Железко Б. А., Синявская О. А.

В статье рассмотрены основные принципы проектирования интеллектуальных систем поддержки принятия решений (СППР) и рекомендательных систем. Сформулированы определение и концепция обобщенного объекта. Проанализированы технологии проектирования и разработки рекомендательных систем. Предложена классификация СППР. Представлен опыт проектирования СППР и рекомендательных систем.

Received 30.03.2024

The Properties Generality Principle and Knowledge Discovery Classification

Viktor Krasnoproshin
Faculty of Applied Mathematics and Computer Science
Belarussian State University
Minsk, Belarus
Email: krasnoproshin@bsu.by

Vadim Rodchenko and Anna Karkanitsa
Faculty of Mathematics and Informatics
Yanka Kupala State University of Grodno
Grodno, Belarus
Email: rovar@grsu.by, a.karkanica@grsu.by

***Abstract*—The paper examines the actual problem of automatic detection of hidden interpretable patterns in intelligent systems. The conceptual basis of the process of learning from examples is determined by the methods of class description and separation. Three basic principles are known: enumeration of class members, generality of properties and clustering. We propose an original method for implementing the principle of generality of properties based on the search for combinations of features that provide class distinction. The effectiveness of the approach is confirmed by the results of numerical experiment.**

***Keywords*—intelligent systems, pattern recognition, learning from examples, data mining**

I. Introduction

The development and large-scale implementation of information technologies has led to the accumulation of huge amounts of data, which today are organized into databases and data warehouses [1], [2]. Currently, the development of new methods aimed at improving the efficiency of representation and automatic knowledge extraction based on the analysis of large amounts of data is an urgent problem in computer science [3], [4].

The experience of using the structured query language (SQL) has shown its very limited capabilities in terms of discovering hidden patterns existing within the data. OLAP technology (interactive analytical data processing) is focused on the preparation of aggregated information on the basis of large data sets structured according to the multidimensional principle. At best, the technology provides for the extraction of knowledge from the data, which should be attributed to the "shallow" level of occurrence. The most interesting in practical terms are the hidden patterns, the detection of which is the focus of Data Mining [5]–[7].

In computer science, the problem of pattern recognition is one of the fundamental problems [8]–[10]. Its successful solution largely determines the progress in the field of artificial intelligence. Pattern recognition is the attribution of initial data to a certain class based on the selection of essential distinguishing features that characterize this data [11]–[14].

If a class is characterized by some common properties inherent in all its members, the construction of a recognition system can be based on the principle of generality of properties. The basic assumption in this case is that patterns of the same class have common properties reflecting their similarity [15], [16].

The paper proposes an original method of implementing the principle of generality of properties. It is assumed that as a result of analyzing the training data set (TS) it is possible to identify such a combination of features that ensure the distinction of classes. That eventually make the procedure of building a classification algorithm quite trivial. The effectiveness of the method is confirmed by the results of numerical experiment.

II. On Pattern Recognition Principles and Classification Problem

The basis of the idea of building automatic recognition systems are the methods of describing and separating classes (Fig. 1) [17].

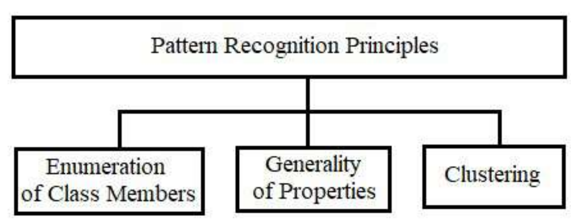


Figure 1. Principles of Pattern Recognition.

When a class is defined by an enumeration of its constituent objects, the construction of a pattern recognition system can be based on the principle of belonging to this enumeration. A set of objects of the class is memorized by the recognition system, and when a new object is presented to the system, it assigns it to the class to which the object located in the system's memory that matches the new one belonged (Fig. 2).

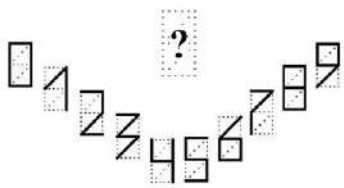


Figure 2. Enumeration of Class Members.

If all objects of one class have a number of common properties or features that are absent or have other values in all representatives of other classes, then the construction of the recognition system can be implemented on the basis of the principle of generality of properties (Fig. 3).

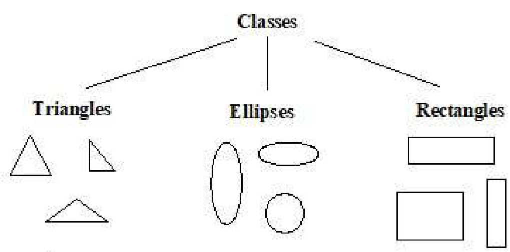


Figure 3. Generality of Properties.

When the objects of a class are vectors in the feature space, the class can be considered as a cluster. If clusters of different classes are separated far enough from each other, then the construction of the recognition system can be carried out using the clustering principle (Fig. 4).

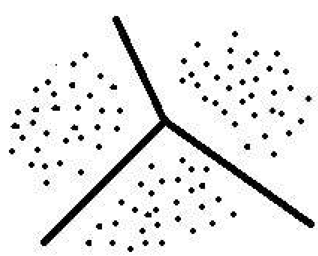


Figure 4. Clustering.

Traditionally, when building automatic pattern recognition systems, three main problems are solved. The first one is devoted to the issues of representation of the initial data obtained as a result of measurements of the recognized object. The second task is related to the extraction of essential features and properties from the initial data. The third one consists in finding optimal decision rules for classification [18], [19].

In [19], the author, discussing the problem of the simplicity of the learning process in pattern recognition, notes the existence of two different approaches to its implementation. In author's opinion, in the vast majority of studies (the first group), the learning process is aimed at constructing solving rules that ensure the extremum of a pre-selected criterion. In the second group, the focus is on understanding the principles of forming the description of recognition objects, within which the recognition process becomes extremely simple. Learning in this case is seen as a process of constructing a space that is universal, if not for all, then for a wide class of tasks. Unfortunately, in the author's opinion, this group of studies is very few and such an approach to solving the recognition problem is still poorly studied.

Today, pattern recognition is dominated by an approach in which training is reduced to solving an optimization problem. The training process begins with the selection of an initial model (a parametric family of algorithms), and then it is assumed that the "training + testing" scenario is repeatedly executed. In fact, training is an iterative process in which positive and negative reinforcements are used to form the desired patterns of classifier behavior.

In this case, it should be pointed out that there are at least two serious problems. First, model selection is a non-trivial task performed by a data science specialist, and therefore the training process can be implemented only in an automated mode. Second, the only result of training is a classification algorithm, which is an uninterpretable "black box".

It is proposed to consider an alternative approach, when the construction of the classification algorithm is performed not within the framework of the optimization problem, but on the basis of the analysis of the properties of the considered classes. As a result of such analysis, the distinguishing properties are determined by the mutual placement of the areas of class definition — patterns.

Before proceeding to the presentation of the alternative approach, let us consider the classical version of the mathematical formulation of the recognition problem.

In the paper by Y. I. Zhuravlev [20], the following formulation of the recognition problem (*classification or Z problem*) is given:

Let there be a set of admissible objects M. The set is covered by a finite number of subsets $K_1, \dots, K_l : M = \bigcup_{i=1}^{l} K_i$*, called classes. The partitioning of the set M is not completely defined, only some information* I_o *about classes* $K_1, \dots, K_l$ *is given. Similarly, an admissible object S is defined by the values of some characteristics. The set of given values defines the description I(S) of the object S.*

The main problem (problem Z) is to compute the values of predicates $P_j(S) - S \in K_j, j = 1, 2, \dots, l$*, from the information* $I_o(K_1, \dots, K_l)$ *and the description of the admissible object I(S).*

The information I_o is commonly called training information, and the predicates $P_j(S)$ are called elementary predicates.

In this formulation of the problem, it is actually required to construct some algorithm $A(I_0(K_1,\ldots,K_l),I(S)) = (\alpha_1^A(S)\ldots\alpha_l^A(S))$, where $\alpha_1^A(S)=P_j(S), j=1,2,\ldots,l$.

Obviously, the result of solving the problem Z is an algorithm with certain properties. In machine learning, the construction of such an algorithm is done within a scenario:

1) Some parametric family (model) of algorithms is selected;
2) The initial values of the parameters are fixed, and thus a specific algorithm is set;
3) The final setting of the algorithm to the subject domain is carried out during its training based on the training set data.

In this case, the learning process is reduced to the construction of algorithms (*decision rules*) that ensure the extremum of some criterion. Such a criterion, for example, can be the value of the average risk in a special class of decision rules. That is, at the beginning, the class of decision rules is defined up to parameters, and the training is reduced to finding the values of parameters that provide the extremum for the selected criterion.

Thus, in the most general form, the recognition problem can be written as follows: *The object description space is given, in which it is necessary to construct surfaces separating classes.*

In this formulation, the emphasis is on the construction of the algorithm *(on the construction of surfaces separating the classes)*, and therefore the problem has a pronounced algorithm-centric character.

In a more detailed analysis of the problem statement Z, we can propose an alternative variant of its formulation, when to find a solution the emphasis is shifted to the study of the property of classes. The new problem statement in this case is as follows.

Let two sets $I_0, I(S)$ be given, i. e., admissible training information $I_0(K_1,\ldots,K_l)$ and descriptions $I(S)$ of admissible objects $S\in M$, respectively.

It is required, based on the analyses of the of information $I_0(K_1,\ldots,K_l)$, to find the set of distinguishing qualities of classes $Q(K_1,\ldots,K_l)$ such that $K_i\cap K_j=\emptyset, \forall i\neq j$ (where $i,j=1,2,\ldots,l$) and using then the set $Q(S)$ to compute the values of predicates $P_j(S), j=1,2,\ldots,l$.

In this formulation, the solution of the problem is emphasized on the study of the property of classes and identification of features that provide class distinction. The recognition problem in this formulation is proposed to be called the Knowledge Discovery Classification Problem (KDC problem).

III. Method for Solving the KDC Problem

Let *M* be a set of objects, called admissible objects, and let it be covered by a finite number of subsets $K_1,\ldots,K_l : M=\bigcup_{i=1}^{l} K_i$ called classes. The partition *M* is not completely defined. Let an a priori dictionary of features $F=\{f_1,\ldots,f_n\}$ be given and on its basis only partial information $X=\bigcup_{i=1}^{l} X_i$ about classes $K_1,\ldots,K_l$ is given. Similarly, an admissible object *S* is defined on the basis of the features of the a priori dictionary.

The classification problem is to compute the values of predicates $P_j(S)-S\in K_j, j=1,2,\ldots,l$, based on the partial information X about classes $K_1,\ldots,K_l$ and the description of the admissible object *S*.

In the framework of the classical approach, the mathematical formulation of the classification problem is as follows: *Let X be the set of object descriptions and Y be the set of admissible classification answers. There is an unknown target dependency y*: X → Y, the values $X^m=\{(x_1,y_1),\ldots,(x_m,y_m)\}$ of which are known only for the objects of the training set. It is necessary to construct an algorithm a: X → Y, which would approximate this target dependency not only on the objects of the finite set, but also on the whole set X.*

To solve the problem, first the model of algorithms is specified up to parameters, and then training is carried out by finding such values of parameters that provide the extremum of the selected criterion. The experience of practical use of this scenario has revealed a number of problematic points.

The choice of a model of algorithms *A = a : X → Y* is actually a non-trivial task. In this case, it is not so much about science as about the art of algorithm construction [21], [22]. Moreover, learning can be realized only in an automated mode. And the final algorithm *a: X → Y* is a "black box". It approximates an unknown target dependency, which cannot be interpreted in terms of the subject domain.

The above drawbacks are avoided by using an alternative approach, which is based on the idea of the compactness hypothesis that classes form compactly localized subsets in the object space [23]–[25].

The mathematical formulation of the *Knowledge Discovery Classification Problem* in this case is as follows: *Let X be the set of object descriptions and Y be the set of admissible answers for their classification. There is an unknown target dependency y* : X → Y, the values of which $X^m=\{(x_1,y_1),\ldots,(x_m,y_m)\}$ are known only for the objects of the training set. It is required to find feature spaces in which classes do not intersect, and on their basis to construct an algorithm a : X → Y which would approximate this target dependency not only on the objects of the finite set, but also on the whole set X.*

The KDC problem is solved in two steps. First, the feature spaces in which the class patterns do not intersect are searched. After that, the construction of the classification algorithm becomes a simple procedure.

The initial data of the *Knowledge Discovery Classification Problem* are the alphabet of classes $K = \{K_1, \ldots, K_l\}$, a priori dictionary of features $F = \{f_1, \ldots, f_n\}$, training set $X_m = \{(x_1, k_1), \ldots (x_m, k_m)\}$, where k_i is the label of one of the classes of the alphabet K.

Let us denote by $V = \{v_1, \ldots, v_q\}$ the set of all possible combinations of features from F. In total V contains $q = \sum_{i=1}^{n} C_n^i = 2^n - 1$ different subsets.

At first glance, the solution to the KDC problem should involve performing a brute-force search on the set V. However, using the properties of combinations of features the search can be significantly reduced.

Let us demonstrate by concrete examples what properties of class distinction can be possessed by features and combinations of features of the dictionary F. Suppose that two classes of objects ***** and **+** are given, and for some pair of attributes f_i and f_j the distribution of objects of these classes is as follows (Fig. 5).

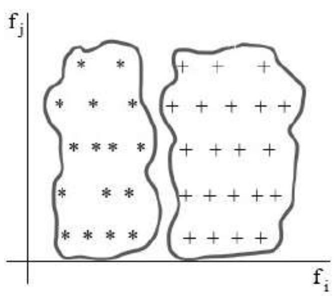


Figure 5. The first variant of mutual placement of objects.

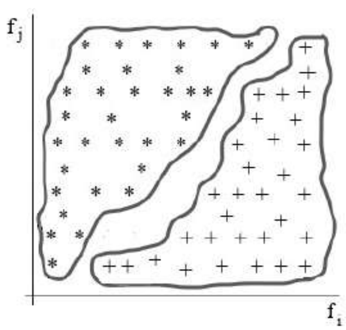


Figure 6. The second variant of mutual placement of objects.

Obviously, the feature f_i in Figure 5 has the property of distinguishing two classes, while the feature f_j has no such property. In addition, all combinations of features containing f_i, have the property of distinguishing two classes, i. e. the patterns of classes in the corresponding feature space do not intersect.

Figure 6 shows that each of the features f_i and f_j individually does not have the property of distinguishing between the two classes, while the combination of features f_i and f_j has such a property.

If each of the features of the original a priori dictionary has the property of class distinction, then the solution of the KDC problem is reduced from the brute-force search to the consideration of n variants.

The algorithm of searching for combinations of features on the set $V = \{v_1 \ldots, v_q\}$ that provide class distinction, as a result of the following steps:

Step 1. Select a subset $V^+ = \{v_1^+, \ldots, v_+^i\}$ of V, where v_i^+ contains only one feature.

Step 2. For each v_i^+ we construct class patterns (class definition areas) and compare their mutual placement [26].

Step 3. If class patterns do not intersect, then feature v_i^+ is included in the set $V^* = \{v_1^*, \ldots, v_k^*\}$.

Step 4. Exclude from the set $V = \{v_1, \ldots, v_q\}$ the subset $V^+ = \{v_1^+, \ldots, v_n^+\}$ and get $V^\Delta = \{v_1^\Delta, \ldots, v_p^\Delta\}$.

Step 5. Exclude from V^Δ all combinations of v_i^Δ, that contain any combination from $V^* = \{v_1^*, \ldots, v_k^*\}$.

Step 6. Take the next combination v_i^Δ from V^Δ and build a feature subspace based on it.

Step 7. In this feature subspace, we construct class patterns and compare their mutual placement.

Step 8. If the class patterns do not intersect, we include the combination of features v_i^Δ in the set V^*, and exclude from V^Δ all combinations that contain v_i^Δ.

Step 9. Repeat the process until V^Δ is empty.

The algorithm will result in the set $V^* = \{v_1^*, \ldots, v_t^*\}$, where $0 \leq t \leq q$. Based on the combinations $v_i^* \in V^*$, we formulate the previously hidden and empirically revealed regularities: «in the feature space of a subset v_i^* classes do not intersect».

Note that within the framework of solving a particular applied problem, all combinations of features v_i^* can be interpreted in terms of the subject domain.

Combinations of features $v_i^* \in V^*$ define decision spaces in which class patterns do not intersect. In such spaces, the compactness hypothesis condition is satisfied and classification is performed by the rule:

- for each combination of features $v_i^* \in V^*$ (where $i = 1, 2, \ldots, t$) and based on the training set data we build cluster structures $P_1^i, \ldots, P_l^i$ — patterns of classes $K_1, \ldots, K_l$ [26];
- investigated object $S \in K_m$ if $S \in P_m^i \forall i = 1, 2, \ldots, t$.

To exemplify the demonstration of the generality principle in Figure 3, the following variant of the classification algorithm construction can be proposed:

There is an alphabet of classes K={Triangles, Ellipses, Rectangles} and an a priori dictionary of features F={area of figure, perimeter of figure, number of angles}.

It is obvious that only the feature "number of angles" has the property of class distinction, because for all figures of class Triangles the value of the feature is equal to 3, for class Ellipses - equal to 0, for class Rectangles - equal to 4. Hence $V^* = \{number\ of\ angles\}$ and the classification algorithm is as follows:

IF (number of angles = 3) THEN Triangles

ELSE IF (number of angles = 0) THEN Ellipses

ELSE Rectangles

IV. Example of Solving the KDC Problem

Let's demonstrate the results of solving the KDC problem based on model data. Let's say we're given:

- classes **Not3** (there is no digit 3 in the number) and **Yes3** (there is at least one digit 3 in the number);
- a priori dictionary of features *F = {Units, Tens}*;
- training set, which consists of 250 two-digit integers, among which 200 have no digit 3 and 50 have at least one digit 3.

Table I shows the feature values of *Units* and *Tens* in the training set used in the numerical experiment.

Table I. Values of Units and Tens in the training set

Units / Tens	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
0	2	2	2	3	2	3	2	3	2	3
1	3	3	2	3	2	3	3	2	3	2
2	2	3	2	3	2	2	2	3	3	2
3	3	2	2	2	3	2	3	2	2	3
4	2	3	2	3	2	3	2	3	3	2
5	3	2	3	3	3	3	3	2	3	2
6	3	2	3	2	3	2	3	2	3	2
7	2	3	3	3	2	3	2	3	2	3
8	3	2	2	3	3	2	2	2	3	2
9	2	3	2	3	3	3	2	2	2	3

Table II summarizes the results of class pattern intersection study based on the features *Units*, *Tens*, where $Not3_i$ is the number of $Not3_i$ class representatives for the i-th digit; $Yes3_i$ is the number of $Yes3_i$ class representatives for the i-th digit.

Table II shows that neither the Units feature nor the Tens feature provides an absolute separation between the **Not3** and **Yes3** classes.

Table III summarizes the results of the study on the intersection of class patterns based on the combination of features *(Tens, Units)*.

Table III shows that:

- all numbers of class *Not3* have no digit 3, and all numbers of class *Yes3* have at least one digit 3;
- combination of features *(Tens, Units)* provides absolute separation of *Not3* and *Yes3* classes;

Table II. Results for features Units and Tens

	Units		**Tens**	
Digit	**Not3**	**Yes3**	**Not3**	**Yes3**
0	22	3	21	3
1	23	3	23	2
2	21	3	21	2
3	0	2	0	24
4	22	3	22	3
5	24	3	24	2
6	21	2	23	3
7	22	3	23	2
8	24	3	21	2
9	21	3	22	3

Table III. Results for the combination of features (Tens, Units)

T,U	N3	Y3	T,U	N3	Y3	T,U	N3	Y3	T,U	N3	Y3
0,0	2	0	2,5	2	0	5,0	3	0	7,5	3	0
0,1	2	0	2,6	2	0	5,1	2	0	7,6	2	0
0,2	2	0	2,7	3	0	5,2	3	0	7,7	3	0
0,3	0	3	2,8	3	0	5,3	0	3	7,8	2	0
0,4	2	0	2,9	2	0	5,4	3	0	7,9	3	0
0,5	3	0	3,0	0	3	5,5	3	0	8,0	3	0
0,6	2	0	3,1	0	2	5,6	3	0	8,1	2	0
0,7	3	0	3,2	0	2	5,7	2	2	8,2	2	0
0,8	2	0	3,3	0	2	5,8	3	0	8,3	0	3
0,9	3	0	3,4	0	3	5,9	2	0	8,4	3	0
1,0	3	0	3,5	0	2	6,0	3	0	8,5	2	0
1,1	3	0	3,6	0	3	6,1	2	0	8,6	2	0
1,2	2	0	3,7	0	2	6,2	3	0	8,7	2	0
1,3	0	3	3,8	0	2	6,3	0	2	8,8	3	0
1,4	2	0	3,9	0	3	6,4	3	0	8,9	2	0
1,5	3	0	4,0	2	0	6,5	2	0	9,0	2	0
1,6	3	0	4,1	3	0	6,6	3	0	9,1	3	0
1,7	2	0	4,2	2	0	6,7	2	0	9,2	2	0
1,8	3	0	4,3	0	3	6,8	3	0	9,3	0	3
1,9	2	0	4,4	2	0	6,9	2	0	9,4	3	0
2,0	2	0	4,5	3	0	7,0	2	0	9,5	3	0
2,1	3	0	4,6	2	0	7,1	3	0	9,6	2	0
2,2	2	0	4,7	3	0	7,2	3	0	9,7	2	0
2,3	0	3	4,8	3	0	7,3	0	3	9,8	2	0
2,4	2	0	4,9	2	0	7,4	2	0	9,9	3	0

- the definition areas of *Not3* and *Yes3* classes do not intersect and are presented below (Fig. 7).

The classification algorithm is built on the basis of a rule: *IF (Units = 3 or Tens = 3) THEN Yes3 ELSE Not3.*

So, as a result of solving the KDC problem in automatic mode, the training data set were analyzed. The initially hidden regularity *«the combination of features (Tens, Units) provides the distinction between classes Not3 and Yes3 by the rule IF (Units = 3 or Tens = 3) THEN Yes3 ELSE Not3»* was found, and the classification algorithm was built on its basis.

V. Conclusion

The paper presents an original approach for solving the problem of learning from examples which is based on the use of the properies generality principle. The method of the principle implementation is proposed which provides

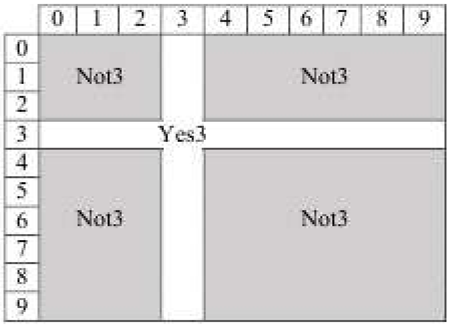


Figure 7. Definition areas of Not3 and Yes3 classes.

for automatic detection of hidden interpretable patterns in the training data set. The revealed patterns can be used then to construct a classification algorithm.

The learning algorithm for identifying combinations of features that have the property of class distinction is described. As a result of analyzing the training data set, the informativeness estimates of combinations of distinguishing features (from the point of view of classes) are automatically calculated and a classifier is built.

Based on model data, the results of applying the developed method to solve the classification problem are presented.

References

[1] D. Gurvits, A. N'yudzhent, F. Khalper Prosto o bol'shikh dannykh [Big Data For Dummies], Moskow, Eksmo. 2015. 400 p.

[2] A. Vaigend, Big Data. Vsya tekhnologiya v odnoi knige [Big Data. All the technology in one book], Moskow, Eksmo. 2018. 384 p.

[3] N. Marts, Dzh. Uorren Bol'shie dannye: printsipy i praktika postroeniya masshtabiruemykh sistem obrabotki dannykh v real'nom vremeni [Big Data: Principles and best practices of scalable realtime data systems], Moskow, OOO "I.D. Vil'yams", 2016. 368 p.

[4] D. Silen, A. Meisman, M. Ali Osnovy Data Science i Big Data. Python i nauka o dannykh [Introducing Data Science: Big Data, Machine Learning, and more, using Python tools], Saint Petersburg, Piter. 2017. 336 p.

[5] A.A. Barsegyan, M.S. Kupriyanov, I.I. Kholod Analiz dannykh i protsessov [Data and Process Analysis], Saint Petersburg, BKhV-Peterburg. 2009. 512 p.

[6] A.G. D'yakonov Analiz dannykh, obuchenie po pretsedentam, logicheskie igry, sistemy WEKA, RapidMiner i MatLab [Data analysis, learning by precedents, logic games, WEKA, RapidMiner and MatLab systems], Moskow, Izdatel'skii otdel fakul'teta VMK MGU imeni M.V. Lomonosova. 2010. 278 p.

[7] Dzh. Gras Data Science. Nauka o dannykh s nulya [Data Science from Scratch], Saint Petersburg, BKhV-Peterburg. 2017. 416 p.

[8] K. Fukunaga Vvedenie v statisticheskuyu teoriyu raspoznavaniya obrazov [Introduction to Statistical Pattern Recognition], Moskow, Nauka, 1979. 368 p.

[9] K.M. Bishop Raspoznavanie obrazov i mashinnoe obuchenie [Pattern Recognition and Machine Learning], Moskow, Dialektika, 2020. 960 p.

[10] V.V. Myasnikov Raspoznavanie obrazov i mashinnoe obuchenie. Osnovnye podkhody [Pattern recognition and machine learning. Basic approaches], Samara, Izdatel'stvo Samarskogo universiteta, 2023. 196 p.

[11] A.D. Zakrevskii Logika raspoznavaniya [Recognition logic], Minsk, Nauka i tekhnika, 1988. 118 p.

[12] N.G. Zagoruiko Prikladnye metody analiza dannykh i znanii [Applied methods of data and knowledge analysis], Novosibirsk, IM SO RAN, 1999. 270 p.

[13] A.V. Bobkov Sistemy raspoznavaniya obrazov [Pattern recognition systems], Moskow, MGTU imeni N.E. Baumana, 2018. 187 p.

[14] V. Krasnoproshin, A. Karkanitsa, V. Rodchenko Pattern Recognition Based on Classes Distinctive Features. Proc. of 15-th International Conference "Pattern Recognition and Information Processing", 2021, pp. 22-25.

[15] V.I. Vasil'ev Problema obucheniya raspoznavaniyu obrazov : Printsipy, algoritmy, realizatsiya [The problem of teaching pattern recognition: Principles, algorithms, implementation], Kiev, Vyshcha shkola, 1989. 64 p.

[16] V. Krasnoproshin, V. Rodchenko Obuchenie po pretsedentam na osnove analiza svojstv priznakov [Learning by precedents based on the analysis of the features]. Doklady BGUIR, 2017, no 6, pp. 35-41.

[17] Dzh. Tu, R. Gonsales Printsipy raspoznavaniya obrazov [Principles of pattern recognition], Moskow, Mir, 1978. 411 p.

[18] S. Rashka Python i mashinnoe obuchenie [Python and machine learning], Moskow, DMK Press, 2017. 418 p.

[19] V.V. Krasnoproshin, V.G. Rodchenko Klassifikatsiya na osnove prostranstv reshenii [Classification based on decision spaces]. Doklady BGUIR, 2019, no 6, pp. 20-25.

[20] Yu.I. Zhuravlev Ob algebraicheskom podkhode k resheniyu zadach raspoznavaniya ili klassifikatsii [On an algebraic approach to solving recognition or classification problems]. Problemy kibernetiki [Problems of cybernetics], 1978, no 33, pp. 5-68.

[21] P. Flakh Mashinnoe obuchenie. Nauka i iskusstvo postroeniya algoritmov, kotorye izvlekayut znaniya iz dannykh [Machine learning. The science and art of building algorithms that extract knowledge from data], Moskow, DMK Press, 2015. 400 p.

[22] V. Krasnoproshin, A. Karkanitsa, V. Rodchenko Implementation of the KD-Agent for Knowledge Ecosystem. Otkrytye semanticheskie tekhnologii proektirovaniya intellektual'nykh system [Open semantic technologies for intelligent systems], 2021, pp. 59-62.

[23] L.A. Belozerskii Sovremennyi vzglyad na gipotezu kompaktnosti [Modern view of the compactness hypothesis]. Iskusstvennyi intellekt [Artificial intelligence], 2005, no 3, pp. 6-12.

[24] V. Rodchenko Pattern Recognition: Supervised Learning on the Bases of Cluster Structures. Proc. XIII International Conference "Pattern Recognition and Information Processing", 2016, pp. 106-113.

[25] V. Rodchenko Automatic Detection of Hidden Regularities Based on the Study of Class Properties. Pattern Recognition and Image Analysis, 2020, vol. 30, no 2, pp. 224-229.

[26] V.V. Krasnoproshin, V.G. Rodchenko Klasternye struktury i ikh primenenie v intellektual'nom analize dannykh [Cluster structures and their application in data mining]. Informatika [Informatics], 2016, no 2, pp. 71-77.

ПРИНЦИП ОБЩНОСТИ СВОЙСТВ И KD-КЛАССИФИКАЦИЯ

Краснопрошин В. В., Родченко В. Г., Карканица А. В.

В работе исследуется актуальная проблема автоматического обнаружения скрытых интерпретируемых закономерностей в интеллектуальных системах. Концептуальную основу процесса обучения по прецедентам определяют способы описания и разделения классов. Известны три базовых принципа: перечисления членов класса, общности свойств и кластеризации. Предлагается оригинальный метод реализации принципа общности свойств, основанный на поиске сочетаний признаков, обеспечивающих различение классов. Эффективность подхода подтверждается результатами численного эксперимента.

Received 14.03.2024

Interoperability as a Critical Component of the Educational Process in Secondary Schools

Alena Kazlova and Alexander Halavaty
Belarusian State University
Minsk, Belarus
Email: kozlova@bsu.by

***Abstract*—The paper presents some results of an analysis of the role of the development of interoperability, cognitive abilities and emotional intelligence in children in a modern school. The importance and ways of introducing technological tools with capabilities for interaction and data exchange to optimize the educational process are discussed. The significance of the development of cognitive abilities and emotional intelligence of students and the impact of this on their academic achievements and social adaptation are also considered.**

***Keywords*—Education, interoperability, cognitive abilities, emotional intelligence, intelligent educational systems**

I. Introduction

School education is an initial and very important stage, which forms the basis for all further development of an individual's education, exerting a significant influence not only on his future activities as a specialist in a particular sector of the economy, but also as an individual, a member of society. It is at school that "the intellect is formed, which is a combination of various functions (sensory-perceptual, mnemological and attentive)" [1], and the personality itself is formed. It is the complex development of the individual, the combination of professional, technical, and personal, universal knowledge and skills that determine the success of a modern person, his ability not only for his own development and improvement, but also for the development and improvement of society in all forms of its existence [2].

In the process of digitalization of education, a variety of tools, methods and technologies are used, largely based on the use of artificial intelligence algorithms. Such types of training as network and electronic are being developed, which include not only the direct educational part, but also means of automating the learning process itself. At the same time, often "behind the scenes" there remains such an important issue as the education of an individual capable of thinking creatively, being able to organize one's own learning process, and also working in a team, distributing tasks, negotiating, explaining his point of view, justifying his decisions. In fact, when considering the development of digital platforms and intelligent educational systems, we should talk about the human interoperability. It is becoming increasingly clear that interoperability is necessary both to create a single barrier-free information space based on the principles of openness, transparency, multi-purpose use of data, technological neutrality, and to ensure the priority of user interests, information security and protection of privacy [3]. The study of the development of the properties of interoperability, cognitive abilities and emotional intelligence starting from childhood is increasingly relevant with the development of society and the transition to the sixth technological order, in which the leading role is given, along with information and nanotechnologies, to cognitive sciences and socio-humanitarian technologies.

The main goal of the present paper is to give an analytical overview of the role and methods of development of secondary school students' cognitive abilities along with the emotional intelligence and interoperability. Another idea was to determine some basic concepts within the interoperability as a subject area and the property of a school student.

II. The cognitive abilities level and the interoperable behaviour. State of art

According to the American Psychological Association Dictionary of Phychology, cognitive ability is defined as the skills involved in performing the tasks associated with perception, learning, memory, understanding, awareness, reasoning, judgment, intuition, and language [4].

People are engaged at every step in the data value chain in collecting, analyzing, interpreting, and using data. In many cases, people themselves are data points. All these people bring perspectives, values, world views, and expectations, which are also embedded in political and organizational cultures. If we want data to work together, we need people to work together. We need human interoperability [5], [6].

Emotional Intelligence (EI) is the ability to manage both your own emotions and understand the emotions of people around you. There are five key elements to EI: self-awareness, self-regulation, motivation, empathy, and social skills. People with high EI can identify how they are feeling, what those feelings mean, and how those emotions impact their behavior and in turn, other people. It's a little harder to "manage" the emotions of other people — you can't control how someone else feels or behaves. But if you can identify the emotions behind their behavior, you'll have a better understanding of where they are coming from and how to best interact with them [7].

Let us consider some characteristics of preschoolers and schoolchildren, depending on their age, from the point of view of developing the ability for interoperable behavior.

Even in preschool age — about 4-5 years — the child begins to understand that other people may have opinions, thoughts, and desires that are different from his own. This ability to understand and accept differences in people's thinking develops as we grow older. Some researchers note that the better developed such mental empathy is, the higher a person's academic performance at school and university [1], [2], [4], [5], [6], [7], [8], [9], [10]. This ability helps build relationships of mutual understanding and involvement, participation between mentors and classmates, leading to the perception and understanding of tasks and requirements.

The younger schoolchild (6 – 10 years old) is characterized primarily by readiness for educational activities, i. e. he is ready to study systematically. It is also the ability to accept new responsibilities, which underlies the educational motivation of a primary school student. This period is the most important for the development of aesthetic perception, creativity and the formation of a moral and aesthetic attitude towards life, which is fixed in a more or less unchanged form for the rest of life. In elementary school, the younger student develops forms of thinking that ensure the further assimilation of various knowledge and the development of thinking. At this age, you can also develop the student's self-organization and self-discipline skills, for example, through group games, encouraging healthy curiosity, and interest in all kinds of creative activities.

In middle school age (from 10–11 to 14–15 years), communication with peers plays a decisive role. The leading types of activities are educational, social and organizational, sports, creative, and labor [1]. During this period, the child acquires significant social experience and begins to comprehend himself as an individual in the system of labor, moral, and aesthetic social relations. He has a deliberate desire to take part in socially significant work, become socially useful, and interact in a team.

In the period of early adolescence (from 14–15 to 17 years), value-orientation activity, which is determined by the desire for independence, acquires key importance. The main components of this period are friendship and trusting relationships. This is the stage that many authors call the final stage of personality maturation [1], [2], when professional interests are clearly formed, theoretical thinking, the ability for self-education, the ability to reflect are developed, the level of aspirations is formed. At this stage, a person is able to formulate his demands, interact with other members of the team, forming personal connections, for example, friendship, sympathy.

Each of the stages of learning and growing up should be accompanied by the acquisition of interaction skills in teams, not so much for the purpose of competitively achieving the result of joint activity, but for the purpose of teaching a person to look for the best, including joint, solutions to assigned tasks. That is, the main goal and task of an interoperable person is the search and implementation of the best solution in the given conditions and given the available opportunities, and not a competitive struggle for the implementation of one's own idea. At the same time, this approach should not teach children to abandon their own position, opinion, or belittle their ideas and achievements. An interoperable personality with a high level of emotional intelligence must be able to combine the ability to appreciate the personal and the collective.

Successful learning, cognitive abilities and emotional intelligence are deeply interconnected. There are many studies aimed at studying cognitive functions and their importance in the cognitive process [7], [8], [9], [10], [11]. Thus, work [7] shows that academic performance in various subjects (mathematics, reading, writing) has a strong dependence on the level of development of cognitive functions. The same work indicates that working memory affects academic performance more than intelligence level (IQ).

In [8], the authors provide research data on the development of cognitive abilities depending on age. The results were obtained based on an analysis of data from four large research projects, which tested about 11,000 people aged 8 to 35 years. It was noted that the most rapid development of the executive abilities of the brain occurs at 10-15 years of age; at 15-20 years of age, development slows down, and by the age of 20, cognitive functions begin to stabilize and reach their maximum level of development. Next, the person uses those skills—the executive functions of the brain—that he acquired at an earlier age. This is a clear confirmation of the need to develop cognitive abilities starting from preschool and especially at school age. It is during this period that a strong foundation can be laid for further successful cognitive and creative human activity.

Interoperability refers to the ability of different systems, programs and technologies to interact and exchange data without communication and semantic difficulties. From a technical point of view, in the context of school education, this could mean that different educational platforms, applications and resources need to be interoperable and able to exchange information with each other. This allows students and educators to use a variety of tools and resources to enhance learning and knowledge sharing. According to information technology standards, for example [12], "interoperability is the ability of two or more information systems or components to exchange information and to use information obtained as a result of the exchange". With the addition of the ability to discuss and negotiate, this concept can be extended to both users and development engineers who work on creating information systems for various purposes, involving in the development process specialists from areas related to information technology, as well as from those areas, for which this or that system is being developed. And in their work, it is the ability to interact, i. e., interoperability, that plays a key role in obtaining a high-quality new product.

Currently, not a single area of human activity can do without the use of information and communication technologies. The interoperability is one of the basic properties, without which further formation and development of the information society will become impossible. Today, in all spheres of human activity, there is an ever-accelerating transition to working with a large number of information systems, network resources, and computing power. This leads to interpenetration and the necessary increasingly conscious and deep interaction of knowledge of specialists from various fields, to the need to understand requests and solutions, and collaborate on tasks. And the level of such interaction will only increase with increasing diversity and integration of knowledge.

A. The role of intelligent information technologies in the development of cognitive abilities, emotional intelligence and interoperability

Let us dwell further on the impact of the use of digital information space on learning. Today, educational systems are built on the basis of methods and technologies of artificial intelligence, which is increasingly penetrating our lives, including education. All of the above puts forward additional requirements for the information systems used in the learning process.

Artificial intelligence, of course, gives us access to broad and detailed information for a various range of issues, but most information systems are organized as search engines. The introduction artificial intelligence technologies, including neural network technologies, into the educational process and human learning activities is a very successful and promising solution. Far instance, there are more than 200 publicly available neural networks. Certain types of neural networks are designed to search for information, check and edit texts, generate images, create curricula, programs and presentations. These neural networks are capable to help teachers draw up curriculum, create tests, organize the structure of a lesson/lecture/seminar, and provide assistance in assessing knowledge. For students, they can help with identifying topics to study and selecting materials on these topics [17].

A very illustrative example of how artificial intelligence should be used in training is the example given by Eric Ofgang [18] when comparing the Gemini and ChatGPT neural networks. The author of the article asked the neural network to write an essay for him on a certain topic. In response, the Gemini chatbot suggested a possible outline for the essay. When asked again to write the text, it offered to help as a "writing coach" and gave some tips on how to write an essay on a given topic himself. This example clearly demonstrates progress in the development of approaches to the use of intelligent technologies for obtaining and processing information in the learning process. Neural networks should not do the original, creative part of the work for a person. They should be assistants both in searching for information and in learning, acquiring new knowledge and skills. The same approach should be implemented in general when creating intelligent teaching systems for any level of education.

III. Semantic and cognitive-emotional approach as a basis for educational intelligent systems building

In our days there is no doubt about the crucial necessity of the intelligent learning systems application at all levels of the education system. Along with that the wide variety of different technologies of such systems construction and operation drives to a number of problems for their actual application at schools — the users are demanded to start from the beginning each time with some new system using some new technology. Thus, the problem has its solution - usage of common technology to systems development and application, which would give an opportunity for the users with different skill-level apply such systems while their studies. Thus, the interoperability is not only technical systems property. It should be considered as one of the main skills of any specialist, and the basis for such skills must be formed within the school education process. Considering the issue of developing emotional intelligence and interoperability along with cognitive abilities, we can consider options for using modern intellectual technologies in this direction at school. The first option is to develop a single digital platform that would integrate various educational resources, applications and tools, providing students and teachers with the ability to access a variety of materials and tasks without the need to constantly switch between different systems.

The OSTIS-Technology [19] is the example of such a basic, common technology of interoperable systems building. It provides a unified semantic platform making it possible to exchange results and data, not only within the technical part of the educational systems, but also present information to the users — children, teachers and parents — and ensure interoperability among them, as long as among the specialists and the public.

Within the framework of this work, it is proposed to take as a basis the approaches to ontology development. Some main features of the cognitive skills corresponding and deeply influencing the development of emotional intelligence and interoperability of the school students will be discussed as the objects for formal clarifying and agreement within the framework of the corresponding set of semantic interpretation.

IV. The process of cognition, the development of emotional intelligence and interoperability

To formalize the basics of any field it is important to find those that play the most important role and agree on their understanding. Considering the interoperability as a property of a student we should clarify the essential aspects that impact its formation. In [20], we have already presented some views on this point — the individual approach to learning, planning, drawing up programs at various levels of education and interdisciplinary connections. But without the ability of students to perceive this information, there will be no result; at graduation we will not see a trained person capable of solving the problems facing him. Therefore, now we would like to touch on some general approaches and methodological techniques aimed at more successful acquisition of knowledge and development of mental executive abilities of school students. The main points that we would like to draw attention to are the modeling of reality, associativity, completeness and consistency of knowledge, stimulation of cognition.

Let's first consider reality modeling. When the new material taught or the formulas used are related to what is encountered and can be applied in real life, then such material is learned much more successfully. The child learns to count, divide, etc. on apples, candies, just common things he uses every day. This skill remains firmly in his head for the rest of his life. Therefore, at school, even in high school, it is necessary to provide material that is confirmed directly or indirectly by examples from life, which the student can discuss with parents and friends. There is no need to strive to present the latest scientific achievements at school, even if they are super important for some area of knowledge. For example, if you present in a botany lesson about 20 layers that make up the bark of a tree, then a diligent student will, of course, memorize their names to answer the test, but they will not remain in memory for a long time, because there is no confirmation in ordinary life around them, there is no immediate interest in such depth of the issue. Unless this student is deeply interested in botany, but this is a different case. Some researchers attribute better performance in humanitarian subjects and the fact that the majority of graduates want to connect their future activities with the humanitarian field precisely to the fact that humanitarian subjects at school, unlike natural sciences, consider issues that are close to most people and related life situations.

Another approach to teaching methods is directly related to the modeling of reality — associativity. If you learn to remember new information based on associations, then this process will become faster, more effective for storing in long-term memory and applying this knowledge at the right time.

The issues considered in the learning process must be complete and complete at the level of possible perception of the audience. You cannot give part of the conclusion, then reduce it to the fact that the formula is too complex and therefore we use a simple approximation or do not take this into account.

Typically, educational material is taught in sequence: a descriptive part, applied formulas and laws, and completed tasks. There must also be feedback. This means that when going through subsequent topics, we must return to this material in order to justify new information on its basis. Repeated repetition of material over an increasing period of time, application of previously studied material in other topics, subjects and complex tasks is one of the foundations for its successful memorization and assimilation.

Also, for a greater degree of involvement in the learning process in order to obtain a higher level of knowledge, it is necessary to use various types of stimulation for cognition — additional points for completing a task, access to the next tasks, the next level (additional functions, like in games), competition tables for completing and leaderboards, etc.

Interoperability and emotional intelligence are two key aspects that are essential in school education today. Emotional intelligence refers to the ability to manage your own emotions and the emotions of others. In school education, it becomes important because it helps students develop skills in self-regulation, empathy, social competence and conflict resolution. Emotional intelligence training can be built into curricula through social-emotional learning programs that include skills such as managing stress, developing an understanding of emotions, and how to interact effectively with others.

Integrating interoperability and emotional intelligence in school education can lead to more flexible and adaptive educational environments that take into account the needs and individual characteristics of each student. Technology can assist in this process by providing access to a variety of educational resources and tools, as well as supporting the development of social-emotional skills.

The use of common data exchange standards to transfer information about the progress, learning process and individual needs of students between different educational platforms will allow, for example, data on a student's progress in any online course to be automatically transferred to the school information system for teachers. The use of adaptive learning algorithms in intelligent teaching systems will make it possible to analyze data on a student's performance and learning style, and then offer personalized tasks and recommendations. Such systems can be integrated with various educational platforms to ensure a continuous flow of data. At the same time, the use of interoperable tools for students to collaborate on projects and exchange materials, such as documents, presentations and videos in real time, will contribute to the development of interoperability among students in the process of interacting with each other and with the system itself.

In addition to directly teaching functions, intelligent systems in educational activities can also perform administrative and technical ones. For example, implement adaptive management of classes and schedules. The use of a classroom management system, which can be integrated with a class schedule, electronic journal and other educational resources, will ensure effective planning and organization of the educational process in educational institutions of different levels. Thus, interoperability can improve the efficiency and functionality of educational systems, providing more flexible and adaptive learning for students and teachers.

To develop emotional intelligence, and as a result, interoperability, students can be taught awareness, empathy, social competence, and develop self-regulation and conflict management skills. Mindfulness training will help one develop his ability to consciously perceive his thoughts, emotions, and physical sensations. This may include breathing exercises, exercises for awareness in movement and other techniques that can be taught in physical education classes, creating a culture of interaction with oneself and acceptance of oneself on different levels - physical, emotional, spiritual. This will in turn promote the teaching of empathy and social competence through group exercises, role plays, case discussions or collaborative projects across school disciplines where students are provided with opportunities to understand others' emotions, develop empathy and improve social interaction skills.

The development of self-regulation includes learning strategies for managing one's emotions and reactions in various situations, time management, planning actions in stressful situations, developing problem-solving strategies, etc. Learning conflict management — constructive resolution based on the skills of active listening, expressing one's emotions and searching compromises. This may include educational games, role-playing games and conflict resolution training.

Teaching emotional intelligence can be integrated into various academic subjects, for example through analyzing literary characters and their emotional reactions, discussing ethical issues, or studying historical events with an emotional perspective. An important part of this work is also providing support and training to teachers and parents so that they can support the development of students' emotional intelligence. It should be recognized that for many teachers and parents, the experience of developing emotional intelligence and interoperability is a new, unknown matter that also needs to be learned. To train them, it is possible, for example, to conduct seminars, training, and provide teaching resources to teachers and parents. These methods and strategies can be integrated into the learning process and school life to support the development of students' emotional intelligence and help them successfully cope with emotional challenges, increasing their competencies, strengthening their abilities, and becoming successful and happy members of society.

V. Basic concepts to be formalized

The task of formalizing concepts in the field of interoperability is complex and requires the participation of psychologists, physiologists and teachers. The list of the items to formalize at first stages can be formed according to the discussion above. Two ways are possible - from the definition of the emotional intelligence and cognitive properties down to the basics, and the other one - from the main conceptual terms up to the understanding of emotional concepts and interoperability. Thus the definitions of the concepts like perception, learning, memory, understanding, awareness, reasoning, judgment, intuition should be determined and agreed. The next step is to derive the contribution of each term to the whole understanding of the level of the cognitive skills, as well as the methods to determine this level. There will appear a hierarchical structure of concepts leading to a description of the phenomenon (ability) of cognition.

The human side of interoperability is a hard part to define. It requires time, energy, and money [5]. The words used here are not basics. There is a need to make a precise research on the issues that contribute to each of them and formalize their meanings, make them acceptable for the information system determination and measuring.

Considering the emotional intelligence we are talking about such concepts as self-awareness, self-regulation, motivation, empathy, and social skills, which are complex and require some investigations on a possibility to identify both their formal meanings as well as options and principals of their physical measuring the degree of their manifestation in a person.

VI. Заключение

The ability for interoperable behavior is not genetically embedded in people. It must be raised, starting from an early age, in accordance with children's ability to perceive tasks that require interaction and joint activity to solve them. Interoperable behavior requires mastery of conceptual apparatus from different scientific fields, the ability to formulate descriptions of phenomena, problems, tasks that a person faces, ways to solve them, taking into account the interests of different people, the characteristics of their thinking and behavior. Thus, modern school education should not just provide knowledge in individual sciences, but educate a cognitive personality interested in knowledge, its development, creativity, and interaction with the environment, be it intellectual systems or people. Interoperable intelligent training systems based on new generation computer systems can play a significant role in this.

The article presents an analysis of the importance and ways of developing cognitive abilities, emotional intelligence and the properties of interoperability at school age. Some proposals are presented for the introduction of these methods and methods into the educational process of secondary school. Some descriptions of basic concepts, processes and methods of the subject area "human interoperability" is proposed for their formalization and further construction of an ontology, knowledge bases and solvers of relevant problems within the framework of the technology for the development and use of OSTIS systems.

Acknowledgment

The authors would like to thank the scientific teams of the departments of intelligent information technologies of the Belarusian State University of Informatics and Radioelectronics and the department of intelligent systems of the Belarusian State University for their assistance in the work and valuable comments.

References

[1] A. N. Leontiev Psikhologicheskie osnovy razvitiya rebenka I obucheniya (digest) // Izdatelstvo "Smysl", 2009, P. 426 (in Russian)

[2] Available at: linkedin.com/in/tatsiana-kandrashova-business-analyst-a906b9131. Published February 24, 2021 (access date 2024-03-03).

[3] Zhuravlev M. S. Interoperabelnost kak factor razvitiya prava v sfere elektronnogo zdravookhraneniya. Pravo. Zhurnal Vysshey Shkoly ekonomiki. 2019. №3. (in Russian) Available at: https://cyberleninka.ru/article/n/interoperabelnost-kak-faktor-razvitiya-prava-v-sfere-elektronnogo-zdravoohraneniya (access date: 2024-03-03).

[4] American Psychological Association Dictionary of Phychology. American Psychological Association (electronic resource). Available at: https://dictionary.apa.org/cognitive-ability (access date 2024-03-20)

[5] Steven Ramage, Jenna Slotin Why people are essential in data interoperability. Global Partnership for Sustainable Development Data (electronic resource). Published August 25, 2021. Available at: https://www.data4sdgs.org/blog/why-people-are-essential-data-interoperability (access date 2024-03-20)

[6] World Development Report 2021. The World Bank (IBRD-IDA) (electronic resource). Available at: https://wdr2021.worldbank.org/stories/the-social-contract-for-data/ (access date 2024-03-20)

[7] What is emotional intelligence and how does it apply to the workplace? Mental Health America (electronic resource). Available at: https://mhanational.org/what-emotional-intelligence-and-how-does-it-apply-workplace (access date 2024-03-24)

[8] Wolf, Julia. (2022). Implications of pretend play for Theory of Mind research. Synthese. 200. 10.1007/s11229-022-03984-5.

[9] Aydin, Utkun & Ozgeldi, Meric. (2019). Unpacking the Roles of Metacognition and Theory of Mind in Turkish Undergraduate Students' Academic Achievement: A Test Of Two Mediation Models. 21. 1333-1365. 10.15516/cje.v21i4.3303.

[10] Smogorzewska, Joanna & Grzegorz, Szumski & Bosacki, Sandra & Grygiel, Pawel & Karwowski, Maciej. (2022). School Engagement, Sensitivity to Criticism and Academic Achievement in Children: The Predictive Role of Theory of Mind Short title: Cognitive consequences of ToM development. Learning and Individual Differences. 93. 10.1016/j.lindif.2021.102111.

[11] Alloway TP, Alloway RG. Investigating the predictive roles of working memory and IQ in academic attainment. J Exp Child Psychol. 2010 May;106(1):20-9. doi: 10.1016/j.jecp.2009.11.003. Epub 2009 Dec 16. PMID: 20018296.

[12] Luna B., Tervo-Clemmens B., Calabro F. J. Considerations when characterizing adolescent neurocognitive development. Biological psychiatry, 2021, Vol. 89, №2, pp. 96-98.

[13] Kail R. V., Ferrer E. Processing speed in childhood and adolescence: Longitudinal models for examining developmental change. Child development, 2007, Vol. 7, №6, pp. 1760-1770.

[14] Susan M Sawyer et al. The age of adolescence. Prof, Susan M Sawyer MD, Peter S Azzopardi, PhD, Dakshitha Wickremarathne, MDS, Prof George C Patton, MD. The Lancet Child & Adolescent Health. Vol. 2, Iss. 3, pp. 223-228, March 2018. DOI: https://doi.org/10.1016/S2352-4642(18)30022-1

[15] Tervo-Clemmens, B., Calabro, F. J., Parr, A. C. et al. A canonical trajectory of executive function maturation from adolescence to adulthood. Nat Commun 14, 6922 (2023). https://doi.org/10.1038/s41467-023-42540-8

[16] GOST R 55062-2021 Natsionalny standart Rossiyskoy Federatsii "Informatsionnye tekhnologii. Interoperabelnost" OCS 35.240.50. Date of introduction 2022-04-30

[17] Sravnitelnye harakteristiki neyrosetey i ih primenenie v obrazovanii [Comparative characteristics of neural networks and their application in education]. Available at: https://habr.com/ru/articles/778044/. Published 2023-12-04. (access date 2024-03-04)

[18] Erik Ofgang «Gemini: Teaching With Google's Latest AI». https://www.techlearning.com/news/gemini-teaching-with-googles-latest-ai. Published 20 December 2023 (access date 04-03-2024)

[19] TTehnologija kompleksnoj podderzhki zhiznennogo cikla semanticheski sovmestimyh intellektual'nyh komp'juternyh sistem novogo pokolenija [Technology of complex life cycle support of semantically compatible intelligent computer systems of new generation]. Minsk, Bestprint, 2023, P. 1064 (in Russian)

[20] Kozlova E. I. Grakova N. V., Golovaty A. I. Automation of educational activities within the OSTIS Ecosystem. Open semantic technologies for intelligent systems: scientific papers collection, Minsk, BSUIR, 2023, Iss. 7, pp. 257-262.

ИНТЕРОПЕРАБЕЛЬНОСТЬ КАК ВАЖНЕЙШИЙ КОМПОНЕНТ ИНТЕЛЛЕКТУАЛЬНОЙ ОБРАЗОВАТЕЛЬНОЙ СРЕДЫ В СРЕДНЕЙ ШКОЛЕ

Козлова Е. И., Головатый А. И.

В работе представлены некоторые результаты анализа роли развития интероперабельности, когнитивных способностей и эмоционального интеллекта у детей в современной школе. Обсуждается важность и способы внедрения технологических средств с возможностями взаимодействия и обмена данными для оптимизации образовательного процесса. Также рассматривается значимость развития когнитивных способностей и эмоционального интеллекта учащихся и влияние этого на их академические достижения и социальную адаптацию.

Received 24.03.2024

OSTIS Glossary — the Tool to Ensure Consistent and Compatible Activity for the Development of the New Generation Intelligent Systems

Nikita Zotov, Tikhon Khodosov, Mikhail Ostrov, Anna Poznyak, Ivan Romanchuk,
Kate Rublevskaya, Bogdan Semchenko, Daria Sergievich, Artsiom Titov, Filip Sharou
Belarusian State University of
Informatics and Radioelectronics
Minsk, Belarus
Email: n.zotov@bsuir.by

***Abstract*—This paper includes a detailed analysis of the problems of organising various types of collective activities, a comparative analysis of current solutions to ensure the consistency and compatibility of information from different knowledge areas, as well as an analysis of methods and technologies for creating unified information spaces to ensure consistent and compatible storage, processing, accumulation and dissemination of knowledge. The paper proposes one of the options for realising a unified information resource to ensure consistent and compatible activities in the development of new generation intelligent computer systems — the *OSTIS Glossary*. It describes its structure, rules of structuring, placement and identification of knowledge in it, as well as its operating principles.**

Keywords*—problem of mutual understanding, knowledge unification, knowledge convergence and integration, knowledge consistency, semantic knowledge compatibility, knowledge standardisation, interdisciplinary synthesis, consistency and compatibility of activities, Artificial Intelligence, semantic knowledge representation, semantic web, knowledge base, intelligent system, scientific knowledge portals, *OSTIS Glossary*, *OSTIS Standard

I. Introduction

In the era of information society, the problem of mutual understanding between people is becoming more and more acute. Due to the existing inconsistency in the definition of terms of concepts, people not only do not understand each other, but (!) also do not have the necessary means to communicate with computer systems and create collectives of computer systems that understand each other [1], [2].

First of all, the reason for this problem lies in the form of information representation and the means by which this information is presented, accumulated, processed, distributed and visualised [1], [3], [4]. As the amount of information increases, not only the number of different forms of representing this information [5] grows, but also the number of different tools and methods to support them.

In addition, due to the rapid development of information technology and the emergence of new fields, knowledge can quickly become outdated or may not correspond to reality and current needs. Therefore, it is important to constantly update and clarify terminology, establish common standards and rules for the use of terms to reduce the likelihood of inconsistency of concepts.

At the current stage of information technology development, the problem of information inconsistency is solved with the help of integrated repositories in the form of reference books, encyclopaedias, standards and online resources. However, even with the use of these tools, the problem of conceptual inconsistency remains relevant for several [1] reasons:

- In different fields of knowledge, terms of concepts often have different meanings, leading to misunderstandings between the people using these terms [5].
- The understanding of the terms of concepts may vary from person to person depending on their experience, education and culture.
- With the development of the internet and digital technology, information has become more accessible. Among the variety of information available, it is often difficult to understand which term is used in which context, what its exact meaning is, and to which field it belongs.
- There are many sources of information, some of which may be inaccurate, distorted or misinforming [6].
- Modern natural languages are constantly evolving, the terms of concepts in them are rapidly changing their meanings or acquiring new ones depending on context and usage, which complicates the task of creating information resources, such as encyclopaedias or reference books, which could fully reflect the unified meaning of terms [1].

- There is a problem of translating terms of concepts from one language to another, which also leads to misunderstanding and misuse of these terms.
- Concepts are described and represented in different ways, which makes it difficult for both humans and computer systems to use them in problem solving [7], [3].

Most often the main problem is not the concepts themselves, but (!) the terms of the concepts with which we name these concepts. All the above problems are related not so much to the current capabilities of the technology as to the current state of modern information technologies and the means realised by them. To solve these problems, it is necessary to switch to methods and technologies of a new level [3], namely:

- to develop standards for the unified interpretation of concepts and their terms in different fields of knowledge;
- to create and implement accessible, unified, integrated information resources that are continuously updated and adaptable to changing meanings and contexts of the terms of concepts;
- to develop the conditions and capabilities to build collectives of people and computer systems and to enable them to enhance their understanding, agreement and coordination capabilities;
- to create and implement human-centred systems that reduce the requirements and improve the conditions for their adaptation and knowledge acquisition.

In simple words, to solve these problems it is necessary to comprehensively standardise the existing terminologies [3], [4] used in various fields of knowledge and to create a unified information space for quick access to information in it, as well as the possibility of using and processing this information not only by a human but also by computer systems [4], [8].

The objective of this paper is to address these issues by creating a single ***integrated glossary*** that can be used to:

- to provide a single source for the interpretation of concepts from different fields of knowledge;
- to provide a single source for obtaining relevant and reliable information;
- to integrate knowledge from different information sources;
- to systematise concepts and the links between them in the form of hierarchies of these concepts;
- to unify the representation form of concepts and their terms, i.e. to describe concepts in the same language understandable both to a human and a computer system;
- to standardise the descriptions of these concepts and their terms by introduction and consistent use of common rules for their identification, specification and placement;
- to provide collective consistency and supplement of existing concepts and introduction of new ones;
- to provide open access to all knowledge and facilitate its exchange;
- and, finally, to develop semantically compatible intelligent systems of various kinds.

Such an integrated glossary needs to be developed as an intelligent system, which will provide:

- usability of this glossary, for example, will allow to explain the difference between terms, to give advice, to automate (!) the search for synonyms and homonyms on the basis of some secondary features, and so on;
- convenience of this glossary development, which will allow not only to collectively develop such a glossary, but (!) also will help the system to improve itself.

In the next section we will consider in detail the main problems associated with the organisation of various types of collective activity and the problems associated with the representation, processing, accumulation and transmission of information, since these are the problems that are the object of study in this paper.

II. Analysis of modern solutions to the problems of consistency and compatibility of different types of activities

A. History of the development of the problem of consistency and compatibility of different activities

To understand modern problems of consistency and compatibility of different types of activities it is necessary to consider the problems associated with the presentation, processing and application of different kinds of information, namely, it is necessary to study the history of the development of society in the direction of improving the ways of interaction between people, organisation of their activities and ways of transferring and accumulating knowledge from the older generation of people to the new generation.

As far back as in ancient times, people began to face the problem of understanding each other. First of all, the need to understand each other was the need for self-preservation (survival) in the environment of similar people. The only way of communication between people for a long time remained communication by means of gestures, pictures and other similar means. For a long time people exchanged knowledge with the help of these means. And not all gestures came only from the man himself. Man also perceived the gestures of nature and made strategically important decisions based on them.

With the development of mental capabilities and the increasing needs of each individual, people began to look for more flexible forms to realise their self-preservation (survival) in nature. People began to realise that it is possible to negotiate with each other to achieve a certain goal through mutual benefit. People co-operated and coordinated their actions on the basis of common interests, traditions and customs, which helped to coordinate actions and establish rules of interaction between people. The first languages appeared, in which people could express their thoughts and transmit these thoughts to other people. People began to organise themselves into groups, thanks to which they became stronger among other groups of people and nature in general. In each group, so-called leaders were formed, usually distinguished among other people by their physical and mental abilities. Thus a form of people management appeared — power, with the help of which one person could organise the work of other people. This form of people management has evolved a lot since then.

Today, people manage not just people, but the environment in which they exist: resources, relationships, knowledge, etc. Whereas in ancient times the question of survival was of primary importance to humans, today the main question is how to properly accumulate, organise, use and transfer knowledge from one person to another and from one generation to another generation of people. This question is also related to the question of human survival. If we, people, do not or cannot develop, i.e. accumulate and multiply the knowledge that we have, how will we be able to solve those constantly emerging problems that any person or society faces, how will we be able to build a future in which every person will be happy and will be able to get what he or she wants. Even today, questions about the future of humanity remain a priority and open for discussion.

B. Organisation and consistency of collective activities through standardisation and legal regulations

For a long time of human and social existence, people have invented and realised quite a large number of forms by means of which they can communicate with each other, achieve goals by reaching common understanding and agreement, solve problems, relying on common sources of knowledge that have passed through time. The issue of organising common activities between people remains the most important. It is the coordinated collective activity of people that contributes to the rapid flow of the creative process, which allows to solve problems faster.

Compared to primitive society, the level of organisation is much higher nowadays. Nowadays, people use a great variety of methods and tools to organise and coordinate their activities:

- organisational structures are established within businesses, organisations and public institutions that define the hierarchy, roles and functions of employees;
- information technologies such as e-mail, messengers, video conferencing, project management systems and others are used to exchange information, coordinate actions between people;
- all rules of behaviour and processes within communities are standardised to improve efficiency and ensure the quality of their products or services.

It can be seen that the degree of the consistency of actions is also much higher compared to earlier stages of human development.

Regulations and standards were the first to address the problem of collecting and systematising knowledge, as they represent formalised norms and rules that regulate the behaviour of people and organisations in society. The introduction of regulations and standards helps to create a unified and ordered approach to knowledge organisation and provides a common basis for information exchange.

Regulations and standards enable:

- to ensure unity, accuracy and reliability of information;
- to organise knowledge, i.e. classify and structure information;
- to protect people's rights and interests.

Regulations and standards are documents that establish rules, norms and requirements in a particular area (e.g. product quality standards, building codes, safety regulations, etc.). They are aimed at ensuring uniformity, quality and safety in different areas of activity. The task of any standard in general is to describe a consistent system of concepts (and corresponding terms), business processes, rules and other regularities, ways of solving certain classes of tasks etc.

However, they cannot completely solve all existing problems because:

- Knowledge is constantly evolving and changing, making regulations and standards quickly outdated or unable to adequately reflect new knowledge.
- The process of adding new knowledge is too resource-intensive because adding new knowledge requires searching for similar existing knowledge and manually integrating that knowledge with existing knowledge.
- Modern normative legal acts and standards describe only rather narrowly specialised knowledge, when the rest of knowledge is not standardised in any way. That is, the question of interdisciplinary organisation of knowledge remains open.
- The most urgent problem remains the problem related not to the form, but to the essence (semantics) of standards - the problem of inconsistency of systems of concepts and terms between different standards, which is relevant even for standards within the same field of activity.

C. Use of common encyclopaedias in organising and coordinating collective activities

In addition to regulations and standards, there are encyclopaedias [9] and dictionaries that cover knowledge from different subject domains and are interdisciplinary

[10]. Encyclopaedias and dictionaries are designed to integrate information related to a common topic.

Encyclopaedias are reference publications that contain information on a wide range of topics and subjects. They are intended for general familiarisation with different fields of knowledge and may contain general information, historical facts, descriptions of phenomena, etc. The purpose of any encyclopaedia is to collect knowledge scattered across disciplines and bring it into a system understandable to the individual.

While regulations and standards cannot always provide complete information or cover all aspects of a particular topic, encyclopaedias can be useful to provide a broader context or additional information on a given topic. Today, regulations, standards and encyclopaedias complement each other, but are also used for different purposes.

D. Digitalisation of information for simplifying and accelerating the organisation and consistency of collective action

With the development of technology, it has become much easier to describe and popularise knowledge by creating topic-based websites. One of the representatives of this kind of knowledge repository is mathprofi — a site that is a resource describing all topics of school and higher mathematics. There is a large number of such online resources on any topic and subject. In them, information is presented and described in an understandable and accessible form for any untrained reader.

In parallel with the development of sites like mathprofi, there occurred an idea of creating a common repository that brings together all the knowledge of mankind, categorised by topic, and to which any person can add new information.

The most widely used computer encyclopaedia at present is Wikipedia — a publicly available multilingual universal Internet encyclopaedia with free content. Its main advantages are its multilingualism and the possibility for users to add and adjust its content.

Traditional wikis based on this approach have a number of disadvantages, including a lack of content consistency, that is, the lack of uniformity in the presentation and formalisation of this content. In wikis, due to frequent duplication of data, the same information may be contained on several different pages. When this information is changed on one wiki page, users must ensure that the data is also updated on all other pages.

Another disadvantage is the difficulty of accessing the knowledge available on wikis. Large wikis contain thousands of pages. Performing complex search queries and comparing information from different pages in traditional wiki systems is a time-consuming task. Traditional wikis use flat classification systems (tags), or classifiers organised into taxonomies. The inability to use typified properties generates a huge number of tags or categories [9].

Open databases and knowledge bases are promising tools for building and retrieving knowledge. Semantic Web-based open databases and knowledge bases [11] are resources that use Semantic Web principles and technologies to represent and organise data and knowledge in a structured and semantic form [12]. The Semantic Web is an extension of the World Wide Web in which information has semantic meaning that allows computers to understand and process data efficiently and accurately. Examples of such Semantic Web-based databases and knowledge bases are:

- DBpedia — one of the best known Semantic Web projects. It extracts structured data from Wikipedia and presents it in RDF format. DBpedia contains an extensive set of knowledge, including information about people, places, organisations, scientific articles and more.
- Linked Open Data (LOD) is an initiative that aggregates and provides access to open data in RDF format from a variety of sources. LOD brings together data from fields such as geography, biology, culture, economics and others, and makes it possible to share and analyse these data.
- Wikidata is an open knowledge base developed by the Wikimedia Foundation. It contains structured data about various entities, including people, places, books, films, scientific terms, and more. Wikidata uses the RDF language to represent the data and provides an API to access this data.
- Cyc — a project to create an ontological knowledge base that allows computer systems to solve complex Artificial Intelligence problems based on logical inference.
- GeoNames — a geographic database that contains information about places from all over the world. It provides data on geographic coordinates, population, administrative units, geographic objects, and other information. GeoNames uses Semantic Web standards such as RDF and OWL to represent and organise data.
- MusicBrainz — an open source music database that contains information about music artists, albums, tracks and other music-related entities. It uses Semantic Web technologies to organise and represent data, and provides an API to access this data.

The listed databases and knowledge bases also do not solve all the problems that modern Wikipedia has.

Unfortunately, modern traditions of presentation of various kinds of documentation, standards, reports, scientific and technical articles and monographs are not only not oriented to their adequate understanding by intelligent computer systems, but also do not contribute to their quick understanding by those people to whom these texts are addressed. The latter circumstance requires the development (writing) of textbooks and teaching aids specially designed for those people who are beginning to master the relevant field of knowledge, who have not yet acquired the necessary qualifications. But it is obvious that this implies a significant duplication of the information presented.

E. Shortcomings of current solutions to ensure consistency and compatibility of different activities

All the tools considered for representing, structuring and accumulating information make it possible to simplify the organisation and coordination of collective activities, but do not allow to solve these tasks in a comprehensive way, because:

- The increase in the number of reference materials presenting and describing the same information in different forms leads to an increase in duplication and, consequently, inconsistency of this information.
- There is quite a lot of information in existing information resources that is characterised by inaccuracy, unstructured, incomplete, incoherent and unreliable information.
- Information becomes obsolete rather quickly, i.e. becomes irrelevant and unclaimed due to finding new methods of solving existing problems. All irrelevant information is quickly accumulated in the Internet space. That is why there is a lot of so-called "junk information" in Internet resources, which is this irrelevant and unclaimed information.
- The input of information in information resources is done by intermediaries - people who do not have the necessary competence to modernise and disseminate this information, which directly affects the quality of all information. This is also due to the fact that a person who receives information from one source interprets and transmits it to another source in his or her own way. Different people describe concepts from different sources by synonymous terms, which leads to the loss of the original meaning of these concepts. Thus, new contradictions in information appear.
- Working with huge amounts of information implies working with several sources of this information. In such sources it is difficult to search for necessary (relevant) information, as there is a huge number of different categories, which implies the use of complex search operations.
- This, in turn, is related to the language of knowledge representation. The format of knowledge representation and description in reference materials is understandable only to a human being and cannot be processed by a computer system, and as a consequence, cannot be used for solving problems by a computer.
- Knowledge is most often structured in the form of books, encyclopaedias, dictionaries and reference books on specific subject domains, which allows one to learn a particular subject domain quickly. However, this makes it difficult to understand information at the "junctions" of subject domains, so that a person is not well-versed in interdisciplinary knowledge. The so-called "mosaicism" of perception is formed in a human being, as a human being during training and work gets used to artificial division of knowledge areas and has difficulties in solving problems at "junctions".
- To integrate information from different sources, algorithms for matching and merging data, identifying and resolving duplicates, and algorithms for converting to common presentation formats are used, but even these algorithms do not completely eliminate inconsistencies and duplication of existing information.
- The existing information resources do not standardise and do not apply general principles of presenting information for a wide range of readers. Each reader perceives information in his/her own way, and consequently, there are differences in the understanding of the same information.
- The popularisation of knowledge is carried out with the help of specialised Internet resources that not only simplify but also distort the presentation of information for professionally untrained readers. The increase in the number of such Internet resources contributes to the duplication of information and the development of contradictions in it.

To solve these problems, methods and technologies must be utilised which can:

- present any information in the <u>same</u> form;
- <u>integrate</u> information from different information sources;
- describe and <u>structure</u> information both from one subject domain and information at "junctions" between subject domains;
- <u>standardise</u> the description and visualisation of various types of information;
- <u>re-use</u> existing knowledge and accumulate new knowledge;
- present information in a form that is <u>understandable</u> to both (!) humans and computers;
- develop tools to <u>quickly find</u> the information you need;
- create a <u>personalised</u> experience for any user;
- <u>develop</u> methods and tools to improve these methods and technologies.

In other words, it is necessary to create such unified integrated information resources, with the help of which it is possible to quickly obtain existing information and to integrate new information and it would be easy to coordinate various activities, including activities on the development of intelligent systems. It is also necessary to develop methods and tools by which such information resources can be continually modernised and updated.

F. Methods and technologies for solving the problem of collective activity organisation and consistency

Today, some information is combined, structured and used in the form of common information repositories — Internet resources with their own databases or knowledge bases [13]. Databases and knowledge bases allow storing and processing information in the same place and automating the solution of information tasks of various kinds. That is, unlike non-digital analogues, information in databases or knowledge bases can be interpreted not only by a human but also by a computer.

This approach is the most promising in comparison with other existing approaches, as it allows solving some of the above mentioned problems. However, it is necessary to approach the solution of these problems in a methodologically correct way, in other words, to use, integrate and develop methods and technologies with the help of which it is possible to develop tools that allow solving all the problems listed above.

In order to create unified information resources with the help of which specific activities could be easily made consistent, it is necessary to apply modern technologies and methods that enable effective data and knowledge management [14].

Knowledge-Based Engineering.

A popular area of research is ***knowledge-based engineering***. Knowledge-based engineering (KBE) [15] is a research area that explores methodologies and technologies for capturing and reusing knowledge in product development. The goal of KBE is to reduce product development time and costs, which is primarily achieved by automating repetitive design tasks while getting, storing, and reusing knowledge about already designed products.

One of the hallmarks of the KBE approach is the automation of repetitive, non-creative design tasks. Automation not only provides significant time and cost savings, but also frees up time for creativity, allowing more of the design domain to be explored. This is facilitated by another advantage of KBE: it allows knowledge to be reused. As the researchers note, "about 20% of a designer's time is spent searching for and assimilating information." This means that development information and knowledge is not represented in a common and easily accessible knowledge base. Obviously, in such cases, reusing knowledge according to the established KBE framework can save considerable time and effort.

The authors of the paper [15] highlight the challenges of using knowledge-based engineering and, first and foremost, emphasise the need to bridge the "technology gap" — lack of tools and technologies to support cost-effective KBE development and its application in the development of information systems for various purposes. According to the authors of the paper, such tools and technologies should provide search and reuse of knowledge, allow standardisation of this knowledge, describe the meaning of information and ensure transparency of information systems developed according to KBE principles.

Ontology-based approach.

Among the methodologies and technology of knowledge engineering, the description and structuring of knowledge by means of ontologies is most often given more attention [16]. The **ontology-based approach** is a methodology based on the use of ontologies to organise and represent knowledge about a particular subject domain [17]. In this context, "ontology" refers to the formal description of concepts, relationships, and attributes in a subject domain.

The widespread use of the ontological approach is explained by the fact that [18], [19], [20]:

- Ontologies make it possible to provide consistency of concept systems between participants in some process.
- Ontologies allow to organise and classify knowledge from different disciplines, establishing links between them, thus making knowledge more structured and usable [21].
- Ontologies allow knowledge integration by combining information from different sources and with different representations, which helps to eliminate semantic incompatibilities between different systems and facilitates information sharing [22].
- Ontologies allow knowledge to be represented in a form that computers can understand. Thus systems can automatically analyse and make decisions based on the knowledge in ontologies.
- Ontologies provide a formal basis for developing expert systems that can provide recommendations and advice in complex and multidimensional problems [23].

The use of ontologies for interdisciplinary knowledge synthesis and integration helps in gathering and analysing information from different domains and ensuring that it is properly understood and used in practice, thus contributing to better knowledge utilisation and more informed decision making.

A series of works [24] emphasises the importance of using ontologies to structure engineering knowledge. According to the authors, ontologies reveal the semantics of the information presented, eliminate heterogeneity in the representation of multiple information sources, provide a common knowledge base for multi-agent systems, provide semantics and structure for trust and reputation systems, privacy-based systems, and codify common knowledge across business and scientific domains. The authors believe that the use of semantics as a central mechanism will revolutionise the development and consumption of software and lead to the development of software as a service — Software engineering 2.0.

According to the authors, an important challenge remains the problem of ensuring that knowledge bases are consistent and coherent with each other, one solution to which may be to develop a hierarchy of ontologies for all systems in the form of a common knowledge base. Knowledge bases help to simplify the storage of different types of knowledge [23], and the ontology approach helps to structure knowledge and the relationships between them [19]. This provides a deeper understanding of the information context and improves data availability and integrity.

The authors also highlight the importance of multi-agent systems in ontology processing. A set of agents in a multi-agent system can use this ontology as a common knowledge base. This will greatly facilitate communication and coordination between agents when solving tasks together. The [24] describes the problem that methodologies for developing ontologies and methodologies for developing multi-agent systems are completely separate and have no connection with each other. The authors believe that combining multi-agent systems and ontologies for mutual use could revolutionise information technology.

Knowledge graph.

Another popular method for describing complex structured knowledge and knowledge from different subject domains is the ***knowledge graph*** [25], [26]. A knowledge graph is a semantic network that stores information about different entities and the relationships between them. An entity or "node" of a graph can be anything: any material object or abstract concept. Predicates or "edges" reflect the relationships between different entities in the graph. For example, Albert Einstein and the city of Ulm in Germany are two separate entities, and the fact that Einstein was born in Ulm is a predicate.

The use of graph models has the following advantages [12], [27], [11], [13]:

- Graph data models have tremendous expressive power. Graph databases offer a <u>flexible</u> model of data and a way to represent it. Graphs are additive, providing the flexibility to add new data relationships, new nodes and new subgraphs to an existing graph structure without compromising its integrity and coherence.
- The diversity of data representation is minimised by reducing the number of syntactic aspects, as graph data models allow different types of knowledge to be written in the <u>same</u> way.
- To understand the meaning of knowledge, it is necessary to represent this knowledge in an <u>understandable</u> form for everyone: both for a person and for the system. Speaking about unification of representation of all kinds of knowledge, it is considered important to use graph models not just as means for storing structured data, but for storing semantically coherent and interconnected knowledge.
- <u>Performance</u> of data processing is improved by one or more orders of magnitude when representing data in the form of graphs, which is explained by the properties of the graph itself. Unlike relational databases, where query performance degrades with increasing query intensity as the dataset grows, the performance of the graph data model remains constant even as the dataset grows. This is because data processing is localised in some part of the graph. As a result, the execution time of each query is only proportional to the size of the part of the graph traversed to satisfy that query, not the size of the entire graph.
- Graph models enable <u>efficient</u> semantic <u>search</u>, i.e. finding data and information based on the relationships between them, which helps to improve the quality and accuracy of search queries, as well as provides a deeper understanding of these relationships and dependencies between data.

Knowledge graphs, like ontologies, help to link large amounts of data from different sources into one common knowledge collection. They can be general, i.e. storing information about different types of data, and specialised, focusing on a single subject domain.

Today, one of the largest knowledge graphs that stores information from different subject domains is Wikidata. Another good example of a knowledge graph is BioPortal – the largest specialised graph with over 140 billion facts about biotechnology and medicine. These graphs are publicly available and accessible to all Internet users.

Systems that use knowledge graphs to represent and process information are widespread. Google Knowledge Graph — is a system used by the Google search engine to provide structured data about user queries. Google Knowledge Graph combines data from various sources such as Wikipedia, Freebase and others to provide information about objects in various fields such as history, culture, science and technology.

Collaborative design.

Obviously, to ensure consistent and compatible activities, only technologies and tools for representing, structuring, accumulating different kinds of knowledge used in these activities are not enough. In addition, it is necessary to create conditions and means to improve the activity itself and its integration with other activities to solve more problems.

One of such approaches that promotes the convergence of activities of specialists from different fields is the methodology of ***collaborative design***. Collaborative design implies collective decision-making, open communication and active participation of all stakeholders throughout the design process. By utilizing the power of collaboration, this approach aims to create innovative and effective solutions that meet the needs and desires of end users.

IEEE researchers in their paper [28] examine the problems of knowledge management and sharing, which are of considerable interest in the field of organisational management. The authors note that much of the knowledge within organisations exists in the minds of individuals and is difficult to document or transfer.

American researchers Stephanie E. Hampton and John N. Parker in their paper [29] consider the importance of co-operation between scientists in the process of research and solving scientific problems. They use and describe the concept of scientific synthesis. Scientific synthesis is the process of combining knowledge from different fields to create new knowledge and develop innovative solutions. With the increasing amount of information and complexity of scientific problems, the role of collaboration is becoming more and more relevant for successful scientific synthesis. The paper also emphasises that collaboration involves not only sharing knowledge from different fields, but also joint research, discussion, synthesis, analysis and interpretation of results. Highly organised collaboration improves the quality of scientific synthesis by providing a broader overview of the problem [30].

The authors offer a number of recommendations for successful collaboration and increased productivity. One of them is to create resources and platforms for sharing knowledge and ideas that are accessible to all scientific researchers. This would reduce the information formats used and facilitate the sharing of information and discussion of results [31].

One way of creating centralised access to information is through ***knowledge portals***. Such portals can contain information related to processes, documentation, procedures, tutorials, and answers to frequently asked questions [32], [33].

One of the key advantages of knowledge portals is their ability to collect and store information from various sources, such as databases, document management systems, project management systems and so on, allowing users to obtain complete and up-to-date information in one place. Based on the knowledge portal, the ability for users to interact with each other by creating forums, discussions and collective editing of documents is provided. This facilitates the sharing of knowledge and experience among the staff of the organisation and enhances their work efficiency.

G. Conclusions

If we consider society as a multi-agent system consisting of independent intellectual agents, it is obvious that the important factors determining the improvement of the quality (level of development) of society are:

- increasing the efficiency of utilisation of the experience accumulated by society, the efficiency of mankind's use of knowledge and skills;
- increasing of the rate of acquisition, accumulation and systematization of knowledge and skills effectively used by mankind.

The solution of the above problems becomes quite possible if for this purpose intelligent computer systems of new generation are used, with the help of which the knowledge and skills accumulated by mankind will be organised as a systematized distributed library of reusable information resources (knowledge and skills).

Consequently, systematisation and automation of reusable information resources accumulated by mankind requires their convergence, deep integration and formalisation. A special place in this process is occupied by mathematics as a basis for systematisation and formalisation of knowledge and skills at the level of formal ontologies.

III. Proposed approach. Semantic electronic glossary based on the *OSTIS Technology* — the *OSTIS Glossary*

A. About the relation between the OSTIS Standard and the OSTIS Glossary. Basic requirements for the OSTIS Glossary

In the analytical part of this paper, the current state of the solution of the problem of ensuring consistency and interoperability of different types of activities was demonstrated, shortcomings and ways to solve them were highlighted. The authors of this paper do not deny all the accumulated experience in solving this problem, but they see significant shortcomings of existing technologies and means by which knowledge is presented, processed, accumulated and disseminated, and which significantly slows down the pace of development of various areas of activity in society.

The intention of the authors and the purpose of this paper is to create a unified information resource that allows:

- to present information of various kinds in a comprehensible and accessible form both for a human being, including untrained professionals, and for a computer;
- to reuse existing information to solve problems of any information complexity;
- to accumulate information from various sources,

that is:

- to contribute to the consistency of different points of view and different activities of society;
- and to promote integration and convergence (interoperability) of one activity between its parts ("horizontally") and between different activities ("vertically").

The following methods and technologies are also considered appropriate for the creation of such a unified information resource:

- As a basis for structuring information of different types and different subject domains, as well as

for ensuring the consistency and interoperability of this information and different types of activities that use this information, it is proposed to use an ontological approach based on the representation of information in the form of semantic networks (knowledge graphs);
- An agent-based approach is proposed as tools for retrieving already existing information, reusing it for problem solving and accumulating it from various information sources;
- Semantic user interfaces using ontological, graphodynamic and agent-based approaches are proposed as means to realise personalised representation and transfer of information.

Such an information resource can be realised in the form of an ***Electronic glossary*** for all available spheres of human activity in society. In the traditional sense, such a glossary can be a dictionary of highly specialised concepts and their terms in various subject domains, the text of which will be clear to the end user. In a broader sense, such a glossary will be understood as a system or subsystem with the help of which:

- you can quickly find already existing information: concepts, their terms, definitions, connections with other concepts, and so on;
- reuse information by both humans and computer systems that know how to communicate with this glossary;
- accumulate information, i.e. new information from various sources can be entered both manually and automatically;
- information can be visualised depending on the user's learning level.

The ***OSTIS Technology*** is proposed to be used as a technology that allows to realise such systems and has all the necessary methods and means for their implementation. The *OSTIS Technology* is an open semantic technology of complex life cycle support of semantically interoperable intelligent computer systems of new generation. The purposefulness of using this technology is determined by the tasks that can be solved with the help of this technology, and the high level of scientific problem, which is aimed at solving this technology.

The *OSTIS Technology* is known for its basic principles [34], [35], [36]:

- all knowledge is described by means of a unified knowledge representation language — *SC-code*, which ensures syntactic and semantic interoperability of this knowledge and makes it possible to interpret knowledge not only by humans but also by computer systems;
- All knowledge is structured by means of a hierarchy of subject domains and their corresponding ontologies, through which the consistency of this knowledge is ensured;
- knowledge processing of various kinds is based on the principles of graphodynamic models, with the help of which it is possible to understand the meaning of this knowledge efficiently and flexibly;
- all knowledge is accumulated in the form of semantically powerful libraries of reusable components, with the help of which the reuse of already existing knowledge in problem solving is realised;
- all knowledge is open and transparent in use and modification by both humans and systems.

SC-code is the main internal formal universal abstract language for representing information constructs in ostis-systems. *SC-code* supports various data types, including numbers, strings, lists and other data structures. It also provides facilities for working with knowledge bases and performing logical operations. There are 3 main external languages of ostis-systems: SCs-code, SCn-code and SCg-code [37]. They provide a way for ostis-systems to communicate with their users and other ostis-systems.

This technology is considered to be a new generation technology. And in the context of the *OSTIS Technology*, all computer systems, in particular intelligent systems, developed on the basis of this technology are called new generation intelligent systems or ostis-systems. The *OSTIS Technology* is realised in the form of a special ostis-system, which is called the *OSTIS Metasystem* [38] and the knowledge base, which this system contains:

- The formal theory of ostis-systems;
- The standard of ostis-systems (standard of ostis-systems knowledge bases, ostis-systems problem solvers, ostis-systems interfaces);
- The standard of methods and tools for ostis-systems life cycle support (the core of the Library of reusable ostis-system components (the OSTIS Library), methods for supporting the life cycle of ostis-systems and their components, tools for supporting the life cycle of ostis-systems).

The *OSTIS Standard* defines general principles and rules for the development and use of knowledge bases and intelligent systems based on the *OSTIS Technology*. It defines the structure of knowledge bases, data formats, ways of knowledge organisation and other aspects that should be taken into account when developing intelligent systems based on the *OSTIS Technology*. In addition, the *OSTIS Standard* defines the rules of organisation and representation of knowledge in knowledge bases. The *OSTIS Standard* helps to ensure interoperability and uniformity within the projects using the *OSTIS Technology* and provides convenience for developers and researchers to work together.

In the context of this paper, the *OSTIS Metasystem* [38], [39] attracts attention because it is some computer version of the *OSTIS Standard*, additionally implementing tools for processing, visualising and using this standard, the aim of which is to provide a consistent and compatible activity both for the development of intelligent systems and for the improvement of methods, models and tools for their development, which is a subobjective of the Electronic glossary conceived by this work.

In addition, all ostis-systems are combined to form an integrated ecosystem — the *OSTIS Ecosystem* [40], containing knowledge from all subject domains of society in a consistent, semantically interoperable and understandable form. If we compare the Electronic glossary and the *OSTIS Ecosystem*, it is obvious that the Electronic glossary is a variant of the *OSTIS Ecosystem* knowledge base. However, in the context of this work, it is more appropriate to specify the Electronic glossary to the glossary, which is a variant of displaying the *OSTIS Standard*, because if we consider the Electronic glossary as a variant of displaying the knowledge base of the *OSTIS Ecosystem*, it is necessary in this work to focus in more detail not on the principles of structuring information in this glossary, but on the principles of coordination of information from different subject domains, which are united in the form of a single knowledge base of the *OSTIS Ecosystem* and which is the subject of the *OSTIS Standard*.

Also due to the identified dependence of the Electronic glossary on the *OSTIS Standard* and the *OSTIS Technology* as a whole, we will refer to this Electronic glossary as the ***OSTIS Glossary***. And also it is important to understand that in this case the *OSTIS Glossary* is not some separate ostis-system, but is a form of display of the *OSTIS Standard*, its part, visualised according to certain formalised rules, and by virtue of this it is also a part of the knowledge base of the *OSTIS Metasystem*, to which it is possible to ask various questions.

Therefore, in the following sections, besides the content of the conceived the *OSTIS Glossary*, the rules of its development, visualisation and tools for working with it will be described in detail. That is, in this paper it is important to fulfil the following tasks:

- describe and fix the principles and rules of development and consistency of the *OSTIS Glossary* fragments both with itself and with the *OSTIS Standard*;
- describe and fix the rules of placement, structuring, identification of concepts and fragments of concepts in the *OSTIS Glossary*;
- describe and fix the principles of interaction of users and systems with the *OSTIS Glossary* and the rules of visualisation of concepts and fragments of concepts in it;
- describe the current structure of the *OSTIS Glossary*.

Further on we will consider in detail the principles of consistency of the *OSTIS Glossary* and the *OSTIS Standard*, the rules of structuring and specification of *OSTIS Glossary* objects, as well as the rules of identification of these objects within the *OSTIS Glossary*.

B. Principles of consistency of the OSTIS Glossary and the OSTIS Standard

As stated earlier, the *OSTIS Glossary* is closely related to the *OSTIS Standard*. This is due to the following reasons, which at the same time are the ***basic principles for the development of the OSTIS Glossary***, consistency of its text with itself and the text of the *OSTIS Standard*:

- The *OSTIS Glossary* is not a separate knowledge base or ostis-system. On the contrary, the *OSTIS Glossary* is a semantically compatible and ordered fragment of the *OSTIS Standard*, some variant of its display, which describes with sufficient detail the entities and concepts used in Artificial Intelligence and, in particular, OSTIS Technologies, as well as the relations between them, references to bibliographic sources and authors of these entities and concepts. The *OSTIS Glossary* is the result of a collective consistency of terms both within the *OSTIS Technology* and across Artificial Intelligence.
- In addition to the *OSTIS Glossary*, an important component of the *OSTIS Standard* is the OSTIS Bibliography. The OSTIS Bibliography is a list of the literature used in the *OSTIS Standard*. The OSTIS Bibliography is the result of the analysis of other works and analogues studied during the development of the *OSTIS Technology* and includes brief bibliographic descriptions of both other technologies similar to the *OSTIS Technology* and the technologies, models and tools on which the *OSTIS Technology* itself is based.
- Consequently, the key objects of description in the Glossary may be:
  - concepts (absolute and relative);
  - specific entities that are not concepts, e.g. specific systems, projects, technologies, languages, etc. (the *OSTIS Ecosystem* [40], the *OSTIS Metasystem*, Neo4j Project, RDF Language);
  - specific individuals;
  - specific bibliographic sources (books, articles, electronic resources).
- The *OSTIS Glossary* is not a static structure stored in the knowledge base of the *OSTIS Metasystem*. It is the result of the work of some collective of agents transforming the hierarchy of sections of the *OSTIS Standard* into some simplified from the reader's point of view hierarchy of sections, in which concepts are ordered lexicographically rather than logically, i.e. the *OSTIS Glossary* can be formed by means of explicit or indirect start of a non-atomic agent consisting of agents: forming a particular section of this glossary, concatenation

of glossary sections, filtering of concepts and their specifications according to given criteria, and so on.
- Since the *OSTIS Glossary* is some fragment of the knowledge base of the *OSTIS Metasystem*, it is more appropriate to develop it by the same means that are used to develop any knowledge base of the ostis-system. From this point of view, the development of the *OSTIS Glossary* is reduced to the development of the knowledge base of the *OSTIS Metasystem*, including the *OSTIS Standard*, already loaded in this knowledge base. Thus, manual transformation of the *OSTIS Standard* into the *OSTIS Glossary* is not required and is automated by existing ostis-system knowledge base development tools.
- For visualisation of the *OSTIS Glossary* the existing tools for development of ostis-systems knowledge bases are sufficient. Viewing of the *OSTIS Glossary* from the whole knowledge base of the *OSTIS Metasystem* should be done with the help of a specialised agent. Such an agent should allow to display it in one of the external sc-languages (SCn-code, or SCg-code).

In other words, the *OSTIS Glossary* should not be some other text describing the current state of the *OSTIS Technology*, on the contrary, the Glossary should be consistent and semantically interoperable with the *OSTIS Standard* and should be only some form of its presentation, simplified for the reader, reducing the threshold for new people to enter the *OSTIS Technology*, allowing to quickly find concepts and agree new ones.

C. Rules for structuring and specification of the OSTIS Glossary objects

The main purpose of the *OSTIS Glossary* is to present concise specifications of the *OSTIS Standard* concepts in an organised form, simplified for the reader. The text of the *OSTIS Standard* is presented in the form of a sequence of ordered and organised sections of subject domains and ontologies containing a logical statement of the specifications of the objects considered within the *OSTIS Technology*. The *OSTIS Standard* is based on the ***following principles of text structuring***:

- The *OSTIS Standard* is the main part of the knowledge base of the *OSTIS Metasystem* and is a description of the current state of the *OSTIS Technology*.
- As a formal language for the external representation of the *OSTIS Standard*, SCn-code, which is an external form of *SC-code* representation, is used.
- The *OSTIS Standard* is ontologically structured, i.e. it is a hierarchical system of related formal subject domains and their corresponding formal ontologies, thus ensuring a high level of stratification of the *OSTIS Standard*.
- Each subject domain can be matched:
  - a family of corresponding ontologies of different kinds;
  - a set of semantic neighbourhoods describing the research objects of this subject domain.

  subject domains are the basis for structuring the sense space, a means of localisation, focusing attention on the properties of the most important classes of described entities, which become classes of objects of research in subject domains.
- Each concept used in the *OSTIS Standard* has its own place within this standard, its own subject domain and its corresponding ontology, where this concept is investigated in detail, where all the basic information about this concept and its various properties is concentrated.
- The *OSTIS Standard* also includes files of information constructs that are not *SC-code* constructs (including sc-texts belonging to different natural languages). Such files allow to formally describe in the knowledge base the syntax and semantics of various external languages, and also allow to include in the knowledge base various explanations, notes addressed directly to users and helping them to understand the formal text of the knowledge base.
- From a semantic point of view, the *OSTIS Standard* is a large refined semantic network, which is non-linear in nature and which includes signs of all kinds of described entities (material entities, abstract entities, concepts, relations, structures) and, accordingly, contains links between all these kinds of entities (in particular, links between links, links between structures).
- The *OSTIS Standard* is a hierarchical system of subject domains and their corresponding ontologies specifying these subject domains. Each of the subject domains describes the corresponding classes of research objects with the maximum possible degree of detail defined by a set of relations and parameters defined on the classes of research objects.
- Each section of the *OSTIS Standard* contains the knowledge that is part of the subject domain and ontology that is either fully represented by the specified section or partially represented by the specification of one or more specific objects of study.
- The specification of each subject domain and each section should have a sufficient degree of completeness. At a minimum, the role of each concept used in each subject domain should be specified.

Since the *OSTIS Glossary* is nothing but a part of the *OSTIS Standard*, the structuring principles of the *OSTIS Glossary* are the same as the structuring principles of the *OSTIS Standard*. The exception is that the structuring of the *OSTIS Glossary* should happen automatically, by some agent generating a structure corresponding to this glossary. Hence, it is important to consider the structure to be formed by this agent, i.e. the principles and rules by which this agent should form the structure of the *OSTIS Glossary*.

In this section, the authors of this paper also want to focus only on why it is important and how to structure and stratify information in general. That is, the purpose of this section is to review the principles of structuring and stratification of knowledge in knowledge bases on *SC-code*, because as it was clarified earlier, the *OSTIS Glossary* is part of one common knowledge base of one ostis-system. If necessary, a detailed description of the syntax and denotational semantics of the knowledge representation language can be found in the works of [34], [41], and a detailed description of what knowledge is and how to structure knowledge in knowledge bases can be found in the works of the following authors [42].

So, let us consider in detail the listed structuring principles of the *OSTIS Standard* and describe the structuring principles of the *OSTIS Glossary*.

In the context of the *OSTIS Standard*, an object is usually defined as either a concept, i.e. an abstract entity that combines other abstract or concrete entities, or an instance of a concept, i.e. a concrete entity. Concepts can be absolute or relative. Absolute concepts denote the same attributes of some group of concepts or entities, relative concepts — connections and relations between other concepts or entities. Generally speaking, there are quite a lot of types of knowledge that can be represented on the *SC-code*. Absolute and relative concepts are basic characteristics of other concepts.

The specification of an object is commonly used to denote a set of information describing this object. Depending on the quality of the knowledge base, primarily determined by the quality of its development, concepts and entities can be specified or unspecified.

Suitably specified entities have the following requirements:

- if the entity is not a concept, the following must be specified for it:
  - different variants of the external signs denoting it;
  - the classes to which it belongs;
  - the links by which it is connected to other entities (indicating the relevant relationship);
  - values of the parameters it possesses;
  - those sections of the knowledge base in which the specified entity is key;
  - the subject domains in which the entity is included;
- if the specified entity is a concept, the following must be specified for it:
  - different variants of external labelling of this concept;
  - The subject domains in which this concept is explored;
  - definition of the concept;
  - explanation;
  - sections of the knowledge base in which this concept is key;
  - example description - an example of an instance of a concept.

These requirements can be formulated in another way. For each object in the knowledge base it is possible to fix their specifications by means of structures denoting sets of all relations of these objects with other objects in the knowledge base. For these structures it is possible to introduce classification, i.e. to set classes for these structures, with the help of which it is possible to understand the degree of detail of specification of a particular object of the knowledge base.

Within the framework of the *OSTIS Technology* for each entity it is done so, such structures describing the specification of entities in the knowledge base are commonly called semantic neighbourhood. Semantic neighbourhood is a specification of a given entity, the sign of which is specified as a key element of this specification.

The set of features by which entities can be specified varies. In addition, it may be necessary to specify the same entity in different aspects and to explicitly capture these aspects in the knowledge base (Fig. 1).

For example, the same person can be described from professional, medical, civil and other perspectives, as presented in the figure.

Consider a specific example of the concept of "cybernetic system" and its specification from the *OSTIS Standard*.

cybernetic system
:= [adaptive system]
:= [targeted system]
:= [active subject of independent activity]
:= [a material entity capable of purposefully (in its own interests) influencing its environment as a minimum to preserve its integrity, viability, and safety]
:= [a natural or artificially created system capable of monitoring and analysing its own state and the state of the environment, as well as capable of sufficiently active influence on its own state and the state of the environment]
:= [a system capable of interacting with its environment sufficiently independently to perform various tasks]
:= [information processing based system]
⇒ *note**:
[The level of adaptability, purposefulness, activity in systems based on information processing can be very different.]

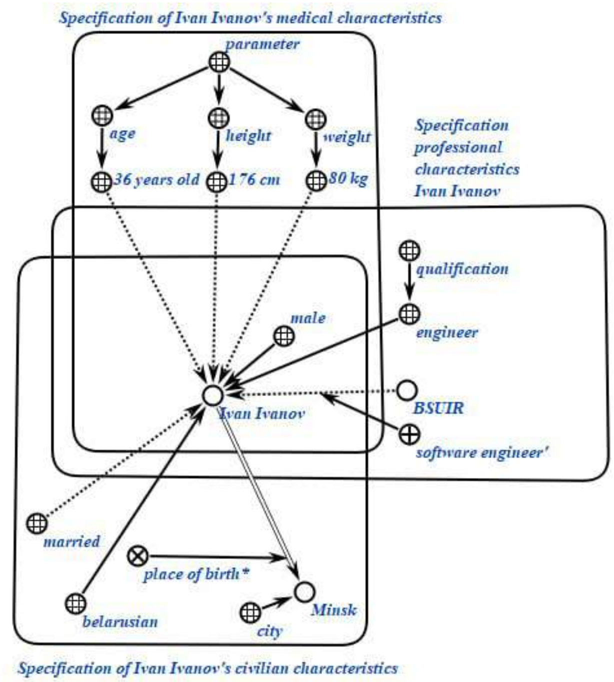


Figure 1. Examples of entity specifications for different sets of attributes of this entity

⇒ *partition**:
A sign of naturalness or artificiality of cybernetic systems
= {• *natural cybernetic system*
⊃ *human*
• *computer system*
• *symbiosis of natural and artificial cybernetic systems*
⊃ *community of computer systems and people*
}
⇒ *partition**:
Structural classification of cybernetic systems
= {• *simple cybernetic system*
• *individual cybernetic system*
• *multi-agent system*
}
⇒ *partition**:
Classification of cybernetic systems on the basis of the presence of a supersystem and the role within that supersystem
= {• *a cybernetic system that is not part of any other cybernetic system*
• *a cybernetic system embedded in an individual cybernetic system*
• *multi-agent agent*
}
⇒ *generalised decomposition**:
{• *information stored in the memory of the cybernetic system*
• *abstract memory of a cybernetic system*
• *cybernetic system problem solver*
• *physical shell of the cybernetic system*
}
⇒ *author**:
Glushkov V.M.
⇒ *Bibliographic source**:
Glushkov V.M.Cyber-1979st.
⇒ *standard bibliographic description**:
[V. Glushkov, "Cybernetics", 1979, pp. 850-856]

Whole presented example is a semantic neighbourhood of the concept "cybernetic system". As we can see, this example illustrates several types of knowledge described within the semantic neighbourhood of a given concept. Thus, for example, within the semantic neighbourhood of a given entity the theoretical-multiple relations of this entity with other entities, didactic relations linking the given entity with the information due to which the content of this entity is revealed and explained, and so on, are specified.

It is important to understand that the variety of types of semantic neighbourhoods indicates the variety of semantic types of descriptions of different entities. In the context of the *OSTIS Standard* we distinguish between full, basic and specialised semantic neighbourhoods, each of which is some kind of description of a particular entity.

For example, the structure of a complete semantic neighbourhood is determined primarily by the semantic typology of the entity being described. Thus, for example, for an absolute concept (class) the following information should be included in the full semantic neighbourhood if available:

- identification options in different external languages (sc-identifiers);
- membership in some subject domain, with an indication of the role performed within that subject domain;
- theoretical-multiple relations of a given concept with other objects;
- definition or explanation;
- statements describing the properties of the specified concept;
- problems and their classes in which this concept is key;
- a description of a typical example of the use of the specified concept;
- instances of the concept being described;
- authors and bibliographic sources of the specified concept;
- and others.

For a relative concept, i.e. a concept that is a relation, the semantic neighbourhood additionally specifies:

- domains;
- area of definition;
- relationship diagram;
- classes of relations to which the described relation belongs.

Obviously, there can be a large number of types of such specifications, since there can be a lot of information describing a particular object, and the need to obtain and visualise this information can be different. For the *OSTIS Glossary*, too, it is possible to specify a kind of semantic neighbourhood, with the help of which it will be possible to specify only what is important from the point of view of the *OSTIS Glossary* itself.

Drawing an analogy with the *OSTIS Standard*, all information in the *OSTIS Glossary* can be presented in the form of ordered and organized sections of the subject domains and ontologies that make up the *OSTIS Standard*. Each section can be presented as a sequence of objects and their specifications arranged alphabetically by the terms of these objects, i.e. in lexicographic order.

In the context of the *OSTIS Glossary*, the specification of each object should specify:

- identification options in various external languages (sc-identifiers);
- object membership in some subject domain with indication of the role performed within this subject domain;
- theoretical-multiple relations of a given object with other objects;
- definition or explanation of a given object;
- description of a typical example of the use of the specified object;
- instances of the described object, if the given object is a concept;
- authors of the given object;
- authors of the specification of the given object;
- analogs of the given object;
- bibliographic sources of the object.

In other words, in the context of the *OSTIS Glossary*, the semantic neighborhood of each specifiable object should include not just information defining and explaining the specified object, but also the information that determines the level of significance of this object in comparison with other objects. In this case, the *OSTIS Glossary* acts not only as a tool for consistency of some kind of activity, but also plays an important role in search and comparison of similar objects, i.e. it acts as a tool for convergence of different, but having common features, objects.

Thus, with the help of semantic neighborhoods it is possible to structure ("horizontally") knowledge about other knowledge. In order to stratify ("vertically") this knowledge about other knowledge among themselves, other kinds of structures have to be used. Therefore, all the *OSTIS Standard* objects are grouped by subject domains and their corresponding ontologies, which are used to stratify knowledge in knowledge bases.

The concept of the subject domain is the most important methodological technique, which allows to single out from the whole variety of the investigated World only a certain class of investigated entities and only a certain family of relations defined on the specified class. That is, localization, focusing attention only on it, abstracting from the rest of the World under study, is carried out.

The subject domains and their corresponding ontologies identified within the knowledge base of an intelligent system are semantic strata, clusters, which allow to "decompose" all the knowledge stored in the memory into "semantic shelves" in the presence of clear criteria that allow to unambiguously determine on which "shelf" certain knowledge should be located.

From the point of view of the *OSTIS Standard*, a subject domain is the result of integration (union) of partial semantic neighborhoods describing all investigated entities of a given class and having the same (common) subject of investigation (i.e. the same set of relations to which the mappings belonging to the integrated semantic neighborhoods should belong). That is, the subject domain is a structure that includes:

- the main studied (described) objects — primary and secondary;
- different classes of studied objects;
- different links, the components of which are the studied objects;
- different classes of the above-mentioned links (i.e. relations);
- different classes of objects that are neither the studied objects nor the above-mentioned links, but are components of these links.

Each concept corresponds to at least one subject domain in which the concept is a studied concept and in which the main characteristics of the concept are dealt with. When describing any subject domain, it is important that all classes declared by the studied concepts should be fully represented within the given subject domain together with their elements, elements of elements, etc. up to terminal elements.

For effective collective development and operation of the knowledge base of the ostis-system not just structuring is important, but such structuring, which is as objective as possible, having a clear semantic interpretation and allowing, on the basis of semantic links between structurally selected fragments of the knowledge base, to easily determine (localise) the "location" of either the knowledge being sought or new knowledge being introduced into the knowledge base. Such semantic structuring of the knowledge base, the formation of a system of semantically related "semantic shelves" on which specific knowledge satisfying clearly defined requirements is placed, significantly simplifies navigation through the knowledge base and clearly localises the evolution of knowledge located on each "semantic shelf". From a meaningful point of view, the subject domain is a set of factual statements describing all elements of a given set of objects of research with the help of a given set of relations and parameters (characteristics).

A section — is a sign of a set of all possible sections included in different knowledge bases. Each section represents a conditionally didactically distinguished fragment of the knowledge base, possessing logical integrity and completeness. In the limit, the whole knowledge base of a particular ostis-system is also one large non-atomic section.

For each partition it is necessary to explicitly specify the belonging to a set of atomic or non-atomic partitions. Atomic partition — a sign of the set of all possible atomic partitions included in various documentation, i.e., partitions not decomposable into more private partitions. Non-atomic section — the sign of the set of all possible non-atomic sections that make up the various documentations, that is, sections that are decomposed into more private sections.

In the context of the *OSTIS Glossary*, it may be appropriate to identify sections that will describe some portion of that information described in the relevant sections of the *OSTIS Standard*. But nobody forbids to consider the *OSTIS Glossary* as some atomic section in which all objects are listed according to the external natural language alphabet together with their specifications. As for subject domains, it is more expedient to use them not as a means for knowledge stratification, but as a means for searching already existing information to form sections with this information.

So, let's list the ***basic rules of structuring and specification of the OSTIS Glossary objects***:

- All information about the *OSTIS Glossary* objects should be represented in the form of semantic neighbourhoods of these objects, in which will be listed:
  - identification variants (variants of terms) of a given object in various external (natural) languages;
  - membership of the object in some subject domain with indication of the role performed within this subject domain, including membership of this object in the corresponding section of the knowledge base or the *OSTIS Standard*;
  - theoretical-multiple relations of the given object with other objects, including:
    - partition and decomposition of a given object into other objects;
    - including a given object into other objects;
    - and other possible unions, intersections that form a given object;
  - definition of the given object;
  - explanation of the given object;
  - description of a typical example of using the specified object;
  - instances of the described object, if the given object is a concept;
  - authors of the given object;
  - authors of the specification, i.e. the authors who described the given object in the knowledge base;
  - analogues of this object, including:
    - close analogues of the given object;
    - differences and similarities with other objects, including listed analogues;
  - bibliographic sources of the object;
  - possible quotations, aphorisms, metaphors and epigraphs related to the given object; This variant of specification will be called dictionary specification.
- In this case, all objects and their specifications within the *OSTIS Glossary* can be ordered:
  - as an enumeration of these objects in the lexigraphic order of their terms, forming one single atomic section, respectively being the *OSTIS Glossary*;
  - and the enumeration of sections corresponding to the sections of the *OSTIS Standard*, uniting objects by common features, considered within a particular section, and representing sequences of these objects in the lexigraphic order of their terms.

  For the second case, it is important that the sections do not overlap with each other, i.e., do not duplicate descriptions of the same entity. From this point of view, the first option is easier to implement, because it does not require taking into account the possibility of occurrence of the found object in the already formed sections of the *OSTIS Glossary*.
- Structuring of objects of the *OSTIS Glossary* should be reduced to the formation of text from the already structured text of the *OSTIS Standard*.

D. Rules for identification of the OSTIS Glossary objects

In the previous section the rules of structuring and specifying (standardisation) of the *OSTIS Glossary* objects were fixed. However, it is necessary to standardise not only those texts that are written directly in *SC-code*, but also those texts that are written in external, e.g. natural languages. One of the features of the *SC-code* is that with it it is also possible to record some natural language files, by means of which the comprehension of the text by any human being is improved. Such a possibility is realised with the help of ostis-system files, with the help of which external information constructions which are not text in *SC-code* [43] are denoted. With their help it is possible to specify information in an external natural language for all objects in the knowledge base. A special case of ostis-system files are files denoting identifiers of objects in the knowledge base.

Identifier is a structured sign representing an entity denoted by a string of symbols. An identifier is an information construct (most often a string of symbols) providing unambiguous identification of the corresponding object described in knowledge bases of ostis-systems, and is, most often, a name (term) corresponding to the described object, a name denoting this object in external texts of ostis-systems.

In formal texts, identifiers must be unique to uniquely match an object. Each pair of identical identifiers must denote the same object.

All objects in a knowledge base have the following ***common identification rules***:

- The membership of the identified object in some object classes is explicitly specified in the external identifier of this object (in the sc-identifier) by means of appropriate conditional attributes:
  - if the last character of the sc-identifier is an "asterisk" character, then the identified object belongs to the Class of non-role relation designations;
  - if the last character of the sc-identifier is an apostrophe, then the identified object belongs to the Class of role relationship designations, each of which is a subset of Membership relation;
  - if the last character of the sc-identifier is "^", then the identified object belongs to the Parameter designation class.
- For each object, you can construct an sc-identifier, which is a proper name that always starts with a capital letter.
- If an object is a designation of some class of objects, then this object can be matched not only with a proper name, but also with a common name, which starts with a small (lowercase) letter. The specification of each class (each concept) includes a list of equivalent (synonymous) sc-identifiers, among which there are both proper and nominative names.
- Identification of partitions in knowledge bases is performed at the expense of identification of partition objects. Instances of partition classes within the Russian language are named according to the following rules:
  - at the beginning of the identifier the word Section is written and a dot is put;
  - followed by the name of the section with a capital letter, reflecting its content.

Obviously, the same rules apply to the *OSTIS Glossary* objects. Therefore, there is no need to describe any additional rules for identifying the *OSTIS Glossary* objects.

Besides identifiers of ostis-systems knowledge base objects, it is possible to standardise the form of presentation of information in the specification of these objects: both the style of writing the text itself and the information with the help of which it is possible to explain the basic information in the specification of these objects.

E. Key elements of didactic information in the OSTIS Glossary object specifications

The most important criterion of quality of created ostis-systems of any purpose is to create conditions so that insufficiently qualified users of each ostis-system (both end-users and those responsible for its effective operation and modernisation) could acquire the required qualification quickly enough with the help of the same ostis-system. This means that each ostis-system, irrespective of its direct purpose (automation of specific types of human activities in a particular field) should also be a training system, i.e. it should be able to train its users in the direction of improving their qualification. A qualified user of any category must understand the capabilities of the ostis-system with which he interacts, must understand what the system knows and can do, as well as how its activities can be managed. Lack of understanding between ostis-systems and their users — is a violation of the interoperability requirement [44] imposed on both ostis-systems and their users.

Therefore, the key stage in the development of any knowledge base, and, in general, any information resource is the stage of development and implementation of didactic information. Didactic information should be understood as a specification of the subject domain, which provides additional information designed to enable users and developers (knowledge engineers), who use or improve the specified subject domain and its ontology, to learn their features faster. Didactic information enables [45]:

- to quickly and adequately assimilate the denotational semantics of knowledge stored in the system;
- to provide a deeper understanding and assimilation of the meaning of various kinds of entities (including various knowledge);
- to establish mutual understanding between systems and their users;
- to accelerate the process of formation of the required qualification of users in various fields.

The "didactic" effect of didactic information is provided by:

- by sufficient detail of the studied entity (completeness of the semantic neighbourhood describing the relations of this entity with other entities)
  - decomposition of the entities under consideration;
  - by specifying analogues (similar entities in different senses);

- indicating metaphors (epigraphs);
- indicating antipodes (entities that differ in different senses);

- exercises — solutions to various problems using the entities studied;
- references to knowledge stored within the same knowledge base;
- references to bibliographic sources.

Following one of the goals of the *OSTIS Glossary*, namely to provide a quality and understandable text for the end user, it is necessary to describe and record the ***main elements of didactic information for the objects of the OSTIS Glossary***. Such elements <u>should</u> include:

- information by means of which the basic information in the *OSTIS Glossary* object specification is defined or explained, i.e.:
  - various types of definitions, explanations, and annotations for that object;
  - examples of how to use this object;
  - instances of this object, if it is a concept;
- information describing distinctive and similar characteristics between the *OSTIS Glossary* objects, including:
  - analogies, correspondences, corollaries between objects;
  - differences and similarities between objects;
- information that supports the significance, scientific novelty, and practical applicability of the *OSTIS Glossary* object, including:
  - of the authors of these objects, as well as the authors who specified this object;
  - bibliographic sources of these objects;
  - quotations, aphorisms, metaphors and epigraphs related to the given object;
  - other.

It should be noted that the quality of the information described in the knowledge base directly depends on the quality of the presentation of this information and the quality of the information with the help of which the basic information in the knowledge base is explained. The more qualitatively the information in the knowledge base is described, structured and stratified, the lower the requirements to the readers of this information and the higher the level of understanding of the information in the knowledge base by these readers [46].

Didactic information in the knowledge base determines the level of quality of the information described in the knowledge base. The key elements of didactic information are those elements that contribute to easier and deeper mastering of the basic information in the knowledge base. The authors believe that it is the information that reveals similarities between objects in the knowledge base that is the key to improving the quality of the entire knowledge base.

F. Conclusion

To summarize, it is important to note the following points:

- The *development of the OSTIS Glossary* is reduced to the development of the *OSTIS Standard* and a set of tools that allow to form and improve the *OSTIS Glossary* on the basis of the *OSTIS Standard*;
- When developing the *OSTIS Glossary* as part of the *OSTIS Standard*, it is important to fix and improve the principles and rules:
  - specification and formalisation of objects;
  - identification of objects and their specifications;
  - of structuring and stratification of object specifications;
  - development and consistency of objects and their specifications;
- The *quality of the OSTIS Glossary* is determined by:
  - quality of the *OSTIS Standard*, which is defined by:
    * the quality of the information it contains;
    * the quality of the means to improve this information, which is determined by:
      · the quality of the methods, principles and rules for developing this information;
      · the quality of didactic information explaining this information;
    * the competence and level of training of the developers of the *OSTIS Standard*.

In the final section the current state of the *OSTIS Glossary* within the *OSTIS Metasystem* will be considered, the principles of automatic formation of the *OSTIS Glossary*, information retrieval in it will be fixed, and also the current Author team of the *OSTIS Glossary* and requirements to its developers will be considered.

IV. Implementation of the *OSTIS Glossary* within the *OSTIS Metasystem*

A. Current specification and structure of the OSTIS Glossary

It is important to note that the *OSTIS Glossary* is not just a dynamically generated text from the text of the *OSTIS Standard*, which is a simplified representation of the *OSTIS Standard*. The *OSTIS Glossary* acts as a means to provide a consistent and interoperable activity for the development of new generation intelligent computer systems, and thus has documentation for its use and development. Therefore, first of all, the *OSTIS Glossary* is documentation on how to properly form a dictionary representation of the *OSTIS Standard*.

The OSTIS Glossary
:= [Semantic electronic dictionary of Artificial intelligence]

:= [Glossary of the *OSTIS Standard* terms and concepts]
:= [Semantically interoperable dictionary of concepts arranged in lexicographic order of their terms]
:= [A tool to ensure consistent and interoperable activities for the development of new generation intelligent computer systems]
:= [A variant of displaying the *OSTIS Standard* as a sequence of concepts and their dictionary specifications, presented in lexicographical order of the terms of these concepts]
:= [The result of different perspectives consistency in the field of information technology]
⇐ *form of presentation**:
OSTIS Standard
⊂ *OSTIS Standard*
∈ *knowledge base fragment*
∈ *semantic dictionary*
⇒ *generalised decomposition**:
{• *A sequence of concepts and their dictionary specifications, arranged in lexicographic order of the terms of these concepts*
:= [Dynamically formed atomic or non-atomic section of the *OSTIS Metasystem* knowledge base, representing a sequence of concepts and their dictionary specifications of the *OSTIS Standard*, arranged in lexicographic order of the terms of these concepts]
∈ *dynamic structure*
⊂ *OSTIS Standard*
⊂ *Knowledge base of the OSTIS Metasystem*
⊃ *Documentation on the development and use of the OSTIS Glossary*
• *OSTIS Glossary browsing and navigation subsystem*
:= [Collective of agents providing the *OSTIS Glossary* generation from the *OSTIS Standard* and its viewing in the system]
⊂ *Problem solver of the OSTIS Metasystem*
}

Documentation on the development and use of the OSTIS Glossary
∈ *knowledge base section*
∈ *ostis-documentation*
⇐ *section concatenation**:
⟨• *Principles and rules of the development and consistency of the OSTIS Glossary and the OSTIS Standard*
• *Rules of placement and specification of the OSTIS Glossary objects*
• *Rules of identification of the OSTIS Glossary objects*
• *Rules of visualisation of the OSTIS Glossary object specification*
• *Principles of interaction between users and computer systems with the OSTIS Glossary*
• *Structure of the OSTIS Glossary*
• *Features and advantages of the OSTIS Glossary*
• *Author team of the OSTIS Glossary*
• *Prospects for development of the OSTIS Glossary*
⟩

We will consider the specifications of the *OSTIS Glossary* generation agents from the *OSTIS Standard*, as well as the specifications of the *OSTIS Glossary* browsing agents.

B. Specification of the OSTIS Glossary Formation Agents from the OSTIS Standard

The *OSTIS Glossary* is nothing but a variant of the *OSTIS Standard* display. It is up to a specialised module to display the *OSTIS Glossary* to the end user.

Within the *OSTIS Technology*, the only kind of entities performing transformations in the memory of systems are agents — some entities capable of performing actions in the memory of these systems, belonging to some specific class of actions.

In order to map the *OSTIS Glossary*, the following tasks must be performed automatically:

- to find and transform information from the *OSTIS Standard* into some dictionary form, in which all information is represented as a lexicographic sequence of concepts and their specifications;
- to integrate several sections of the same dictionary among themselves;
- to be able to filter the dictionary depending on the characteristics of the end user.

The following agents have been developed to solve these problems:

- The *OSTIS Glossary* section formation agent, which deals with the transformation of a logically stated the *OSTIS Standard* section into an the *OSTIS Glossary* section in which all objects are ordered lexicographically;
- The *OSTIS Glossary* section concatenation agent, designed to set the sequence between several generated the *OSTIS Glossary* sections;
- The *OSTIS Glossary* section filtering agents, forming:

– a simplified section of the *OSTIS Glossary* that contains objects without their specifications, but arranged in lexicographic order;
– simplified section of the *OSTIS Glossary*, which contains objects with their specifications without theoretical-multiplicity relations, arranged in lexicographic order;
– simplified section of the *OSTIS Glossary*, which contains objects with their specifications without didactic links, arranged in lexicographic order.

The purpose of the *OSTIS Glossary* section agent is to create a sequence of all objects belonging to the corresponding section of the *OSTIS Standard* in lexicographic order. This agent implements the following algorithm:

- Step 1: A section already existing in the knowledge base is specified as an argument.
- Step 2. A new section of the *OSTIS Glossary* is created and knowledge base objects and their corresponding dictionary specifications are added to it.
- Step 3. All objects are organised in alphabetical order.

The *OSTIS Glossary* section filtering agent allows you to simplify the display of the section for convenient use for specific purposes. The principle of operation is as follows:

- Step 1: The *OSTIS Glossary* section and the display option are specified as an argument.
- Step 2: Depending on the display option the agent changes the structure of the *OSTIS Glossary* section.

C. Specification of agents for navigating the OSTIS Glossary

One of the important tasks is the need to search for information in the *OSTIS Glossary* fast enough. To solve this problem, search agents have been developed, in particular:

- An agent for searching authors of a given object;
- An agent for searching analogues of a given object;
- An agent for searching for differences between a given object and its analogues;
- An agent for searching for objects developed by a given scientist;
- An agent for searching the definition of a given concept;
- Agent for searching relations defined on the concept;
- Agent for searching for concepts through which the given concept is defined;
- An agent for searching all entities that are specialized with respect to the given concept;
- An agent for searching authors of the specification of a given entity;
- An agent for searching identification rules for a given entity;
- and others.

D. Author team of the OSTIS Glossary

A key feature and also a key advantage of the *OSTIS Glossary* is its author team. Obviously, the authors of this paper are members of the author team of the *OSTIS Glossary*.

The ***highlight*** of the whole the *OSTIS Glossary* development activity lies in the authors themselves, the developers of the *OSTIS Glossary* — most of those involved in the development of the *OSTIS Glossary* and the writing of this paper are first-year students of the speciality "Artificial intelligence". The impetus for the creation of this creative team is as follows:

- First-year students have sufficient and, most importantly, "fresh" (!) learning experience: they understand and realise the problems related to the form of presentation and search of educational material, "gluing" information from different academic disciplines, application of the learned information in practice. Who but first-year students, who have gone through all the problems of pre-university education, are able to realise, understand and describe them in a form that will be understandable not only to highly qualified specialists, but also to the new generation of first-year students.
- Students are full of ambition and enthusiasm. The aspiration of a young person to learn new information, to create something unique contributes to the development of the whole society. From the school bench, students are the backbone of all activities for the development and improvement of all mankind. The *OSTIS Glossary* is a good starting point for the development of students.
- In addition to all this, students benefit greatly. Studying at such practice-oriented specialities as "Artificial intelligence", students not only study and get all the necessary information for developing themselves as professionals in this field, but also develop social qualities, i.e. become personalities, which is very important for their future work. Scientific-research activity, first of all, is not a creative activity, but a social one, and only effective social interaction will help to create solutions to existing problems in society.
- It should be recognised that the participation of students in such activities is also beneficial for the *OSTIS Technology* itself. Undoubtedly, the quality of any technology is determined by the level of preparedness of its developers. But every technology exists and develops as long as there is not only some need in it, but also when there is a constant replenishing and growing creative team of interoperable people.
- Yes, it should be noted that students are more interoperable than many other existing specialists in the world. It is easier to get in touch with them, it

is easier to communicate, to agree, to solve tasks together. The key problem of the current state of the whole society is that it is not capable of solving problems of a serious enough level.

Creating and organising such a team is not the easiest task. Let's reveal the secret of how to form such a team:

- First of all, it is important that the initiators of all these activities are not just teachers who are motivated by student learning and development, but also teachers who are engaged in both science and application development.
- And most importantly, such specialists should realise that nothing in the world is done by one person, that it is necessary not only to improve oneself, but also to contribute to the development and improvement of others, and only in this case society will be able to develop harmoniously in all its directions.

Students and teachers have the same requirements. They <u>must</u> possess:

- A high level of professional qualities, including:
  - a high level of system culture, i.e. the ability to think abstractly, argue a point of view, draw correct conclusions, etc.;
  - a high level of mathematical culture, i.e. the culture of formalisation;
  - a high level of technological and engineering culture, i.e. the ability to apply theoretical knowledge in practice, invent and implement ideas, etc;
- a high level of social qualities, including:
  - a high level of interoperability (!), as defined by:
    - a high level of social responsibility, i.e. responsibility for the tasks they have to perform;
    - a high level of social engagement, i.e. the ability to make decisions, to create ideas, to be the engine of the team;
    - a high level of agreement ability, i.e. the ability to create mutually beneficial conditions with other team members;
  - a high level of moral and ethical qualities.

The following directions of development of the current Author team of the *OSTIS Glossary* are worth mentioning:

- to create a favourable environment and conditions for the development and training of the members of this team;
- to create conditions for the rapid entry of new people;
- to create conditions for the accumulation and reuse of experience in the team.

E. Prospects for development of the OSTIS Glossary

The following directions can be set for the *OSTIS Glossary*:

- to improve the text of the *OSTIS Glossary*, which implies improving the text of the *OSTIS Standard*, including:
  - to provide a complete specification of the objects described within the *OSTIS Glossary*;
  - to provide sufficiently complete didactic information in the specifications of these objects, including:
    - full description of authors, bibliographic sources of these objects;
    - comparative description with similar objects within the *OSTIS Glossary*;
    - a sufficiently detailed description of the explanatory information to the basic information of these objects.
- translation of concept terms into other natural languages, including English;
- to improve the tools for browsing and navigating the *OSTIS Glossary*;
- improvement of tools to automate the process of updating the *OSTIS Ecosystem* knowledge base;
- development of the Author team of the *OSTIS Glossary* and attraction of new specialists and developers.

V. Conclusion

In summary, the result of this work is a comprehensive tool to ensure <u>consistent</u> and <u>comatible</u> activities both in the development of new generation intelligent systems and in any other field — the *OSTIS Glossary*, <u>allowing</u>:

- to represent any information in the same form which:
  - makes this information understandable and consistent not only for humans, but also for computers;
  - simplifies the processing of this information;
  - simplifies the convergence and integration of different types of knowledge;
  - makes it possible to integrate information from different information sources;
- to describe and structure information both from one subject domain and information at "junctions" between subject domains which:
  - makes it easier to provide consistency of existing information and addition of new information;
  - makes it easier to find this information and integrate new information;
  - makes it possible to build knowledge libraries, i.e. to reuse existing knowledge;
  - makes it possible to describe information in different external languages in a consistent form;
- to standardise the description and visualization of information of various kinds, which:

– simplifies the consistency of existing information and the addition of new information;
– standardizes support for methods and tools to enhance existing information;
– enables the creation of a personalized experience for any end-user.

It is also worth noting that the solution proposed within this paper — *OSTIS Glossary* — is:

- a part of the *OSTIS Metasystem* Knowledge base, as known the *OSTIS Standard*, which allows:
  – to develop the *OSTIS Glossary* by the same means by which any intelligent computer system based on the *OSTIS Technology* is developed;
  – use the same tools to view and navigate the text of the *OSTIS Glossary*;
  – automatic consistent development of the *OSTIS Glossary* and the *OSTIS Standard*;
- a simplified version of the *OSTIS Standard*, which allows:
  – to quickly search and reuse existing information;
  – to quickly provide consistency and integrate new information;
  – to reduce the circle of entry for new people to develop the *OSTIS Technology*;
- an environment for social and creative learning and development of new staff in the field of Artificial Intelligence.

The introduction of such information resources can significantly improve the quality and efficiency of various activities.

The authors believe that this paper will be useful not only for those who are researching innovative methods and technologies for more effective organisation of teamwork, but also for those who are just beginning research in this area.

VI. Acknowledgment

The authors would like to thank the scientific staff of the Department of Intelligent Information Technologies of the Belarusian State University of Informatics and Radioelectronics for their help and valuable comments.

References

[1] V. Golenkov, N. Guliakina, V. Golovko, V. Krasnoproshin, "Methodological problems of the current state of works in the field of artificial intelligence," in *Otkrytye semanticheskie tekhnologii proektirovaniya intellektual'nykh system [Open semantic technologies for intelligent systems]*, ser. 5, V. Golenkov, Ed. BSUIR, Minsk, 2021, pp. 17–24.

[2] A. M. Ouksel, "Semantic interoperability in global information systems," *SIGMOD Rec.*, vol. 28, no. 1, p. 5–12, Mar. 1999.

[3] V. Golenkov, V. Golovko, N. Guliakina, and V. Krasnoproshin, "The standardization of intelligent computer systems as a key challenge of the current stage of development of artificial intelligence technologies," in *Otkrytye semanticheskie tekhnologii proektirovaniya intellektual'nykh system [Open semantic technologies for intelligent systems]*, ser. 4, V. Golenkov, Ed. BSUIR, Minsk, 2020, pp. 73–88.

[4] V. Golenkov, N. Guliakina, V. Golovko, and V. Krasnoproshin, "Artificial intelligence standardization is a key challenge for the technologies of the future," in *Open Semantic Technologies for Intelligent System*. Cham: Springer International Publishing, 2020, pp. 1–21.

[5] P. Hagoort, "Semantic unification," in *The cognitive neurosciences, 4th ed.* MIT press, 2009, pp. 819–836.

[6] A. Iliadis, "The tower of babel problem: Making data make sense with basic formal ontology," *Online Information Review*, vol. 43, no. 6, pp. 1021–1045, 2019.

[7] J. H. Siekmann, "Universal unification," in *International Conference on Automated Deduction*. Springer, 1984, pp. 1–42.

[8] A. P. Sokolov and A. O. Golubev, "System of computer-aided design of composite materials. part 3. graph-oriented methodology of development of means of user-system interaction," 2021, pp. 43–57.

[9] G. Evgenev, "Expertopaedia as a means of creating an ontological internet of knowledge," *Ontology of Design*, vol. 9, no. 3 (33), pp. 307–320, 2019.

[10] V. Polonskiy, "Why do i need a dictionary?" *Nauka i shkola*, no. 1, pp. 214–226, 2019.

[11] M. Yahya, J. G. Breslin, and M. I. Ali, "Semantic web and knowledge graphs for industry 4.0," *Applied Sciences*, vol. 11, no. 11, p. 5110, 2021.

[12] J. Han, S. Sarica, F. Shi, and J. Luo, "Semantic networks for engineering design: state of the art and future directions," *Journal of Mechanical Design*, vol. 144, no. 2, p. 020802, 2022.

[13] Robinson, I., *Graph databases*. O'Reilly Media, Inc., 2015.

[14] V. Syskov and V. Borisov, "Method of organisation of collective activity in complex organisational and technical systems," *Sovremennye nauchnye issledovaniye i innovatsii*, no. 11, pp. 284–310, 2015.

[15] W. J. Verhagen, P. Bermell-Garcia, R. E. Van Dijk, and R. Curran, "A critical review of knowledge-based engineering: An identification of research challenges," *Advanced Engineering Informatics*, vol. 26, no. 1, pp. 5–15, 2012.

[16] T. Gruber, "Toward principles for the design of ontologies used for knowledge sharing," *Intern. J. of Human-Computer Studies*, vol. 43, no. 5/6, pp. 907–928, 1995.

[17] A. Palagin and N. Petrenko, "To the question of system-ontological integration of subject area knowledge," *Mathematical Machines and Systems*, vol. 1, no. 3-4, pp. 63–75, 2007.

[18] Lim, S.C. Johnson and Liu, Ying and Chen, Yong, "Ontology in design engineering: status and challenges," in *International Conference on Engineering Design 2015 (ICED 2015)*, 07 2015.

[19] L. Yang, K. Cormican, and M. Yu, "Ontology-based systems engineering: A state-of-the-art review," *Computers in Industry*, vol. 111, pp. 148–171, 2019.

[20] I. Davydenko, V. Tarasov, and A. Fedotova, "Ontological approach to building knowledge bases based on semantic networks," in *Open semantic technologies for designing intelligent systems (OSTIS-2016) : proceedings of the VI International scientific-technical conference, Minsk, 18–20 Feb. 2016 y.* Minsk, 2016.

[21] O. Ataeva, N. Kalenov, and V. Serebryakov, "Ontological approach to the description of a unified digital space of scientific knowledge," *Electronic Libraries*, vol. 24, no. 1, pp. 3–19, 2021.

[22] Y. Kravchenko, "Synthesis of heterogeneous knowledge on the basis of ontologies," *Investigations of the Southern Federal University. Technical Sciences*, vol. 136, no. 11 (136), pp. 216–221, 2012.

[23] A. Tuzovskiy, "Development of knowledge management systems based on a unified ontological knowledge base," *Investigations of Tomsk Polytechnic University. Engineering of georesources*, vol. 310, no. 2, pp. 182–185, 2007.

[24] T. Dillon, E. Chang, and P. Wongthongtham, "Ontology-based software engineering-software engineering 2.0," 04 2008, pp. 13–23.

[25] V. Ryen, A. Soylu, and D. Roman, "Building semantic knowledge graphs from (semi-) structured data: a review," *Future Internet*, vol. 14, no. 5, p. 129, 2022.

[26] P. Barnaghi, A. Sheth, and C. Henson, "From data to actionable knowledge: Big data challenges in the web of things [guest editors'

introduction],” *IEEE Intelligent Systems*, vol. 28, no. 6, pp. 6–11, 2013.

[27] C. Kellogg, “From data management to knowledge management,” *Computer*, vol. 19, no. 01, pp. 75–84, 1986.

[28] G. Fischer and J. Otswald, “Knowledge management: problems, promises, realities, and challenges,” *IEEE Intelligent systems*, vol. 16, no. 1, pp. 60–72, 2001.

[29] S. E. Hampton and J. N. Parker, “Collaboration and productivity in scientific synthesis,” *BioScience*, vol. 61, no. 11, pp. 900–910, 2011.

[30] I. Sechkina and G. Sechkin, “Synthesis as a goal, method and final result of knowledge integration,” *Omskiy nauchny vestnik*, no. 3 (129), pp. 191–192, 2014.

[31] Y. Zagorulko and O. Borovikova, “An approach to building portals of scientific knowledge,” *Autometry*, vol. 44, no. 1, pp. 100–110, 2008.

[32] P. Van Baalen, J. Bloemhof-Ruwaard, and E. Van Heck, “Knowledge sharing in an emerging network of practice:: The role of a knowledge portal,” *European Management Journal*, vol. 23, no. 3, pp. 300–314, 2005.

[33] R. Mack, Y. Ravin, and R. J. Byrd, “Knowledge portals and the emerging digital knowledge workplace,” *IBM systems journal*, vol. 40, no. 4, pp. 925–955, 2001.

[34] V. V. Golenkov and N. A. Gulyakina, “Graphodynamic models of parallel knowledge processing: principles of construction, implementation and design,” in *Open semantic technologies of designing intelligent systems (OSTIS-2012) : proceedings of the II International scientific-technical conference, Minsk, 16–18 Feb. 2012 y.* Minsk, 2012, pp. 23–52.

[35] V. V. Golenkov, “Ontology-based design of intelligent systems,” in *Otkrytye semanticheskie tekhnologii proektirovaniya intellektual’nykh system [Open semantic technologies for intelligent systems]*, ser. Iss. 1. Minsk : BSUIR, 2017, pp. 37–56.

[36] V. Golenkov, N. Guliakina, I. Davydenko, A. Eremeev, “Methods and tools for ensuring compatibility of computer systems,” in *Otkrytye semanticheskie tekhnologii proektirovaniya intellektual’nykh system [Open semantic technologies for intelligent systems]*, ser. 4, V. Golenkov, Ed. BSUIR, Minsk, 2019, pp. 25–52.

[37] A. Zhmyrko, “Family of external languages of next-generation computer systems, close to the language of the internal semantic representation of knowledge,” in *Otkrytye semanticheskie tekhnologii proektirovaniya intellektual’nykh system [Open semantic technologies for intelligent systems].* Minsk : BSUIR, 2022, pp. 65–80.

[38] K. Bantsevich, “Metasystem of the ostis technology and the standard of the ostis technology,” in *Otkrytye semanticheskie tekhnologii proektirovaniya intellektual’nykh system [Open semantic technologies for intelligent systems]*, ser. Iss. 6, V. Golenkov, Ed. BSUIR, Minsk, 2022, pp. 357–368.

[39] V. V. Golenkov and N. A. Gulyakina, “Open project aimed at creating a technology for component-based design of intelligent systems,” in *Open Semantic Technologies for Designing Intelligent Systems (OSTIS-2013) : proceedings of the III International Scientific and Technical Conference, Minsk, 21-23 Feb. 2013 y.* Minsk, 2013, pp. 55–78.

[40] A. Zagorskiy, “Principles for implementing the ecosystem of next-generation intelligent computer systems,” in *Otkrytye semanticheskie tekhnologii proektirovaniya intellektual’nykh system [Open semantic technologies for intelligent systems].* BSUIR, Minsk, 2022, p. 347–356.

[41] V. Ivashenko, “General-purpose semantic representation language and semantic space,” in *Otkrytye semanticheskie tekhnologii proektirovaniya intellektual’nykh system [Open semantic technologies for intelligent systems]*, ser. Iss. 6. Minsk : BSUIR, 2022, pp. 41–64.

[42] K. Bantsevich, “Structure of knowledge bases of next-generation intelligent computer systems: a hierarchical system of subject domains and their corresponding ontologies,” in *Otkrytye semanticheskie tekhnologii proektirovaniya intellektual’nykh system [Open semantic technologies for intelligent systems]*, ser. Iss. 6, V. Golenkov, Ed. BSUIR, Minsk, 2022, pp. 87–98.

[43] A. Goylo and S. Nikiforov, “Means of formal description of syntax and denotational semantics of various languages in next-generation intelligent computer systems,” in *Otkrytye semanticheskie tekhnologii proektirovaniya intellektual’nykh system [Open semantic technologies for designing intelligent systems (OSTIS-2022)].* Minsk, 2022, pp. 99–118.

[44] A. Zagorskiy, “Factors that determine the level of intelligence of cybernetic systems,” in *Otkrytye semanticheskie tekhnologii proektirovaniya intellektual’nykh system [Open semantic technologies for intelligent systems].* BSUIR, Minsk, 2022, p. 13–26.

[45] O. Parshina, “A question about lexicographic description of didactic terms,” *Vestnik Volzhskogo universitet n.a. Tatishchev*, no. 7, pp. 34–40, 2011.

[46] A. Tazetdinov, “Technology of structuring and visualisation of educational information in tutoring systems,” *Information and control systems*, no. 1, pp. 60–65, 2009.

ГЛОССАРИЙ OSTIS — ИНСТРУМЕНТ ДЛЯ ОБЕСПЕЧЕНИЯ СОГЛАСОВАННОЙ И СОВМЕСТИМОЙ ДЕЯТЕЛЬНОСТИ ПО РАЗРАБОТКЕ ИНТЕЛЛЕКТУАЛЬНЫХ СИСТЕМ НОВОГО ПОКОЛЕНИЯ

Зотов Н. В., Ходосов Т. П., Остров М. А., Позняк А. В., Романчук И. М., Рублевская Е. А., Семченко Б. А., Сергиевич Д. П., Титов А. В., Шаров Ф. И.

Данная работа включает подробный анализ проблем организации различных видов коллективной деятельности, сравнительный анализ текущих решений по обеспечению согласованности и совместимости информации из различных областей знаний, а также анализ методов и технологий для создания единых информационных пространств для обеспечения согласованного и совместимого хранения, обработки, накопления и распространения знаний. В работе предлагается один из вариантов реализации единого информационного ресурса для обеспечения согласованной и совместимой деятельности по разработке интеллектуальных компьютерных систем нового поколения — Глоссарий OSTIS. Описывается его структура, правила структуризации, размещения и идентификации знаний в нём, а также принципы работы с ним.

Received 15.03.2024

Fundamentals for the Intelligent Non-Invasive Diagnostics

Natallia Lipnitskaya
Department of Intelligent Information Technologies
Belarusian State University
of Informatics and Radioelectronics
Minsk, Republic of Belarus
Natasha.lipnitskaya@gmail.com

Vladimir Rostovtsev
Republican Scientific and Practical Center
of Medical Technologies, Informatization,
Management and Economics of Health Care
Minsk, Republic of Belarus
vnrost@kmsd.su

***Abstract*—The article elaborates the needs of the design and implementation of a intelligent non-invasive diagnostics system. Technological basis for development and different variants of non-invasive diagnostics are proposed as two fundamental components of such system.**

The domestic Open Semantic Technology of Intelligent Systems (OSTIS) is proposed to be used as a core technological foundation while designing the intelligent diagnostic system. The adaptation of diagnostic tasks within logical-semantic approach will allow to carry out differential diagnostics (i. e. formulating several diagnostic hypotheses). Various approaches towards the non-invasive diagnostics have been considered: functional-spectral diagnostics (FSD-diagnostics), bioimpedance analysis, preliminary diagnostics based on the assessment of the basic parameters of functional state, diagnostics by Zakharyin-Ged zones, diagnostics by Nakatani method, frequency-resonance diagnostics.

***Keywords*—non-invasive diagnostics, artificial intelligence, diagnostic decision support system**

I. Introduction

Health is the most valuable resource of the state. One of the task of modern society is to timely detect the disease risk. The implementation of this task requires new diagnostic tools based on the latest technologies. Risk diagnosis will provide significant economics savings towards disease prevention and treatment, as well as improve the quality of primary health care. Risk is the probability of developing a disease [1].

The current problem in the area of risk diagnostic is the creation of non-invasive technology for examination and detection of diseases at early stages in order to carry out individualized prevention. The emerging modern technologies provide ample opportunities for solving this problem [2]. At the same time, let us quote a doctor's critical statements about informatization of medicine: "The global problem is the lack of resources. And we are not talking about the shortage of money, but about the shortage of time. The time of professionals is the main world deficit. Information technology offers great opportunities to save money. Telemedicine, for example, has a huge potential. Support for medical decision-making is of enormous value, but it is not being deployed and practiced" [3].

The importance of the intelligent non-invasive diagnostics problem has several aspects. Firstly, it is **caring** for people's health that leads towards the individual health improvement and preventive care. Secondly, it is an **increase in the quality** of individual preventive care to the population. Thirdly, it is **beneficial** from the economic point of view, as the costs of prevention and treatment are minimized. Taking into account the problem of "time shortage", it is important to minimize time costs, as the procedures are carried out quickly enough. It is important that non-invasive diagnostics procedures are safe and painless.

Therefore, there is a need to continue to investigate and develop the intelligent non-invasive diagnostics, with the primary focus on the development of an intelligent system to support the decision making for non-invasive diagnostic.

The proposed architecture for the intelligent diagnostic ostis-system allows to assess the risk of diseases in patients, and creates the "windows of opportunity" not only for patients and doctors, but also for developers in terms of expanding the functionality of the system.

The main aim of this paper is to create an intelligent non-invasive diagnostic system architecture suitable for screening of systemic and nosological risks and early diagnosis of diseases, i. e. for diagnosing latent and initial stages of pathological process development for the purpose of primary and secondary prevention or timely treatment.

II. Overview of Existing Solutions

The quality of medical care depends on the level of doctors' training and on systems that support decision-making, including in the field of diagnosing the diseases at various stages.

While there are many medical decision support systems in various fields, the deployment of such systems into the everyday practice is relatively slow.

The current state of art of existing medical decision support systems is presented in [4]–[7] and demonstrates that almost every system is focused on a specific disease or group of diseases.

To date, a large number of private, highly specialized decision support systems have been developed. For example, SkyChain is designed for diagnosing lung, liver, breast, and melanoma cancer. The IDDAP system identifies potential infectious diseases and disease states based on the constructed ontology of the subject area. MYCIN is an interactive expert system for diagnosis and treatment of infectious diseases [8]. In [9], an intelligent system for personalized human health monitoring based on biomedical signal processing is designed using the Internet of Things, cloud computing, big data processing and neural network.

The number of localized problem statements and their solutions is so large that it is almost impossible to catalog them and have a common source of information about them. Localized solutions can be found in various information and search engines: PubMed, Scopus, Google Scholar, WoS.

It is difficult for medical staff to use several systems at once in practice, and it is expensive to develop and maintain such systems. Most of the existing systems focus on diagnostics on late stages of disease, while virtually none of the existing system considers the diagnosis in the early stages of the disease.

In terms of diagnostic methods, this article focuses on non-invasive methods of examination for early stages of the disease. Since they are highly informative, do not require long additional preparation of the patient, significant time expenditures, and also during the procedure the integrity of the skin is preserved.

III. Problem Statement

The design and implementation of intelligent non-invasive diagnostic system will require the consideration of identification of technological platform capable to process and support the non-invasive diagnostics. The methodological approach should address the solution of at least three steps:

The first step is to investigate promising methodological directions (variants) of non-invasive signals suitable for implementation of intelligent diagnostics: Functional-Spectral Diagnostics (FSD-diagnostics); preliminary diagnostics based on the assessment of basic parameters of functional state (such as electrocardiogram (ECG), arterial blood pressure, heart rate (pulse, HR), temperature distribution in the local skin area, carbon dioxide volume in exhaled air CO_2, arterial blood oxygen saturation SpO_2); bioimpedance analysis; Zakharyin-Ged zone diagnostics; Nakatani method diagnostics; frequency-resonance diagnostics.

The second step is to formulate general requirements to the diagnostic decision support system.

The Decision Support System for Diagnostics is designed for doctors and patients.

When developing a decision support system for human diagnostics, we believe it is reasonable to shift the focus of attention not only to the development of an intelligent system, but also to the issues of integration and compatibility of different solutions and approaches, and their subsequent joint development.

The third step is to justify the choice of technology for the development of an intelligent system.

Following the formulation of methodological approach, the technological requirements of the system can be summarized as following. The proposed intelligent non-invasive diagnostics system should:

- be oriented to the design and development (improvement) of the system;
- provide the possibility of integration of heterogeneous data pipelines:
- process different types of data
- ensure standardization of data representations (forms of representation, information processing models);
- support modular system architecture that provides the possibility of adding new components and new data types, and their integration into the system;
- ensure integration with modern emerging technologies.
- ensure compatibility of interfacing with different systems, their docking.

The choice of technology to be developed should include the possibility of strengthening technological sovereignty. Since full technological dependence of countries on monopolistic corporations, which we are now witnessing, is a "path dependence" in the current situation. One of the solutions is the development of open-source projects and focusing on efficient exploitation of open-source libraries openly available for research and industrial use free of charge.

IV. Proposed Approach

The purpose of the system is to assist physicians in establishing a diagnosis. For this purpose, it is necessary to establish in the system the principles of diagnosis formation, the structure of the diagnosis, as well as to provide a transparent mechanism of reasoning when making a diagnosis or when proposing several diagnostic hypotheses. Knowledge driven systems solve these problems.

To define a technical solution it is necessary to take into account the three key elements: (1) comprehensibility of the reasoning process, (2) heterogeneous data representation, and data and (3) interface standardisation.

Firstly, when designing decision support systems in medicine, it is necessary to take into account that for practical application these systems require explainability of the reasoning process with provision of reasoning on how the results obtained, as well as the ability to modify the knowledge used in the system in a timely manner. Fulfillment of such requirements is ensured by the ontological approach [10], on the basis of which knowledge bases (KB) in terms and structure familiar to specialists can be created.

Secondly, since different types of data are used as input data, it is necessary to integrate heterogeneous data pipelines, which is ensured by the ontological approach.

Thirdly, when building decision support systems, one of the problems is to provide a uniform description and interpretation of data, regardless of the place and time of their receipt in the overall system. One of the ways to solve this problem can be the introduction of ontological modeling technologies.

A number of other advantages of the ontological approach should be emphasized. Ontologies use logical formalisms, which makes them convenient for use in the development of complex systems. Ontologies are characterized by flexibility, which allows combining information from different sources and building new knowledge on its basis. The main purpose of creating any ontology is to model a certain subject area, in turn, this forms the core of the system, and other modules are easily integrated with this module. Also the knowledge stored in ontological form has a high potential for reuse. Each ontology contains some fragment of conceptual knowledge of the subject area and hence ontology systems are called knowledge-based systems.

A diagnosis ontology includes a structure for describing information, rules for interpreting it and applying it to diagnosis.

In view of the above, it is reasonable to use the ontological approach.

Logical-semantic systems are based on ontologies. Artificial intelligence systems based on logical-semantic knowledge processing work with conceptual apparatus. Logical-semantic systems work with knowledge representations in the form of ontologies, realized, for example, in the form of a knowledge graph, where concepts and other objects correspond to the nodes of the graph and relations between them — to the edges of the graph.

As applied to the tasks of non-invasive diagnostics logical-semantic systems will allow to build parallel hypotheses and form a diagnosis (more or less probable), which is an advantage of their use.

Alternative development tools include ontology editors Protégé, Ontolingua, OntoEdit and others. Recently, the number of publicly available ontology editors has been increased. However, the main problem is that ontology editors are considered in isolation from implementation technology.

The most effective technology for creating ontological systems is the domestic technology of complex life cycle support of semantically compatible intelligent computer systems of new generation (Open Semantic Technology of Intelligent Systems — OSTIS) [11], [12].

OSTIS technology is an open semantic technology for component-based design of hybrid and interoperable intelligent systems.

The OSTIS technology is based on the OSTIS standard, which is a standard of semantic computer systems that provides

- semantic compatibility of systems complying with this standard;
- methods of building such computer systems and their improvement in the process of operation;
- means of building and improving these systems, including language tools, libraries of standard technical solutions, as well as tools (means of synthesis and modification; means of analysis, verification, diagnosis, testing; means of [10].

It should be noted that one of the most important features of systems built on the basis of OSTIS technology is their platform independence. It has an orientation on the semantic representation of knowledge, which is completely abstracted from the peculiarities of the technical realization of intelligent systems.

It is important to consider the possibility to use already developed ontologies on medicine (on various resources), pre-transforming them into OSTIS format with further processing of the expert (adaptation to specific tasks). For example, it is possible to use "Knowledge bases of medical terminology and observations", ontology for representing knowledge about diagnostics of diseases and syndromes realized on the cloud platform IACPaaS [13] and others.

V. Description of the system operation principle

Problem Statement. We aim design the architecture of medical decision support system for patient diagnosis based on non-invasive diagnostics.

Tools for realization — OSTIS technology.

The architecture of the OSTIS system for medical applications consists of the following components:

- OSTIS Knowledge Bases — Can describe any type of knowledge, while being easily augmented with new types of knowledge. Can include such ontologies as disease model, etc.
- OSTIS problem solver is based on multi-agent approach and allows easy integration and combination of any problem-solving models.
- The OSTIS interface is a subsystem with its own knowledge-base and problem solver (separate knowledge base and problem solver).

The intelligent diagnostic decision support system will include the following components:

- Block of input data collection and storage
- Data processing unit

- Block of diagnostic conclusions formation (mechanism of logical conclusion, mechanism of "reasoning" when making a diagnosis, allowing to get an idea of what information was the basis for the diagnosis).
- Block of consulting doctors and patient — issuing a response to the end user of the system about the results of diagnostics, recommendation on further actions.

VI. Non-Invasive Diagnostic Methods

The term "non-invasive" can be translated as "without disturbing the skin".

The following criteria are taken into account when for choosing a non-invasive diagnostic method:

- simplicity of the procedure;
- safety and painlessness for the person;
- examination time;
- the number of features (e. g., markers);
- the number of body systems covered.

Intelligent diagnostics of the human condition is focused on a set of data and considers the human body as a whole.

Non-invasive human diagnostics can be carried out in several directions (variants) of functional diagnostics.

1) Functional spectral-dynamic diagnostics (FSD-diagnostics) for early diagnosis of diseases [14].
2) Bioimpedance analysis [15].
3) Preliminary diagnosis based on the evaluation of basic parameters of the functional state, such as electrocardiogram (ECG), arterial blood pressure, heart rate (pulse, HR), temperature distribution in the local skin area, volume of carbon dioxide in exhaled air CO_2, arterial blood oxygen saturation SpO_2. These measurements can be carried out, for example, using the device "Patient Monitor" [2], [16].
4) Diagnosis by Zakharyin-Ged zones [17].
5) Diagnosis by the Nakatani method [18].
6) Frequency-resonance diagnostics (bioresonance).

The choice of features mentioned above is justified by their prevalence and technological efficiency.

For each identified feature the principle of passivity of the main mode of diagnostics (without influence on the organism) or the principle of activity of the mode of diagnostics (influence on the organism is present) becomes an important indicator.

The passive indicators include: FSD-diagnostics; Zakharyin-Ged zone diagnostics; measurement of electrocardiogram (ECG), arterial blood pressure, heart rate (pulse, HR), temperature distribution in the local area of the skin, volume of carbon dioxide in exhaled air CO_2, arterial blood oxygen saturation SpO_2.

The active indicators include: bioimpedance analysis, Nakatani method diagnostics, frequency-resonance diagnostics.

Let's consider the in great detail non-invasive diagnostic methods mentioned above.

A. FSD-diagnostics

The most effective diagnostic technology is the technology of functional spectral-dynamic diagnostics (FSD-diagnostics) applied to solve the tasks of health dynamics monitoring [14]. This is due to the fact that FSD-diagnostics is effective with respect to common infectious and non-infectious diseases, including latent (hidden) stages and actual risks of their development. FSD provides a priori sufficiency (due to markers).

The FSD diagnostic technology is focused on the detection of disease risks (that is often used during early diagnosis). The sensor is a metal electrode that records a wave electromagnetic signal in the sound range.

The core principle of the spectral-dynamic method is to analyze the electrical oscillations of the body field in the frequency range from 20 hertz to 11 kilohertz with an amplitude of 1 millivolt.

FSD diagnosis involves the following operations [19]:

- tool: a sensor (metal electrode) is applied to the patient's skin surface for 35 sec;
- data collection: recording of the body wave signal in the frequency range from 20 Hz to 11 kHz (EMF audio range) is performed;
- signal processing: spectral analysis of the signal based on Dobeshi wavelet transform 3;
- detection problem: recognizing the presence of spectral correspondences with similar spectra of electronic copies of reference diagnostic markers;
- spectral correspondence (similarity of the marker with the corresponding part of the patient's spectrum) expressed in percent is the main indicator for the physician who issues a diagnostic report;
- the diagnostic report contains indications of risks or presence of infectious and non-infectious diseases.

Distinctive features of FSD-diagnostics from existing diagnostic technologies are: the principle of pattern recognition instead of the principle of parameter measurement; the principle of passivity of the main mode of diagnostics (without impact on the body); the possibility of automation of nosological diagnostics (recognition of the disease itself or nosological risk). The examination is performed in less than one minute, and its diagnostic analysis can take up to two hours.

B. Bioimpedance analysis (bioimpedanceometry)

Bioimpedance analysis (from "biological" and "impedance" — complex electrical resistance, "bioimpedance" — electrical resistance of biological tissues) — analysis of the amount of fat and fluid in the body, muscle and bone mass and metabolism, a method of rapid diagnosis of human body composition by measuring the electrical resistance between different points on the human skin.

In bioimpedance analysis, the active and reactive resistances of the human body and/or its segments at different frequencies are measured. Based on these measurements, the body composition characteristics such as fat, cellular and skeletal muscle mass, body water volume and distribution are calculated. To conduct bioimpedanceometry, a device called a bioimpedance meter is used, sensor include two pairs of electrodes in the chain "arm–torso–leg" with the use of probing sinusoidal current of constant frequency and low power (no more than 500-800 μA).

The main parameters evaluated by this method are the amount of fluid in the body, body mass index, basic metabolic rate, bone and fat mass, level of physical development and others, as well as their reference values depending on sex and age.

Bioimpedance analysis is used in the medical practice by doctors of different specialties: nutritionists, endocrinologists, doctors of other directions. The technique provides the doctor with a large amount of valuable information, indicates the need for laboratory and functional studies, and helps in determining treatment tactics [13].

There are several ways of measuring bioimpedance, one of them involves performing the following operations:

- the doctor enters data such as age, sex, weight and height, waist circumference, hip circumference, and wrist circumference into a computer program;
- the person is laid down, special sensors — electrodes are connected to his wrists and ankles, through which a weak alternating current of low power is applied;
- results are analyzed by a computer program and given in the form of convenient screen forms with comments.

Measurements are carried out within less than one minute.

C. Preliminary Diagnosis Based on the Assessment of Basic Parameters of the Functional State

The main parameters evaluated by this method are: electrocardiogram (ECG), arterial blood pressure, heart rate (pulse, HR), temperature distribution in the local area of the skin, volume of carbon dioxide in exhaled air CO_2, arterial blood oxygen saturation SpO_2.

Electrocardiography (ECG) is a ubiquitous method of studying heart function based on a graphic representation of the heart's electrical impulses. The intensity of heart muscle contractions is measured and converted into a graphic image (on a tape in the form of teeth). The results determine the absence or presence of abnormalities in heart function. Curve records the heart biocurrents.

Arterial blood pressure is the pressure of blood on the wall of the artery. The value of blood pressure is denoted by two numerical values. The figures 120/80 millimeters of mercury column for the brachial artery are taken as the norm. Systolic (upper) blood pressure — the level of blood pressure on the arterial wall at the moment of maximum heart contraction (the norm is 100-140 mm Hg). Diastolic (lower) blood pressure — the level of blood pressure on the arterial wall at the moment of maximum relaxation of the heart (normal — 60–90 mmHg). Values of normal BP depend on inheritance and age.

Heart rate (pulse rate, HR) is a physical quantity obtained by measuring the number of cardiac systoles per unit of time, the norm is 60-90 beats per minute. Deviations from the normal regular sinus rhythm are considered a heart rhythm disorder. HR depends on age, sex and external factors.

Temperature distribution in the local area of the skin is a comprehensive indicator of the thermal state of the human body. The body temperature of a person during the day varies within small limits, remaining in the range of approximately 35.5°C to 37.2°C. Changes in body temperature may indicate the presence of an inflammatory process in the body. The patterns of change in skin temperature often provide important diagnostic information about a person's condition.

Volume of carbon dioxide in exhaled air CO_2 , is a physiological stimulant of respiration: affects the cerebral cortex and stimulates the respiratory center. The norm is up to 4%.

Arterial blood oxygen saturation SpO_2 is the percentage of oxygenated hemoglobin in the blood (the amount of oxygen in the blood). This is an important indicator of the state of the human respiratory system. The norm is 95-100 percent.

Respiratory rate (RR) is the number of respiratory movements (inhalation-exhalation cycles) per unit of time (usually a minute). The normal respiratory rate (RR) is 16–20 per minute. Respiratory rate depends on the position of the body, physical activity.

These physiological parameters of a person can be measured using various devices, including the "Patient Monitor" (Fig. 1) [2].

Analysis of the measured parameters will allow the doctor to objectively assess the physiological state of the patient [2], [16].

In non-invasive measurements of basic human physiological parameters, the output signals of sensors have different physical nature and, accordingly, the types of data representation are heterogeneous, which, in turn, demonstrates the necessitates the use of ontological approach.

D. Detection of Organ Pathology by Zakhar'ina-Geda Zones

Zakhar'ina-Geda zones are certain skin areas in which reflected pains, as well as pain and temperature hyperesthesia often appear when internal organs are diseased.

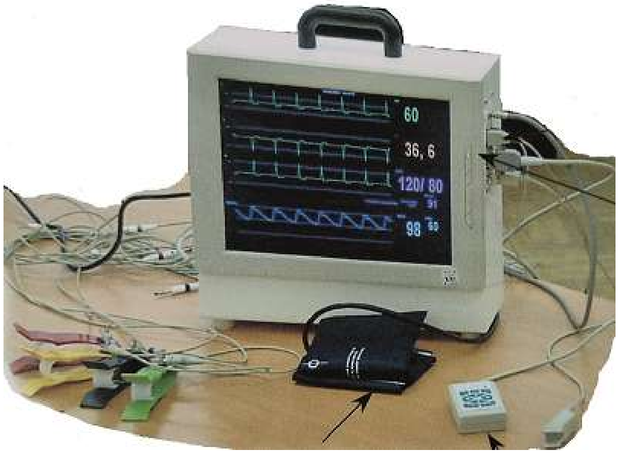


Figure 1. External view of the "Patient Monitor" device

In order to assess the condition of the patient's organs and identify diseased organs, we propose to use Zakharyin-Ged zones.

Initial data is the digitized thermal image of the patient taken in the infrared spectrum (using a thermal camera or heat sensors). It is necessary to compare the obtained image with reference maps of pathologic thermal zones allocation.

The method of detecting diseased organs of a patient involves performing the following operations:

- taking an image of the patient in the infrared radiation range using a thermal camera (obtaining a thermal portrait);
- digitizing the image by comparison and functional transformations;
- comparing the digitized thermal portrait with reference maps of Zakharyin-Ged, which are stored in the computer memory;
- allocation of pathological thermal zones based on comparison of the thermal portrait with reference maps;
- identification of the diseased organ by pathologic zones and output of information on the monitor to the doctor.

Measurements made on Zakhar'ina-Ged zones will allow to give a preliminary assessment of diseases of internal organs [17]. These measurements can be carried out with the help of the device "Patient Monitor".

The use of Zakharyin-Ged zones has been technologically developed in dynamic segmental diagnostics (including Nakatani method diagnostics).

E. Nakatani method diagnostics (riodoraku diagnostics)

Diagnosis using the Nakatani method consists of the studying the segmental cutaneous sympathetic reflex activity and is performed by measuring electrical conductivity values at representative points.

Japanese physician I. Nakatani (Nakatani) developed a method of electropuncture diagnostics of the functional state of meridians based on the measurement of electrocutaneous resistance (ECS) in representative (representative) acupuncture points. By measuring the ECS with an electrical detector in patients with inflammatory kidney disease, Nakatani found points with increased electrical conductivity and called them the electropermeable points.

The lines drawn through the electrically permeable zones are called "riodoraku" (from Japanese: line of good electrical conductivity, where "rio" — good, "de" — (electro) conductivity, "raku" — line). The main purpose of testing cutaneous sympathetic reflexes is to assess the activity of classical Chinese meridians for the subsequent prescription of acupuncture. Nakatani justified the use of galvanic current of 12 V and current strength of 200 μA (with closed electrodes) for diagnostic purposes. In Nakatani's method, current is applied to a point on a 1 cm^2 area of skin.

According to the method, the analysis is focused mainly on the ratio of these indices among themselves rather than using absolute values of current intensity (or ECS). For convenientt use of the method, a "standard riodaraku map" has been developed, where the ratios of the ratio of the current strength indices on skin projections of different "riodaraku", characteristic of healthy people, are graphically laid down. A scale is also used to interpret the indices of cutaneous sympathetic reflex activity in the area of representative zones of each riodoraku [20].

Electropuncture reflexodiagnostics according to Nakatani belongs to the methods of functional research. Through the assessment of the state of acupuncture meridians obtained by measuring the electrical conductivity of a set of representative points, it is possible to determine the functional state of individual internal organs and body systems. According to the Nakatani method, any changes in internal organs are reflected in the electrical characteristics of the skin. Therefore, the parameters of electropuncture measurements can be sensitive indicators, signaling systemic and nosological risks or the development of a pathological process.

The basic principle of this method can be formulated as "treat the person, not the individual disease". Nakatani method testing is widespread in many countries and is even considered mandatory during medical examination in Japan [18].

Electropuncture diagnosis is an integral part of clinical reflexology. The general order of investigations includes several main stages:

- gathering information about the patient;
- examination (measurements) (sensor-electrode);

- analysis and evaluation of measurement results;
- drawing up a conclusion.

The Nakatani method is widely used by reflexologists, mainly to assess the state of the meridian system and subsequent planning of acupuncture [18], [20].

F. Frequency-Resonance Diagnostics (Bioresonance)

The generation of wave diagnostics was created in 1978 by H. Schimmel and was called frequency-resonance diagnostics (bioresonance).

Bioresonance diagnostics is one of the methods of body research, which allows to carry out a complete examination of internal organs and systems in real time, to detect functional disorders at an early preclinical stage, to identify a weak or affected organ, and to determine the pathological process.

The core principle of the method can be formulated as "like cure by like".

Any organism emits electromagnetic vibrations. Cells and organs vibrate with a certain frequency. If we get sick, the vibrations of the affected organ change.

The frequency resonance method of diagnosis is based on the principle of frequency resonance.

This diagnosis involves the following operations:

1) electrodes are "attached" to the patient, which will read the measurements of electrical potentials at the points of skin projection of organs or systems of the body. Diagnostics is carried out on the points: head, hands and feet;
2) measurements of skin resistance by alternating current at various frequencies are started under the action of a very weak electromagnetic current;
3) the doctor, examining biologically active points, sends to each different frequency requests in expectation of resonance. Depending whatever the signal has been , received or not, the doctor finds out if a certain organ or system of the tested patient has a specific set of frequencies characterizing a specifically defined disease;
4) a diagnostic conclusion is formed.

Frequency resonance diagnostics is carried out within 2-2.5 hours.

VII. Non-Invasive Methods of Diagnostics and Periods of Diseases

Several periods (stages) are distinguished in the development of the disease. The most effective, in our opinion, non-invasive methods of diagnostics are proposed for a particular stage of the disease (Table I). It should be noted that diagnosis according to the Nakatani method is not in the periods of the disease, but in the plane of the state of the organism.

Let us emphasize the following periods of the disease:

- prenosologic period (risks) is — the period from the onset of gynesis risk (moment of gestation) to the onset of pathogenesis;
- latent period is the period from the onset of pathogenesis to the appearance of the first clinical signs of the disease.
- prodromal period is — a period of time from the first signs of the disease to the full manifestation of its symptoms (manifestation);
- manifest period (period of pronounced manifestations) has- specific symptoms of the disease are pronounced.

Table I
Disease stages and non-invasive diagnostic methods

Periods of disease	Non-invasive diagnostic methods				
	FSD	Bio-impe-dance	key para-meters	Zakha-ryin-Ged zones	Fre-quency reso-nance
Pre-nosologic period (risks)	+				+
Latent period	+			+	+
Prodromal period	+		+	+	+
Manifest period (period of pronounced manifestations)	+	+	+	+	+

Design and implementation of a diagnostic decision support system is a labor-intensive and time-consuming process. It is expedient to formulate priorities and decompose the tasks.

Analyzing the data shown in the table, FSD-diagnostics allows to identify the risks of diseases and has maximum informativeness and therefore it can form the first stage of the system implementation.

Let us determine the ontologies of the subject level (by diseases) within the framework of this problem.

VIII. Conclusion

Ancient Chinese wisdom says: "If there are no errors in diagnosis, there can be no errors in treatment". The use of intelligent non-invasive diagnostics will improve the quality of preventive medical care for the population.

In the paper, the justification and overall proposed architecture of intellectual non-invasive diagnostics system is discussed in great details.

Several directions (variants) of non-invasive diagnostics are considered: Functional-spectral diagnostics (FSD-diagnostics); preliminary diagnostics based on the assessment of basic parameters of the functional state (such as ECG, arterial blood pressure, HR, temperature distribution in the local area of the skin cover, volume of carbon dioxide in exhaled air CO_2, arterial blood oxygen saturation SpO_2); bioimpedance analysis; Zakharyin-Ged zone diagnostics; Nakatani method diagnostics; frequency-resonance diagnostics.

In non-invasive diagnostics, sensor output signals have different physical nature and correspondingly different types of data representation - and the question arises how to process them. This, in turn, necessitates the use of ontological approach. It is proposed to use the domestic technology of complex life cycle support of semantically compatible intelligent computer systems of new generation (Open Semantic Technology of Intelligent Systems). Logical-semantic approach in diagnostic tasks will allow to carry out differential diagnostics (to put forward several diagnostic hypotheses).

The technological basis for the creation of an intelligent diagnostic system is the technology of OSTIS.

The project of intellectual non-invasive diagnostics is justified by the presence of technological basis of OSTIS, application of logical-semantic approach in diagnostic tasks, as well as different variants of non-invasive diagnostics.

In the future, one of the directions of development of this system is the development of a personal medical assistant for the patient, which is focused not only on early detection of the disease, but also on recommendations to the patient for possible additional examination.

The basis for the functioning of the personal medical assistant is a decision support system. Obtaining knowledge by the patient will contribute to the understanding of his condition and its possible causes.

Important aspects of the system functioning are accumulation of data and knowledge, their systematization and the possibility of system evolution.

References

[1] V. N. Rostovtsev, V. S. Ulashchik. New technology of physical medicine. Zdravookhranenie, 2005, №5, pp. 10-14

[2] M. I.Silkou, R. M.Raviako, N. G.Lipnitskaya Multifunctional device for non¬invasive measurement of functional condition of patients. 6–th International Seminar on science and computing. Moscow, 2003, pp. 488-493.

[3] V. L.Malykh. "Decision support systems in medicine". Software Systems: Theory and Applications, 2019, vol. 10 no. 2(41), pp. 155-184.

[4] Efimenko I. V., Khoroshevsky V. F. Intelligent decision support systems in medicine: retrospective review of the state of research and development and prospects.Open semantic technologies for designing intelligent systems, Iss. 1, pp. 251-260.

[5] Kobrinsky B. A. Artificial intelligence systems in medical practice: status and prospects. Bulletin of Roszdravnadzor. 2020. №3, pp. 37-43.

[6] Sutton R. T. et al. An overview of clinical decision support systems: benefits, risks, and strategies for success. NPJ digital medicine, 2020, Vol.3, №1, P. 17.

[7] Berseneva E. A., Mikhaylov D. Y. The experience of using intelligent diagnostic decision support systems in a multidisciplinary hospital, Ural Medical Journal, №05 (188) 2020, pp. 174-180.

[8] Shchekina E. N. Using a systematic approach to create decision support systems in medicine (literature review). VESTRUCTURE OF NEW MEDICAL TECHNOLOGIES, Vol. 11, №2 2017, RUSSIA, TULA, pp. 356-364.

[9] — Smart system of personal health monitoring. Medelectronics — 2020. Means of medical electronics and new medical technologies : collection of scientific articles XII International Scientific and Technical Conference, Minsk, December 10, 2020. BSUIR, Minsk, 2020, pp. 198-203.

[10] Aminu E. F. et al. A review on ontology development methodologies for developing ontological knowledge representation systems for various domains. International Journal of Information Engineering and Electronic Business(IJIEEB), 2020, Vol. 12, №2, pp. 28-39.

[11] Shunkevich, D. Ontological approach to the development of hybrid problem solvers for intelligent computer systems. Open Semantic Technologies for Intelligent Systems. Minsk, 2021, Iss. 5, pp. 63-74.

[12] Technology of complex life cycle support of semantically compatible intelligent computer systems of new generation. Minsk, Bestprint, 2023, 1064 P.

[13] — Medical diagnostics ontology for intelligent decision support systems. Design Ontology, 2018, Vol. 8, №1(27), pp. 58-73.

[14] V. Rostovtsev Intelligent Health Monitoring Systems. Open Semantic Technologies for Intelligent Systems. Minsk, 2023, Iss. 7, pp. 237-239.

[15] — Bioimpedanceometry as a method of assessing the component composition of the human body (literature review). Bulletin of St. Petersburg State University. Medicine, 2017, Vol. 12, №4. pp. 365-384.

[16] Lipnitskaya, N. G. Synthesis of the information-measuring devices with application of the evolutionary calculations: Cand. Sci. (Techn.) Dissertation: 05.13.05, Minsk, 2005, 115 P.

[17] — Method and device for detection of a diseased organ in a patient by Zakharyin-Ged zones : patent. BY 10905. Published 30.08.2008.

[18] Boytsov I. V. Dynamic segmental diagnostics. Manual for doctors, N.Novgorod, "Typography "Povolzhie", 2014, 460 P.

[19] V. N. Rostovtsev, A. K. Kalmanovich, A. O. Lukyanov Device for recording and compensation of spectral-dynamic processes : patent. BY 6321. Published 30.06.2010

[20] Boytsov I. V., Kononovich V. I. Meridian diagnostics: a modern approach to research. Military Medicine, 2024, №1, pp. 60-70.

ОСНОВАНИЯ ИНТЕЛЛЕКТУАЛЬНОЙ НЕИНВАЗИВНОЙ ДИАГНОСТИКИ

Липницкая Н. Г., Ростовцев В. Н.

В статье предложено обоснование направления работ по созданию системы интеллектуальной неинвазивной диагностики. В качестве ее основополагающих составляющих определены два аспекта: технологический базис для разработки и различные варианты неинвазивной диагностики.

Технологическим базисом для создания интеллектуальной диагностической системы является отечественная Открытая Семантическая Технология Интеллектуальных Систем (ОСТИС). Использование логико-семантического подхода в диагностических задачах позволит осуществлять дифференциальную диагностику (выдвигать несколько диагностических гипотез). Рассмотрено несколько направлений (вариантов) неинвазивной диагностики: функционально-спектральная диагностика (ФСД-диагностика), биоимпедансный анализ, предварительная диагностика на основе оценки основных параметров функционального состояния, диагностика по зонам Захарьина-Геда, диагностика по методу Накатани, частотно-резонансная диагностика.

Received 13.03.2024

Integration and Standardization in New Generation Intelligent Medical Systems Based on OSTIS Technology

Veronika Krischenovich, Daniil Salnikov, Vadim Zahariev
Belarusian State University of Informatics and Radioelectronics
Minsk, Belarus
{krish, d.salnikov, zahariev}@bsuir.by

***Abstract*—The article discusses the integration of international medical standards in Russia, Belarus, and Kazakhstan using semantic technologies. It describes OSTIS (Open Semantic Technology for Intelligent Systems) — an open project for creating common semantic technologies for the component design of intelligent systems. An example of the integration of various medical records standards in intelligent medical systems is given. The advantages of such integration are the improvement of the exchange of medical information, simplification of the diagnosis and treatment process, and the possibility of creating a unified medical space within the region.**

***Keywords*—ontology modeling, OSTIS, medical information systems, standardization of medical data, integration of medical systems, International Classification of Diseases (ICD), flexibility and adaptability in healthcare systems.**

I. Introduction

The possibilities of application of new modern technologies in human life are growing day by day. Implementation of artificial intelligence in medical systems plays an important role in medical data processing. The huge amount of data processed in modern medical systems can be quickly analyzed and processed only using the artificial intelligence (AI) element that is being implemented in modern computer technologies. Processing medical data using AI enables the identification of diseases and the patterns and trends of their development, helps to identify pathologies and risks of disease, as well as forecasting the spread of infections and predicting epidemics. This helps doctors and healthcare organizations in general to make informed decisions [1]–[3].

The implementation of AI in medical systems can solve both complex and simpler problems. The complex challenges include the implementation of AI in robotic surgery. The application of AI in robotic surgery could significantly accelerate the development of surgical robots to perform complex surgeries with a high level of precision, which in turn will reduce risks and shorten patient recovery time after surgery. More straightforward challenges include the personalization of medicine, namely taking into account individual patient characteristics: genetic data and biomarker analysis, when developing personalized treatment plans and systematizing patients' medical histories [4], [5].

Electronic medical histories of patients, as well as digital data from medical examinations, patient monitoring data from medical devices, are part of a unified medical decision support system called a medical information system (MIS). MIS is a key element of the medical system. MIS providing automation of document flow and accounting in medical institutions has an important role in modern medicine. And the introduction of AI in MIS allows to move these systems to a new level of development.

II. Problematics of modern medical information systems

MIS is an information system designed for processing, accumulation, storage and retrieval of a patient's electronic medical record. MIS can be classified depending on the area of activity of medical institutions. For example, MIS for hospital usually collects and processes information from all blocks of the information system, including operating and resuscitation units. They may include modules for managing patients, staff, equipment, as well as for monitoring the performance of medical procedures and treatment. And MIS for outpatient clinics, in turn, usually focus on automating processes such as making appointments, working with waiting lists, and maintaining patient registries. They may also include functions for managing doctors' schedules, tracking patients' medical histories, and sharing medical information between different specialists and departments [6].

Thus, it can be concluded that MISs vary depending on the specific needs of the healthcare facility, but all systems will perform functions such as completing a patient's medical record and tracking medical history.

The first thing patients encounter when moving from one health care facility to another is the need to refill out their data for their medical records. Although personal data remains unchanged, the medical history may be incomplete because the patient cannot access all the data about all the examinations they have undergone.

This can make diagnosis difficult and distort the overall picture. It is worth noting that the more systems in which a patient enters their personal data, the greater the risk of this data being leaked.

Protection of patients' personal data is an important element of MIS operation. When implementing an information system, the staff of a health care institution should make certain efforts, for example, they should follow the algorithm of working with the selected information system, enter information into the system using available templates and forms and consistently maintain electronic medical records.

In 2020, during a roundtable discussion at the BELTA press center, representatives of the Ministry of Health and practical medicine discussed the promising directions of Belarusian e-health. And even earlier, in March 2018, the Concept of e-health development of the Republic of Belarus for the period until 2022 was developed and approved. The purpose of which was to develop e-health and create a centralized health information system (CHIS) for the formation of a unified information archive of patients and exchange of medical data [7].

The activity of the CHIS aims to improve the availability and quality of health care by assisting in clinical decision-making, improving the quality and efficiency of management decisions based on statistical and analytical data.

CHIS consists of functional and support subsystems and other subsystems. Supporting subsystems include software and hardware complexes, information protection system and subsystems that ensure proper technical functioning and interdepartmental information interaction of the CHIS. One of the subsystems included in the CHIS is the electronic medical record of the patient and the clinical decision support system.

Thus, we can conclude that the electronic medical record is a MIS, a subsystem of the CHIS. And the main obstacle to the creation of CHIS is the lack of standards in the field of e-health and regulations for the exchange of electronic medical information

As recommended by the World Health Organization and the International Telecommunication Union, a national eHealth system requires the following components: standards and interoperability (component); an enabling environment (role); and standards that will ensure the holistic and accurate capture and exchange of health information across all health systems and services (functional purpose).There are also a number of principles that should be considered when developing a CHIS:

- utilization of cloud computing technologies;
- use of open source software;
- service-oriented architecture, microservices, modularity, possibility to create additional services through open interfaces;
- elimination of duplication of engineering and telecommunication infrastructure;
- Web client technology;
- ensuring information security and information protection;
- scalability;
- simple and user-friendly interfaces, ergonomic and intuitive to use;
- single entry and repeated use of primary information;
- Interoperability of MIS with CHIS.

Thus, there is a need for technology that meets all the requirements for the realization of CHIS and, in particular, MIS.

The results of the implementation of the Concept in eHealth is the creation of the following systems:

1) National registers:
    - State Register "Diabetes Mellitus";
    - State register of persons exposed to radiation as a result of the Chernobyl catastrophe and other radiation accidents;
    - Belarusian Kancer Register;
    - Republican register of HIV-infected patients;
    - Republican register "Tuberculosis".
2) Medical information systems:
    - AIS "Electronic Prescription". This system is designed to automate the process of prescription writing and control over its fulfillment.
    - RSTMC (Telemedicine). Telemedicine system allows doctors to counsel patients from a distance using video, audio or messenger chat.
    - IAS "Zdravookhranenie". This system is designed to automate the recording and analysis of medical information, including data on the health status of patients.
    - IAS "Drug Supply". This system is designed to automate the process of planning and control of centralized procurement of medicines for healthcare organizations.

All these systems are aimed at improving the quality and efficiency of medical care through the use of modern information technologies. They also promote standardization and centralization of medical information, which facilitates data exchange and collaboration between different specialists and medical institutions. In addition, these systems help to ensure the security and confidentiality of medical data.

There are a number of foreign analogs of MIS. Here are a few MIS popular in Russia:

1) ArchiMed+ is a versatile medical software that is suitable for private physicians, medical centers, dental offices, and chain clinics. ArchiMed+ is easily scalable, offers many integrations including third-party labs, labeling system, telemedicine and more.

2) Medesk — this cloud-based medical information system is used in more than 20 countries around the world and in 72 regions of Russia. It is suitable for healthcare institutions of any size, from private clinics to networks. electronic medical records, CRM, telemedicine and solutions for managers and physicians are available in Medesk.
3) KMIS is a complex MIS suitable for automation of any medical institutions, the feature of which is integration with Federal systems.
4) MEDMIS is a relatively young MIS that entered the market in 2017. In 4 years, MEDMIS has been used by more than 200 medical organizations. MEDMIS is constantly evolving and gaining momentum: updates are released once a week.
5) MedAngel is an MIS with the possibility of individual customization for the specifics of the clinic's work. There is only a boxed version of the program with open code. The system is modular, you can assemble a personalized kit.

The list represents the most popular systems on the market. Each system has its own advantages and features, and the choice depends on the needs of a particular healthcare facility.

An Intelligent Medical System (IMS) is an information system that uses artificial intelligence techniques and approaches, including semantic technologies, to process, store and analyze medical data. Such systems represent a key element in modern healthcare, providing effective management of patient information, improved decision-making in medical practice and personalized patient care.

With the rapid development of artificial intelligence systems, IMS are becoming increasingly in demand. Their ability to analyze and interpret large amounts of data allows to reduce the workload of medical personnel, improve diagnostic accuracy and optimize treatment processes.

However, problems arise due to the variety of formats for storing medical data in different countries. Norms and legislation governing the processing and protection of medical information may differ from country to country. For example, the Republic of Belarus, Russia and the Republic of Kazakhstan have different standards and requirements for storing and processing medical data. This can create difficulties in integrating and sharing information between systems, as well as increase the risk of breaches of patient data privacy and security. For successful implementation of intelligent medical systems, it is necessary to take these differences into account and develop appropriate mechanisms for data standardization and interoperability.

III. Proposed approach

At the heart of any knowledge-based IMS is a formalized knowledge base, the quality and volume of which directly affects the effectiveness of the system. For uniform representation and unambiguous interpretation of knowledge bases, a common set of all terms used in practice is required. Terminology should be generally accepted and understandable to medical specialists, i.e. it should be the result of ontological agreement in the field of medicine.

Ontological modeling of the subject domain includes specification, conceptualization and formalization. At the specification stage, a glossary of terms is built, including all terms important for the subject area and their descriptions. In the conceptualization stage, important objects of the subject domain are identified. In this stage, the hierarchy and relationships between the objects of the subject domain are also defined. In the formalization stage, meta-objects and relationships between meta-objects are created that correspond to the objects and relationships between the objects of the subject domain. As a result, an ontology of the subject domain will be obtained. At the actualization stage, object parameters and their domains are defined (parameter domain is the area of acceptable parameter values), as well as values, classes, subclasses and class instances. Parameters, parameter domains, parameter values, classes, subclasses and class instances are realized as metaobjects (a metaobject is some text representing a definition, concept or some other description.) of appropriate types. After the work at the actualization stage, the ontology is turned into a knowledge base. The result of ontology modeling is the ontology of the subject area. To turn the ontology into a knowledge base, it is necessary to actualize it.

In MIS, the problem of incompatibility of data formats is a significant challenge that can hinder effective information sharing and interoperability between different healthcare systems. OSTIS (Open Semantic Technology for Intelligent Systems) is proposed to address this problem. OSTIS provides innovative tools and approaches for creating semantically interoperable medical systems that can efficiently process and store data, regardless of its original format and structure.

One of the key features of OSTIS technology is its ability to unify the representation of different types of knowledge in a single database. This enables the creation of centralized information repositories where medical data can be organized and structured according to common semantic standards, while ensuring a high degree of compatibility and interoperability.

OSTIS technology is also highly flexible and modifiable, allowing the system to be customized to meet the specific requirements and standards of each country, including the Republic of Belarus, the Russian Federation and the Republic of Kazakhstan. As a result, medical information systems based on OSTIS technology can easily adapt to various regulatory and legislative requirements, which ensures their seamless integration into the existing healthcare infrastructure.

One of the key benefits of using OSTIS technology is the ability to automatically convert and compare data from different formats. This helps eliminate interoperability issues and ensures seamless information exchange between different medical systems and institutions, which ultimately improves system efficiency and the quality of care provided.

Thus, the use of OSTIS technology is an effective and promising approach to solving the problem of incompatibility of data formats in medical information systems. It ensures the creation of modern and innovative healthcare systems capable of adapting to a variety of requirements and changes in the medical field, which is a key factor for improving the quality and accessibility of medical care in different countries.

As an illustrative example, let us give the possibilities of using OSTIS formalization technology for two subject areas related to intelligent medical systems.

Formalizing the subject matter of a patient's medical record using OSTIS technology demonstrates significant convenience and efficiency in processing and storing medical data. The formalization process allows describing various aspects of a patient's health, including medical history, diagnoses, treatments and symptoms of diseases.

One of the features of formalization on OSTIS technology is the ability to create a single semantic model that integrates different aspects of medical information in a unified format. For example, information about gastrointestinal (GI) diseases can be organized in the form of semantic entities such as disease type, symptoms, diagnosis and treatment methods.

The advantage of using OSTIS technology is that information on GI diseases can be stored in a single semantic format, making it easier to access and share data between different medical systems and institutions. In addition, thanks to the semantic approach, information about diseases and their symptoms can be structured and categorized for easy analysis and processing.

Formalizing the subject matter of a patient's medical record using OSTIS technology demonstrates significant convenience and efficiency in processing and storing medical data. The formalization process allows describing various aspects of a patient's health, including medical history, diagnoses, treatments and symptoms of diseases.

One of the features of formalization on OSTIS technology is the ability to create a single semantic model that integrates different aspects of medical information in a unified format. For example, information about gastrointestinal (GI) diseases can be organized in the form of semantic entities such as disease type, symptoms, diagnosis and treatment methods.

The advantage of using OSTIS technology is that information on GI diseases can be stored in a single semantic format, making it easier to access and share data between different medical systems and institutions. In addition, thanks to the semantic approach, information about diseases and their symptoms can be structured and categorized for easy analysis and processing.

IV. The subject area and ontology of a patient's medical record

Medical records are kept according to the order approved by the Ministry of Health. In the Republic of Belarus, this is Order No. 710 of the Ministry of Health of the Republic of Belarus of August 30, 2007 "On Approval of the Forms of Primary Medical Documentation in Outpatient and Polyclinic Organizations". The card of a patient receiving medical care in outpatient conditions must comply with the form 025/-07 "Medical card of an outpatient". In the Russian Federation it is the Order of the Ministry of Health of the Russian Federation from December 15, 20144 № 834n "On approval of the unified forms of medical documentation used in medical organizations providing medical care in outpatient settings, and procedures for filling out the form № 025/u "Medical card of a patient receiving medical care in outpatient settings". In the Republic of Kazakhstan it is the Order of the Acting Minister of Health of the Republic of Kazakhstan from October 30, 2020 № KR DSM- 175/200 "On approval of forms of record documentation in the field of health care" form № 052/u "Medical card of an outpatient". There are similarities and differences. There is a common part: full name, sex, date of birth, place of work, blood group and Rh. There are also differences, for example, in the RF there is a section SNILS, in the RK IIN, and in the RB personal number on the passport [9]–[11].

Based on OSTIS technology, it is possible to effectively formalize various concepts from a patient's medical record, taking into account the peculiarities of regulatory documentation in different countries. Despite the differences in the formats and structure of medical documents in the Republic of Belarus, the Russian Federation and the Republic of Kazakhstan, OSTIS technology can be used to create a universal model for data storage and processing, ensuring their standardization and flexibility.

The advantage of OSTIS technology lies in its graph-based approach to information representation, which allows for quick changes to the system without changing its structure, just modifying the knowledge base. This approach makes the system flexible and adaptable to new requirements and changes in legislation, which is especially important in the medical field, where requirements and standards are constantly changing.

Fig. 1 shows a fragment of a patient medical record ontology formalized using SCG language and including such basic concepts as:

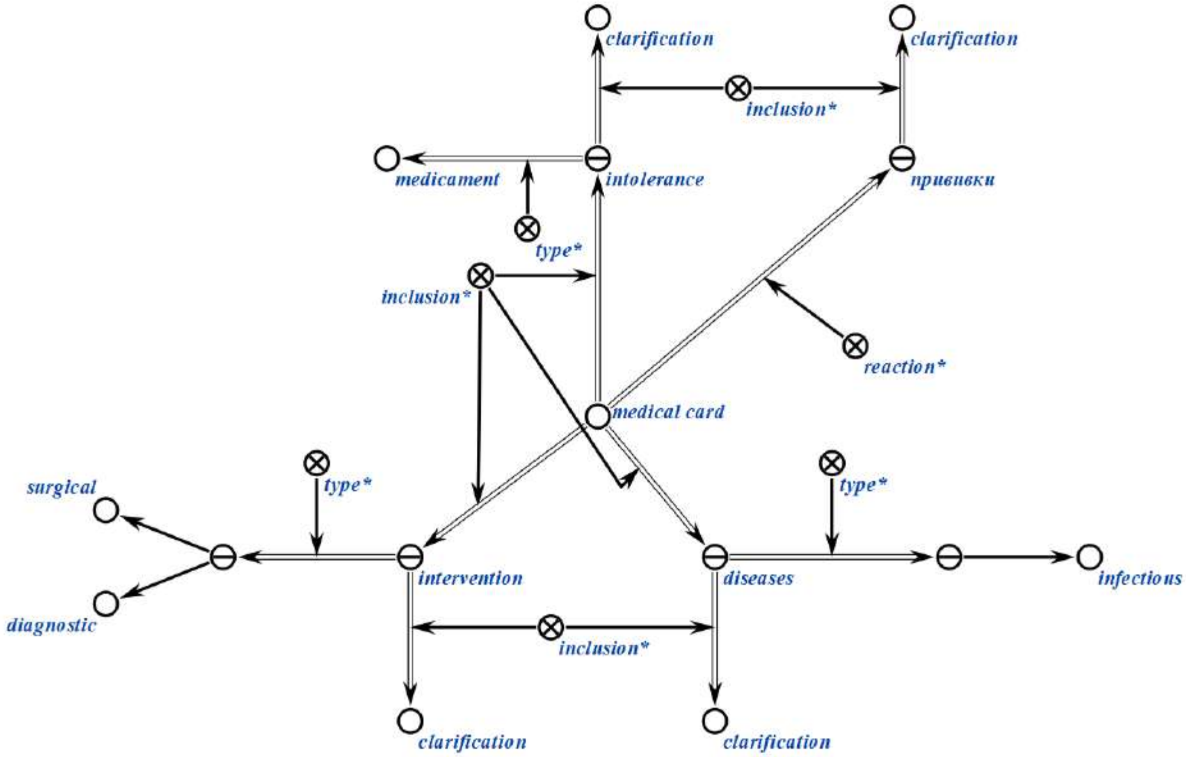


Figure 1. Fragment of medical record ontology

For example, patient information, including personal data and medical history, can be represented in graph structures where each node represents a different concept and the links between nodes reflect the relationships between the data. This approach makes it easy to add new data or modify existing data without having to redesign the entire system.

Using OSTIS technology, it is also possible to bring different formats and structures of data from medical documents to a common standard of presentation. For example, patient information from different countries may contain different fields and formats, such as SNILS in Russia, IIN in Kazakhstan, and personal passport number in Belarus. However, thanks to the flexibility and adaptability of OSTIS technology, these data can be standardized and combined into a common information model, providing a unified view and access to a patient's medical information regardless of where they live or where they are being treated.

In addition, with the help of OSTIS technology, meta-information can be added to the patient's medical data at the level of knowledge base, which allows its use by intelligent system agents in the process of processing and analysis. This makes it possible to create a more flexible and intelligent system of medical data processing than just adding comments in a classical medical information system. An example of a fragment of such data formalization is presented in Fig. 2.

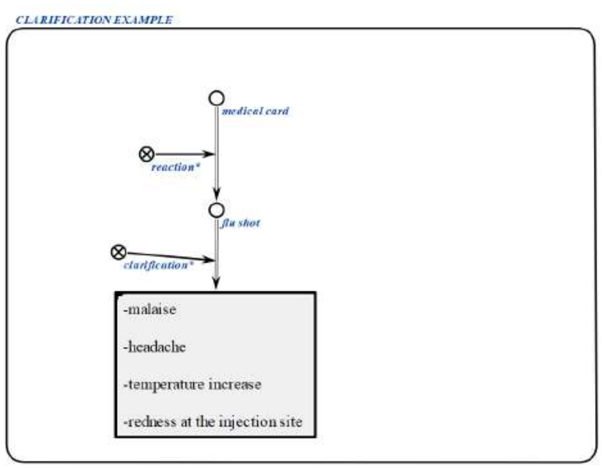


Figure 2. A fragment of the ontology of a medical record that allows you to store clarifying information

Thus, the use of OSTIS technology to formalize data from patients' medical records and information on GI diseases allows to create a flexible and adaptive system that can effectively take into account new requirements and changes in legislation, as well as standardize the presentation of information to simplify its analysis and processing.

V. The subject area and ontology of GI disease

Gastrointestinal (GI) diseases are one of the most common problems in medical practice worldwide. They cover a wide range of conditions, from functional disorders to serious pathologies such as peptic ulcers and cancer. According to the World Health Organization (WHO), GI diseases are the leading causes of death and disability worldwide.

GI disease statistics:

- according to the WHO, in 2020, GI diseases are the cause of death for more than 4 million people worldwide;
- according to studies conducted in different countries, GI diseases account for up to 25% of all reasons for visits to general practitioners;
- some of the most common GI diseases include peptic ulcer disease, gastric and duodenal ulcers, gastritis, colitis, irritable bowel syndrome (IBS), gallstones, pancreatitis, and GI cancer;

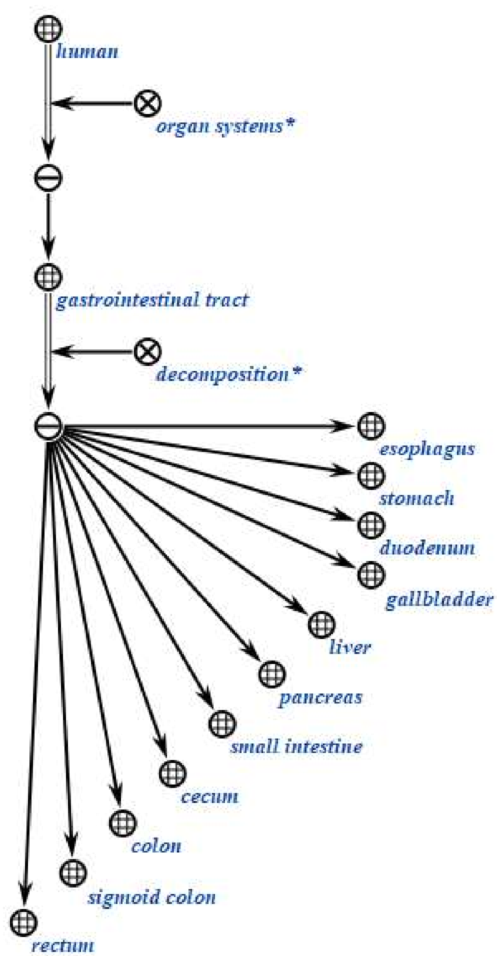


Figure 3. A fragment of the ontology of a medical record that allows you to store clarifying information

The International Classification of Diseases, 10th Revision (ICD-10) provides a coding system for diseases used in medical statistics and diagnosis. GI diseases are described in ICD-10 section K00-K93. This section includes a wide range of conditions, from dental problems to diseases of the liver, pancreas, and other GI organs. Diseases of this area include functional disorders, inflammatory processes, infections, tumors, and other pathologies specific to the GI tract. They can be manifested by various symptoms such as abdominal pain, diarrhea, constipation, nausea, vomiting and others. The definition and classification of GI diseases according to ICD-10 is important for statistical analysis, morbidity studies and health care planning.

Fig. 3 shows the formalization of the digestive organs domain using OSTIS technology. This formalization includes the development of an appropriate ontology structuring information about GI diseases according to the main sections of the International Classification of Diseases 10th Revision (ICD-10) [12].

The first section of the digestive organ ontology covers the anatomical structure and functions of organs including stomach, liver, pancreas, intestine and others. Each organ is presented as a separate entity described by its anatomical features and functions. The subject matter is further divided into various sections, including functional disorders, infections, tumors and other pathologies, in accordance with ICD-10. Each section contains the relevant classes of diseases and their associated medical conditions, symptoms and treatments.

In the context of the study of the subject area of digestive organs, special attention is paid to the stomach, considered on the example of gastritis in its usual and hyperacidic forms. Each disease corresponds to a reference marker set by the expert, which can be tissue or drug-specific. In addition, each disease has etiologic markers, which are multiple indicators that point to possible sources of the disease, such as bacteria, viruses, and other factors.

Organs in the digestive system can be in three states: disease state (more than 80% similarity), risk state (50% to 80% similarity), and non-risk state (healthy organ, less than 50% similarity). This approach allows the system to classify organs according to their current status based on analysis of user data.

The formalization of the ontology fragment and its corresponding knowledge base, presented in the figure, allows not only to treat diseases after their manifestation, but also to carry out the tasks of early diagnosis and prevention of the disease at early stages. This methodology allows integrating reference and etiological markers of diseases into the knowledge base, which provides the system with access to information for analyzing and processing medical indicators at a deeper level, which is discussed in the works of Rostovtsev V. N. [13]–[15].

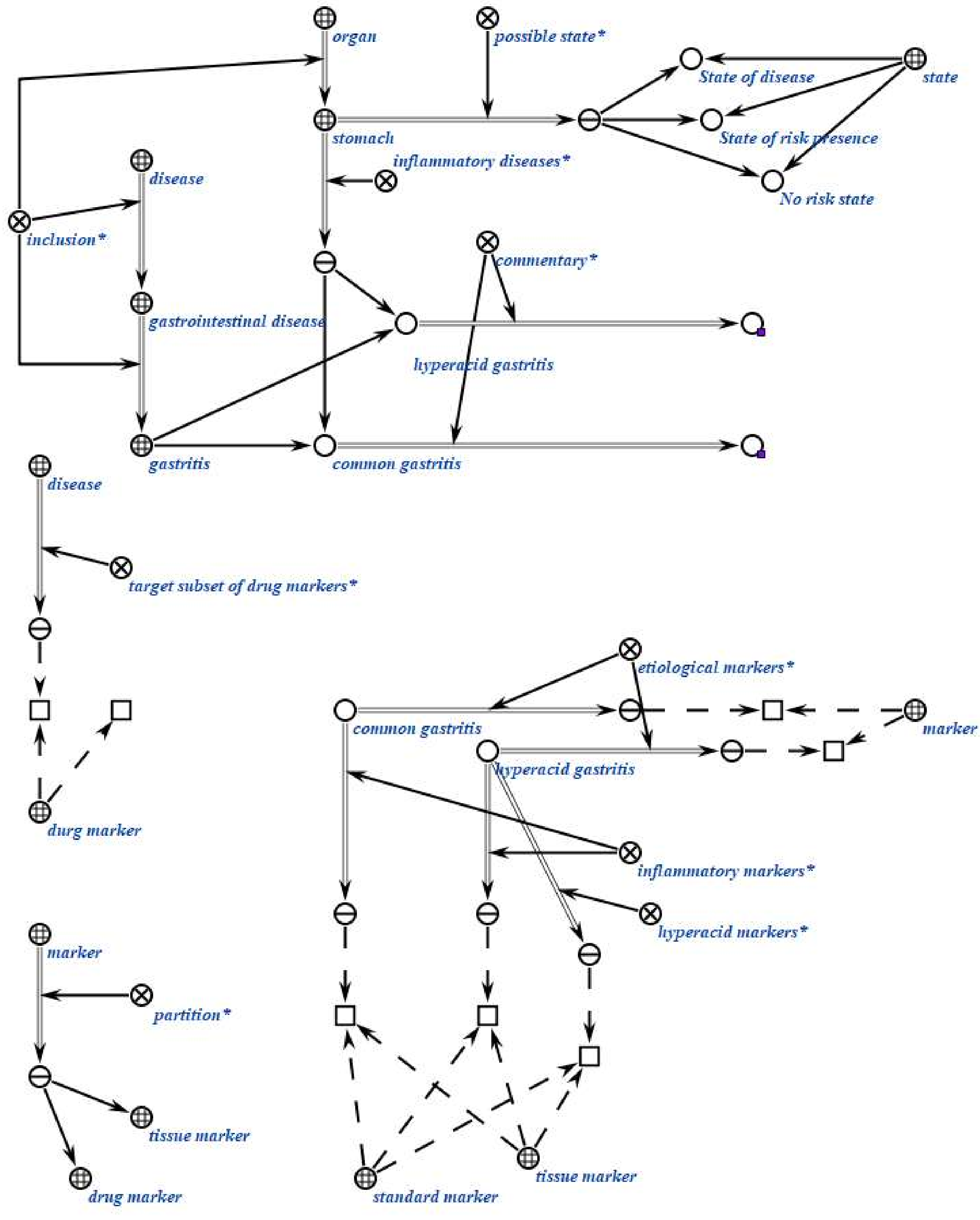


Figure 4. Fragment of medical record ontology

VI. Conclusion

The integration of Open Semantic Technology for Intelligent Systems (OSTIS) into medical information systems presents a promising solution to the challenge of data format incompatibility. OSTIS offers innovative tools and approaches for creating semantically compatible medical systems capable of efficiently processing and storing data regardless of their original format and structure.

One of the key features of OSTIS is its ability to unify various types of knowledge into a single database. This centralized approach allows for the organization and structuring of medical data according to unified semantic standards, ensuring high compatibility and interoperability.

Furthermore, the flexibility and adaptability of OSTIS enable the customization of systems to meet the specific requirements and standards of each country, including Belarus, Russia, and Kazakhstan. This adaptability facilitates seamless integration into existing healthcare infrastructures.

The automatic conversion and matching of data in different formats represent a significant advantage of OSTIS. This capability eliminates compatibility issues and facilitates smooth information exchange between various medical systems and institutions, ultimately enhancing system efficiency and the quality of healthcare delivery.

In summary, the application of OSTIS technology offers an effective and promising approach to addressing data format incompatibility in medical information systems. It fosters the creation of modern and innovative healthcare systems capable of adapting to diverse requirements and changes in the medical field, which is crucial for improving the quality and accessibility of healthcare in different countries.

References

[1] Zhan Y. et al. Investigating the role of Cybersecurity's perceived threats in the adoption of health information systems. *Heliyon*, 2024, Vol. 10, № 1.

[2] Yang P. et al. LMKG: A large-scale and multi-source medical knowledge graph for intelligent medicine applications. *Knowledge-Based Systems*, 2024, Vol. 284, P. 111323.

[3] Yakimov, D. A., Vygovskaya, N. V., Drozdov, I. V. Development of a Medical Information System with Data Storage and Intelligent Image Analysis. *Digital Transformation*, 2024, Vol. 30, № 1, pp. 71–80.

[4] Troqe B., Holmberg G., Lakemond N. Making decisions with AI in complex intelligent systems. *Research Handbook on Artificial Intelligence and Decision Making in Organizations*. Edward Elgar Publishing, 2024, pp. 160–178.

[5] Shawwa L. The use of telemedicine in medical education and patient care. *Cureus*, 2023, Vol. 15, №. 4.

[6] Chauhan P. et al Breaking Barriers for Accessible Health Programs: The Role of Telemedicine in a Global Healthcare Transformation. *Transformative Approaches to Patient Literacy and Healthcare Innovation*. IGI Global, 2024, pp. 283–307.

[7] Kontseptsiya razvitiya elektronnogo zdravookhraneniya Respubliki Belarus' na period do 2022 goda: prikaz Ministerstva zdravookhraneniya Respubliki Belarus' ot 20 marta 2018 g. [Concept for the development of electronic health care of the Republic of Belarus for the period until 2022: order of the Ministry of Health of the Republic of Belarus dated March 20, 2018]. Ministerstvo zdravookhraneniya Respubliki Belarus' [Ministry of Health of the Republic of Belarus], 2018, № 244

[8] Ob utverzhdenii form pervichnoi meditsinskoi dokumentatsii v ambulatorno-poliklinicheskikh organizatsiyakh: prikaz Ministerstva zdravookhraneniya Respubliki Belarus' ot 30 avgusta 2007 g. [On approval of forms of primary medical documentation in outpatient clinics: order of the Ministry of Health of the Republic of Belarus dated August 30, 2007.]. Ministerstvo zdravookhraneniya Respubliki Belarus' [Ministry of Health of the Republic of Belarus], 2007, № 710

[9] Ob utverzhdenii unifitsirovannykh form meditsinskoi dokumentatsii, ispol'zuemykh v meditsinskikh organizatsiyakh, okazyvayushchikh meditsinskuyu pomoshch' v ambulatornykh usloviyakh, i poryadkov po ikh zapolneniyu: prikaz Ministerstva zdravookhraneniya Rossiiskoi Federatsii ot 15 dekabrya 2014 g. [On approval of unified forms of medical documentation used in medical organizations providing medical care in outpatient settings, and procedures for filling them out: order of the Ministry of Health of the Russian Federation dated December 15, 2014.]. Ministerstvo zdravookhraneniya Respubliki Belarus' [Ministry of Health of the Republic of Belarus], 2014, № 834n

[10] Ob utverzhdenii form uchetnoi dokumentatsii v oblasti zdravookhraneniya: prikaz i.o. Ministra zdravookhraneniya Respubliki Kazakhstan ot 30 oktyabrya 2020 g. [On approval of forms of accounting documentation in the field of healthcare: order of acting. Minister of Health of the Republic of Kazakhstan dated October 30, 2020]. Ministerstvo yustitsii Respubliki Kazakhstan [Ministry of Justice of the Republic of Kazakhstan], 2020, № KR DSM-175/2020.

[11] O poryadke funktsionirovaniya i ispol'zovaniya tsentralizovannoi informatsionnoi sistemy zdravookhraneniya: postanovlenie Soveta Ministrov Respubliki Belarus', 13 maya 2021 g. [On the procedure for the functioning and use of a centralized health information system: Resolution of the Council of Ministers of the Republic of Belarus, May 13, 2021]. *Natsional'nyi pravovoi Internet-portal Respublsiki Belarus' [National legal Internet portal of the Republic of Belarus]*, 2021, № 267, 15.05.2021, 5/49050.

[12] International Statistical Classification of Diseases and Related Health Problems (ICD-10): Official Version. Tenth revised edition. Geneva: World Health Organization, 1992.

[13] Rostovtsev, V. Intelligent health monitoring systems. *Open semantic technologies for intelligent systems*, 2023, Iss. 7, pp. 237–240.

[14] Rostovtsev, V. N., Kobrinskii, B. A. Principles and possible ways of building an intelligent system of integral medicine. *Open semantic technologies for intelligent systems*, 2021, Iss. 5, pp. 225–228.

[15] Rostovtsev, V. N., Rodionova, O. S. Principles of ostis-system of automatic diagnosis design. *Open semantic technologies for intelligent systems*, 2018, pp.341–346.

ИНТЕГРАЦИЯ И СТАНДАРТИЗАЦИЯ В ИНТЕЛЛЕКТУАЛЬНЫХ МЕДИЦИНСКИХ СИСТЕМАХ НОВОГО ПОКОЛЕНИЯ НА ОСНОВЕ ТЕХНОЛОГИИ OSTIS

Крищенович В. А., Сальников Д. А., Захарьев В. А.

В статье рассматривается интеграция международных медицинских стандартов в России, Беларуси и Казахстане с применением семантических технологий. Предлагается подход к интеграции и стандартизации медицинских данных на основе применения технологии OSTIS. Приводится пример разработки фрагмента онтологии на основе различных стандартов медицинских карт в интеллектуальных медицинских системах. Преимуществами такой интеграции являются улучшение обмена медицинской информацией, упрощение процесса диагностики и лечения, а также возможность создания единого медицинского пространства в рамках региона.

Received 25.03.2024

Neuro-semantic Industrial Control

Dzmitry Ivaniuk
JSC "Savushkin Product"
Brest, Republic of Belarus
Email: Ivaniuk.Dzmitry@savushkin.com

***Abstract*—This article provides a review of the current situation of ontology use in industrial control with help of OSTIS Technology, examines in more detail the issues of combining together classical approach (standards in the field of Industry 4.0, such as ISA-88, ISA-95 and ISA 5.1) and intelligent technologies (neuro-control as part of Artificial Intelligence, AI) between robots (robotic systems) and humans.**

***Keywords*—neural network, ANN, reinforcement learning, standards, ontologies, Industry 4.0, Industry 5.0, OSTIS, ISA-88, ISA-95, ISA-5.1.**

I. Introduction

This work expands on the ideas discussed in [1]–[3] and includes descriptions of current issues and new versions of suitable tools for developing and using standards in industry in relation to modern techniques (neuro-control, semantic technologies).The connection with Industry 4.0 is taken into consideration, which is typically characterized by its complexity and the need for comprehensive knowledge of models and techniques to achieve an integrated solution [4].The arrival of Industry 4.0 has made it necessary for technical systems to consider the reliable and safe interaction of various intelligent systems with each other [5]. By using Artificial Intelligence decision making algorithms, process mapping in the new Industry 5.0 scenario can be enhanced, particularly by defining workflow checkpoints and identifying risks related to production and product quality.

The important role is played by all participants in the process: users — people (operators, masters, supervisors, etc.); devices — sensors and actuators (temperature sensors, pumps, valves, etc.); mechanized systems — conveyor systems, units; robotic systems — hinged robots, delta robots, manipulators; and software systems — SCADA, MES, ERP. Their interaction ensures the achievement of the goal, elimination and prevention of emergency situations. And important influence have both quantitative indicators (number of operators, devices, aggregates, control panels, etc.) and quality (quality of devices, qualification of operators, quality of software systems, etc.). Also important in management systems is the speed of decision-making — making changes quickly to meet plans. Each level is controlled by its own algorithms — often the element may look like a black box: input and output data are known, and the algorithm is hidden from the user. For instance, a neural network can so be described (Fig. 1).

Figure 1. Neural network as a black box

But if you want to have some more subtle configuration, you have to go to the documentation and use specific tools (for example, a change in the structure of the neural network may involve additional training or full new training on new reference data, which may also be unavailable and will require additional creation or/or adaptation of existing ones). The ontological description provides a general description that can be understood by all the participants. An integrated approach is needed to create a fundamentally new generation of intelligent computer systems and a corresponding technological complex.

Standardization describes different aspects of each developed human activity and includes a system of concepts (including terminology), a typology, and a model that describes how to apply appropriate methods and means, production sites, types and structures of project documents, accompanying activities, etc. With the existence of standards, we can solve one of the key problems related to any technology, particularly the rapidly developing computer information technology, **compatibility problem** [6]. Compatibility can be considered in many aspects, from the consistency of terminology in the interactions of process participants to the consistency of actions taken in the process of technology application. On the one hand, the problem with cohesion of digital twin models lies in the fact that a large number of disparate, unrelated and heterogeneous models are required. On the other hand, connecting digital twins in a single system [7] requires their interaction, and awaits conceptual unification of this interaction. It also require from Supervisory Control And Data Acquisition (SCADA) systems a higher level of integration, scalability and technological modernity [8].

Despite advances in information technology, most standards are now presented in the form of traditional linear documents or Web resources containing a series of static pages connected by hyperlinks. This approach to expressing standards has many serious drawbacks, and ultimately the overhead costs of maintaining and using standards actually outweigh the benefits of using them [9].

II. Problems and state of art

An analysis of the work has made it possible to formulate the most important and common problems related to the development and application of modern standards in various fields [9], [10]:

- Above all, the complexity of maintaining the standards themselves due to the duplication of information, especially the complexity of changing terminology.
- Duplicate information in the documentation describing the standard.
- Standards Internationalization Issues — translating a standard into multiple languages actually requires supporting and coordinating independent versions of the standard in different languages.
- As a result, inconsistencies in the format of different standards. As a result, automating the process of developing and applying standards is complicated.
- The inconvenience of using the standard, especially the complexity of finding the information you need. As a result, the complexity of studying standards.
- The complexity of automating the verification that an object or process complies with the requirements of a particular standard.
- etc.

These problems are mainly related to the presentation of standards. The most promising approach to solve these problems is the transformation of each specific standard into a knowledge base, which is based on a set of ontologies corresponding to this standard [6], [9]–[12]. This approach allows us to significantly automate the development processes of the standard and its application.

As an example, consider the ***ISA-88*** [13] standard (the basic standard for batch production). Although this standard is widely used by American and European companies and is actively implemented on the territory of the Republic of Belarus, it has a number of drawbacks. Essential ISA batch systems standards are:

- *ANSI/ISA-88.00.01-2010*, Batch Control — Part 1: Models and Terminology;
- *ISA-88.00.02-2001*, Batch Control — Part 2: Data Structures and Guidelines for Languages;
- *ANSI/ISA-TR88.00.02-2015*, Machine and Unit States: An implementation example of *ANSI/ISA-88.00.01*;
- *ISA-88.00.03-2003*, Batch Control — Part 3: General and Site Recipe Models and Representation;
- *ISA-TR88.0.03-1996*, Possible Recipe Procedure Presentation Formats;
- *ANSI/ISA-88.00.04-2006*, Batch Control — Part 4: Batch Production Records;
- *ISA-TR88.95.01-2008*, Using ISA-88 and ISA-95 Together;
- *IEC 61512-1*, The European version approved in 1997, based on the older version *ISA-88.01-1995*;
- *GOST R IEC 61512-1-2016* — Russian version of the standard, identical to *IEC 61512-1*.

Another standard often used in the context of Industry 4.0 is ***ISA-95*** [14]. ***ISA-95*** is an industry standard for describing high-level control systems. Its main purpose is to simplify the development of such systems, abstract from the hardware implementation and provide a single interface to interact with the ERP and MES layers. Consists of the following parts:

- *ANSI/ISA-95.00.01-2010*, Enterprise-Control System Integration — Part 1: Models and Terminology;
- *ANSI/ISA-95.00.02-2018*, Enterprise-Control System Integration — Part 2: Objects and Attributes for Enterprise-Control System Integration;
- *ANSI/ISA-95.00.03-2013*, Enterprise-Control System Integration — Part 3: Activity Models of Manufacturing Operations Management;
- *ANSI/ISA-95.00.04-2018*, Enterprise-Control System Integration — Part 4: Objects and Attributes for Manufacturing Operations Management Integration;
- *ANSI/ISA-95.00.05-2018*, Enterprise-Control System Integration — Part 5: Business-to-Manufacturing Transactions;
- *ANSI/ISA-95.00.06-2014*, Enterprise-Control System Integration — Part 6: Messaging Service Model;
- *ANSI/ISA-95.00.07-2017*, Enterprise-Control System Integration — Part 7: Alias Service Model;
- *ANSI/ISA-95.00.08-2020*, Enterprise-Control System Integration — Part 8: Information Exchange Profiles.

Models help define boundaries between business and control systems. They help answer questions about which functions can perform which tasks and what information must be exchanged between applications. The ISA5 standards development committee is often referred to as the ISA-5.1 standard among practitioners. However, the ISA5 committee, "Documentation of Measurement and Control Instruments and Systems," has a broader scope—namely to develop standards, recommended practices, and technical reports for documenting and illustrating measurement and control instruments and systems suitable for all industries. ISA5 standards consist of the following:

- *ANSI/ISA-5.1-2022*, Instrumentation Symbols and Identification;
- *ISA-5.9 working group*, Controller Algorithms and Performance;
- *ISA-5.4*, Instrument Loop Diagrams;

- *ISA-5.5-1985*, Graphic Symbols for Process Displays;
- *ISA-5.6*, Documentation for Control Software Applications;

This standard is useful when a reference to equipment is required in the chemical, petroleum, power generation, air conditioning, metal refining, and many other industries. The standard enables anyone with a reasonable level of plant knowledge to read flow charts to understand how to measure and control a process without having to go into the details of instrumentation or the knowledge of an instrumentation expert.

III. Neurocontrol

Neurocontrol is a relatively young field of research that became independent in 1988. However, research in this area began much earlier. One of the definitions the science of "cybernetics" considers this as a general theory control and interaction not only of machines, but also also biological beings. Neurocontrol tries to achieve this position through the construction of control systems (decision-making systems), which can be trained during operation, and thus improve its performance. In this case, such systems use parallel mechanisms of information processing, like the brain of living organisms [15].

For a long time the idea of building a perfect control system - a universal controller that would look like a «black box» from the outside was popular. It could be used to control any system, with connections to sensors, actuators, other controllers, and a special link to the «efficiency module» — a system that determines the management efficiency based on given criteria. The user of such a control system would only set the desired result, the further trained controller would manage himself, perhaps following a complex strategy of achieving the desired result in the future. It would also constantly adjust its management based on the management object's response to achieve maximum efficiency. An outline of such a system is given below (Fig. 2).

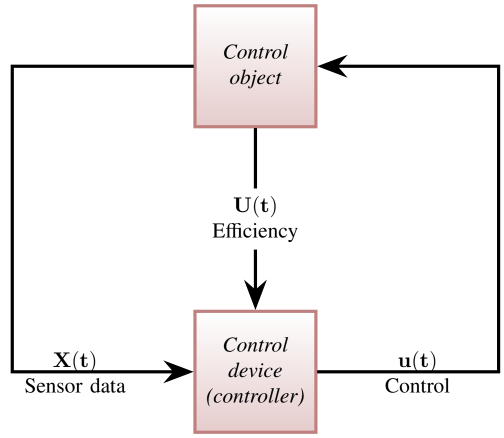


Figure 2. Reinforcement learning

IV. Developed Neuro-PID controller

The overall structure of the self-tuning neuro-PID controller is shown in Fig. 3, where the neural network (NN) outputs are proportional (K), integral (T_I), and differential (T_d) components [16].

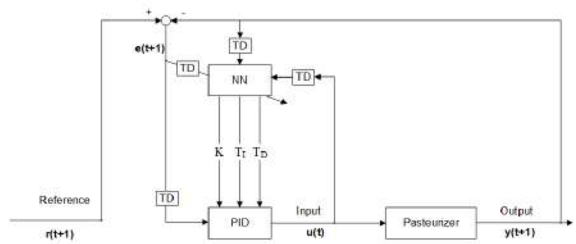


Figure 3. The developed neuro-PID controller, TD means delay operator

To control the Pasteurizer [16] PID is configured with a multilayer perceptron (MLP, neuro-PID adjuster) with the following structure: 20 input, 10 hidden and 3 output neural elements; the function of activating the hidden and output layers is sigmoid (Fig. 4).

The discrete-time **PID controller** can be described by equation 1 [16], where P, T_I and T_D are proportional factors, integral and differential constituents respectively, u_n determines the input of a control object at a time of $t = nT_0$ and e_n — an error between the desired output value of r_n and the real output of $e_n = r_n - y_n$. T_0 defines a unit time interval.

$$
\begin{aligned}
u_k &= u_{k-1} + \Delta u_k, \\
\Delta u_k &= q_0 e_k + q_1 e_{k-1} + q_2 e_{k-2}, \\
q_0 &= \mathbf{K}\left(1 + \frac{T_D}{T_0}\right), \\
q_1 &= -\mathbf{K}\left(1 + 2\frac{T_D}{T_0} - \frac{T_0}{T_I}\right), \\
q_2 &= \mathbf{K}\frac{T_D}{T_0}.
\end{aligned} \tag{1}
$$

Algorithm for operation of **neuro-PID adjuster** [16]:

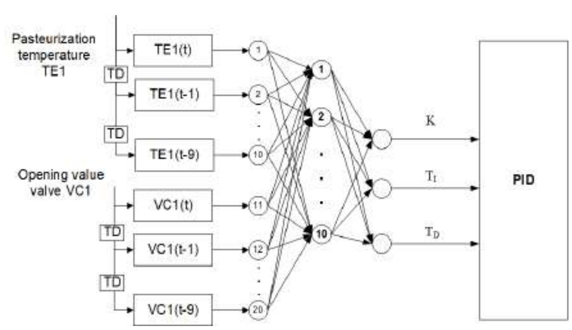


Figure 4. Neuro-adjuster PID

Figure 5. Start page

$$
\begin{aligned}
y_j &= F(\sum_i \omega_{ij} y_i - T_j),\\
\gamma_j &= y_j - t_j,\\
\gamma_i &= \sum_i \gamma_i F'(S_i)\omega_{ji},\\
\omega_{ij}(t+1) &= \omega_{ij}(t) - \alpha\gamma_j F'(S_j) y_i,\\
T_j(t+1) &= T_j(t) + \alpha\gamma_j F'(S_j),\\
E &= \frac{1}{2}\sum_{k=1}^{L}\sum_j (y_j{}^k - t_j{}^k)^2.
\end{aligned}
\tag{2}
$$

To use the back propagation algorithm, we must select the E function, which must be minimized. It will be the management error e_n at the time $t = nT_0$ — get $E_n = \frac{1}{2}e_n^2$. To accumulate errors, we store the data we have previously obtained — $E_{n-p}, ...E_{n-2}, E_{n-1}, E_n$, where p determines the number of previously saved images used for network learning (2).

V. Examples of system operation with natural language information display

For information to be clear and understandable to the reader, it must be presented in a consistent manner. The recipe authoring system interface allows the structure of domains and ontologies to be expressed in natural language. This process of converting an internal knowledge representation to an external knowledge representation is performed by a graphical interface component. On the main page general information (in 4 languages) is displayed, Fig. 5.

Fig. 6 shows resulting ontologies for standards (ISA-88, SCg).

VI. Integration of third-party solutions with a knowledge base

A standard system built on the basis of OSTIS Technology can be easily integrated with other systems in the workplace. To integrate ISA-88, ISA-95 and ISA-5.1 standards system with other systems running on JSC "Savushkin Product", a web-oriented approach is used —

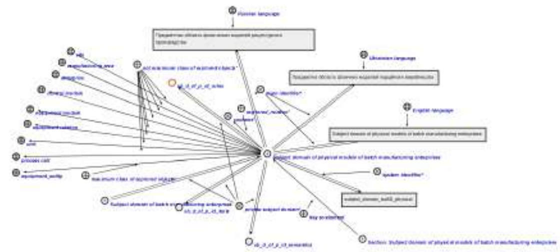


Figure 6. Ontologies for standard ISA-88

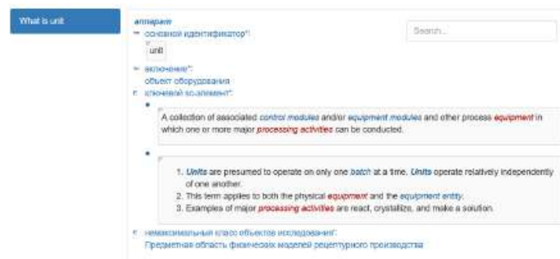


Figure 7. Unit

the ostis-system server is accessed with the use of the following queries:

```
http://industry.ostis.net?sys_id=unit
```

where "sys_id=unit" defines a term (the name of an entity) whose value we want to find out (in this example, in fact, the answer to the question "What is a "unit"?). This approach makes it relatively easy to add support of the knowledge base for current control systems projects, for this it is enough to indicate the names corresponding to the entities in the knowledge base within the control system. The answer is shown on Fig. 7.

Thus, an interactive intelligent help system for control systems projects is implemented, allowing employees to simultaneously work with the control system and ask questions to the system directly during the work.

Another example is the integrated help subsystem within corporate Add-In **EasyEPLANner** [17] for CAD EPLAN. It helps to describe technological objects (Tank, Boiler, etc.), operations, etc. according to the ISA-88 standard. Fig. 8 shows UML-model of EasyEPLANner objects to be described in OSTIS. The **PID controller** is at the lower level — the control module (highlighted by the green). It can be replaced by the development **neuro-PID controller**.

VII. Use in control systems

It is very important to correct and fast react on different events during process control, especially on critical accidents. But when we have complex distributed system it is rather complicated and in normal way require help of the human operator. It may leads to variety of

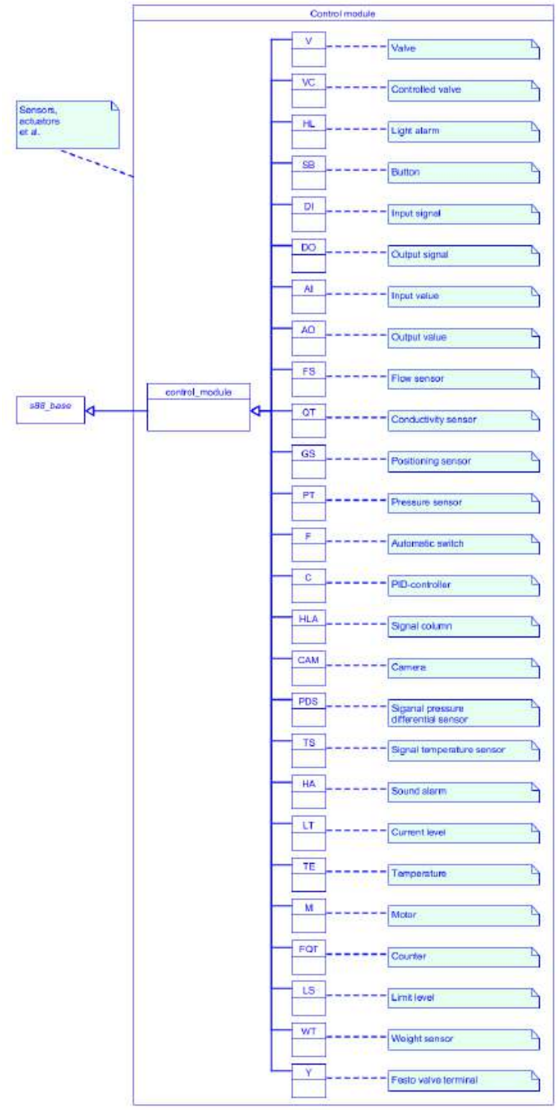


Figure 8. **EasyEPLANner** objects

problems. So usage OSTIS-based system can helps to solve as described on Fig. 9. Project #2 has a valve failure but the project does not know what to do. Then it makes a request to the OSTIS server, which already knows which projects also use this line (with this valve). The OSTIS server polls the rest of the projects (projects #1 and #2). Each project has information about which operations are currently active and gives an answer on what to do — pause the operation, do nothing, etc. After that OSTIS-server sends back to project #2 answer with result actions to be used. These are going in automatic way — no need of human operator.

VIII. Future development

Current project issues can be found on GitHub ([18], [19] and [20]). Main problems to be solved are:

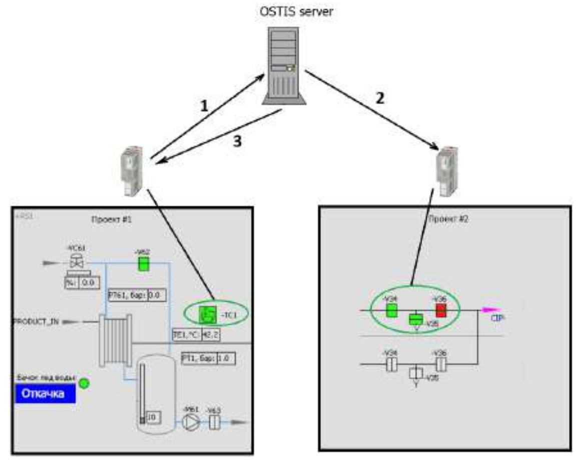


Figure 9. OSTIS in control systems

- Improving system performance and especially accelerating system response time to user requests. It is connect with productivity and overall user satisfaction.
- Continuous updating and refactoring ontological models (further formalization of missing concepts, fix typos and etc.);
- Enhancing PFC-visualization - not only displaying, but also editing diagrams. Adding rich navigating between PFC-diagram and according text representation;
- Further formulation of questions (typical) to the system from the user and their formalization at the level of the existing knowledge base;
- Adding more description of parts of real control projects based on the existing knowledge base.

The implementation of answers to complex questions is necessary to make easier the work of not only process operators, but also maintenance personnel — instrumentation engineers, mechanics, electricians, etc. Therefore, it is planned to implement the system's answer to the question of the following type — in what operations of which objects this actuator is used (for example, valve **"T1V1"**). This question is very important when a device failure occurs and it is necessary to determine the criticality of this situation. For analysis, it is necessary to compare the time of the accident and the history of operations. Since, for example, an accident of the mix-proof valve during the line washing operation and the active product dosing along the other line, should lead to a stop of these operations and stop the preparation of the batch in the corresponding unit. The operator must report this to the appropriate maintenance specialist to fix it. After the fault has been eliminated, the operator continues to perform operations. This is the correct events order, which is very important to avoid mixing of detergent and product. If the device malfunction occurred within the line, which is now inactive, then this situation has a low priority, does not lead to a stop in operations and can be eliminated later if the service personnel have free time.

IX. Conclusion

The paper considers an technique to automating the process of creating, developing and making use of standards primarily based on OSTIS Technology. Using the instance of the ISA-88, ISA-95 and ISA-5.1 standards used on the Savushkin Product enterprise, the structure of the knowledge base, the features of the problem solver and the user interface of the support system for these processes are considered. It is proven that the developed system can be easily integrated with other enterprise systems, being the basis for building an information service system for employees in the context of Industry 4.0. The approach proposed in the work allows us to provide not only the ability to automate the processes of creation, agreeing and development of standards, but also allows us to significantly increase the efficiency of the processes of applying the standard, both in manual and automatic way.

Acknowledgment

Author would like to thank to the research teams of the Departments of intelligent information technologies of the Belarusian State University of Informatics and Radioelectronics and the Brest State Technical University.

References

[1] N. Lutska, O. Pupena, A. Shyshak, V. Taberko, D. Ivaniuk, M. O. Nikita Zotov, and L. Vlasenko, “Ontological model of digital twin in manufacturing,” in Open semantic technologies for intelligent systems, ser. 5, V. Golenkov, Ed. BSUIR, Minsk, 2022, p. 310–335.

[2] V. V. Taberko, D. S. Ivaniuk, D. V. Shunkevich, and O. N. Pupena, “Principles for enhancing the development and use of standards within Industry 4.0,” in Open semantic technologies for intelligent systems, ser. 4, V. Golenkov, Ed. BSUIR, Minsk, 2020, pp. 167–174.

[3] V. A. Golovko, A. A. Kroshchanka, V. V. Golenkov, V. P. Ivashenko, M. V. Kovalev, V. V. Taberko, and D. S. Ivaniuk, “Integration of artificial neural networks and knowledge bases,” 2018.

[4] S. Gil, G. D. Zapata-Madrigal, and G. L. Giraldo-Gómez, “An ontological model to integrate and assist virtualization of automation systems for industry 4.0,” Smart and Sustainable Manufacturing Systems, vol. 5, p. 10, 09 2021.

[5] V. R. Sampath Kumar, A. Khamis, S. Fiorini, J. L. Carbonera, A. Olivares Alarcos, M. Habib, P. Goncalves, H. Li, and J. I. Olszewska, “Ontologies for industry 4.0,” The Knowledge Engineering Review, vol. 34, p. e17, 2019.

[6] V. Golenkov, N. Gulyakina, I. Davydenko, and D. Shunkevich, “Semanticheskie tekhnologii proektirovaniya intellektual’nyh sistem i semanticheskie associativnye komp’yutery [Semantic technologies of intelligent systems design and semantic associative computers],” Open semantic technologies for intelligent systems, pp. 42–50, 2019.

[7] “Digital Twins for Industrial Applications, an Industrial Internet Consortium White Paper,” https://www.iiconsortium.org/pdf/IIC_Digital_Twins_Industrial_Ap\ps_White_Paper_2020-02-18.pdf, 2018.

[8] M. Sverko, T. G. Grbac, and M. Mikuc, “Scada systems with focus on continuous manufacturing and steel industry: A survey on architectures, standards, challenges and industry 5.0,” IEEE Access, vol. 10, pp. 109 395–109 430, 2022.

[9] P. Serenkov, V. Solomaho, V. Nifagin, and A. Minova, “Koncepcija infrastruktury standartizacii kak bazy znanij na osnove ontologij [the concept of a standardization infrastructure as an ontology-based knowledge base],” Novosti. Standartizacija i sertifikacija. [News. Standardization and certification.], vol. 1, no. 5, pp. 25–29, 2004.

[10] V. Uglev, “Aktualizacija soderzhanija standartov proektirovanija slozhnyh tehnicheskih ob’ektov: ontologicheskij podhod [updating the content of design standards for complex technical objects: ontologic approach],” Ontologija proektirovanija. [Ontology of designing], vol. 1, no. 1, pp. 80–86, 2012.

[11] C. Dombayci, J. Farreres, H. Rodríguez, A. Espuña, and M. Graells, “Improving automation standards via semantic modelling: Application to ISA88,” ISA Transactions, vol. 67, 01 2017.

[12] M. Vegetti and G. Henning, “Isa-88 formalization. a step towards its integration with the isa-95 standard,” in 6th Workshop on Formal Ontologies meet Industry, vol. 1333, 02 2015.

[13] “ISA-88 standard,” https://www.isa.org/standards-and-publications/isa-standards/isa-standards-committees/isa88/, (accessed 2022, October).

[14] “ISA-95 standard,” https://www.isa.org/standards-and-publications/isa-standards/isa-standards-committees/isa95/, (accessed 2022, October).

[15] A. A. Uskov and A. V. Kuz’min, Intellektual’nye tekhnologii upravleniya. Iskusstvennye nejronnye seti i nechetkaya logika. Goryachaya liniya - Telekom, 2004.

[16] D. S. Ivaniuk, “Neuro-pid controller for a pasteurizer,” in XV International PhD Workshop OWD, 2013.

[17] “EasyEPLANner project on GitHub,” Available at: https://github.com/savushkin-r-d/EasyEPLANner/, (accessed 2022, Jun).

[18] “EasyServer-4.0 project on GitHub,” Available at: https://github.com/savushkin-r-d/EasyServer-4.0, (accessed 2024, Feb).

[19] “s88-ostis project on GitHub,” Available at: https://github.com/savushkin-r-d/s88-ostis, (accessed 2024, Feb).

[20] “isa-5.1-ostis project on GitHub,” Available at: https://github.com/savushkin-r-d/isa-5.1-ostis, (accessed 2024, Feb).

НЕЙРО-СЕМАНТИЧЕСКОЕ УПРАВЛЕНИЕ В ПРОМЫШЛЕННОСТИ

Иванюк Д. С.

В работе рассмотрен онтологический подход к пониманию, интеграции и развитию современных подходов к управлению (нейроуправление, семантические технологии, современные международные стандарты) с использованием Технологии OSTIS. Уточнены формальные трактовки основных понятий, используемых в стандартах, что позволяет упростить описание реальных задач. Также описаны варианты интеграции базы знаний в используемые программные средства разработки и сценарии её использования непосредственно в системах управления.

Received 01.04.2024

Applied Aspects of Using OSTIS Technology in Information Support of Digitalisation of Water Use Processes of Dairy Processing Enterprises

Vladimir N. Shtepa
Minsk, Belarus
Email: tppoless@gmail.com

Eduard N. Muslimov
Minsk, Belarus
Email: muslimoven@mail.ru

***Abstract*—The conceptual approaches of digitalisation of manufactures based on the e-Manufacturing ideology are evaluated, which are proposed to be used for modelling the processes of water use of milk processing enterprises; the organisational and technological processes of formation of pollutants in their wastewater are analysed. In IDEF0 methodology functional modelling of water use of such productions is carried out, that allowed to reveal complexity and multidirectionality of interrelations of parameters and to justify the use of OSTIS technology for tasks of formation of intellectual information and reference system. On the example of a biological pond as a node of wastewater treatment, an element of the proposed approach of practical implementation of OSTIS-solutions in the segments of digital modelling of the dairy industry and environmental management is implemented.**

***Keywords*—Digitalisation, water management, milk processing, information and reference system**

I. Introduction

The term Digitale Fabrik (digital factory) is used to describe production in the context of informatisation, but today the essence of Digitale Fabrik is more often expressed by the term e-Manufacturing. At the heart of the idea of e-Manufacturing is the continuous application of digital models in the design and operation of production systems. Not only the products themselves are modelled, but also production equipment, material flows, as well as production and logistics processes, taking ergonomics and human factors into account. The goal of e-Manufacturing is to achieve such a level of object and process modelling that the real manufacturing process starts only after all its elements have been studied and optimised with the help of models [1].

Digital manufacturing is one of the components of product lifecycle management (PLM) technology, its main task is to improve complex manufacturing processes. A set of digital manufacturing solutions belongs to the class of MRM-systems (Manufacturing Process Management) [2].

If in CAD/CAM/CAE-systems in most cases the application of a particular software tool is associated with obtaining a digital layout of the product and the distribution of roles is clearly deterministic by the content of the work performed (surface designer, layout designer, solid geometry designer, etc.), in MRM-systems this dependence is much more flexible [3]. This is due to the fact that technological processes in different industries can differ significantly, even enterprises of the same industry can have different technological processes. Such operations include the functioning of complexes ensuring the fulfilment of environmental requirements, including the removal (treatment, discharge) of industrial wastewater (WW) [4], [5].

This problem is especially acute for enterprises of milk processing industry.

II. Technological problems of formation and treatment of wastewater of milk processing enterprises

Contaminated wastewater of milk processing plants is a product formed after washing of equipment, technological piping system, transport tanks of different volumes, including road and railway tanks, flasks and other containers [6], [7]. Also the sources of pollutants formation include effluents after cleaning of production facilities, washing of panels and floors. The amount of polluted wastewater is 20% — 50% of the total volume of water use. Wastewaters of milk processing enterprises belong to the category of highly concentrated organic pollutants: they contain significant concentrations of organic pollutants (fat, protein, lactose), polluted also with inorganic compounds (including acids and alkalis) and synthetic surface active substances (detergents). Their composition and concentration of pollutants depend on the profile and productivity of enterprises [7], [8].

Analysis of literature sources [4]–[8] has shown that there is an intensive search for rational and highly efficient methods and technologies of wastewater treatment for food industry enterprises (including dairy industry). The most common solution in this area is the combination of classical treatment methods with new methods [8], while an adequate choice of equipment for a particular enterprise is impossible without a high-quality design task and the preparation of appropriate models (first of all, water technological passport), including those based on digital technologies [9].

III. Problem formulation for an intelligent information and reference system water use by dairy processing enterprises

To solve the problem, we initially perform functional modelling in IDEF0 methodology (Fig. III).

The following categories of parameters (according to IDEF0 terminology) are selected on the basis of technological analysis:

- **input factors**
  - coming from measuring instruments in operational mode:
    - actual flow rate of process water (PW):;
    - about 2-3 process water quality indicators (e.g., pH, electrical conductivity);
    - actual wastewater flow rate;
    - about 3-5 quality indicators of WW: pH, temperature, redox potential (ORP), chemical oxygen demand (COD), electrical conductivity;
    - information on equipment status (based on the production scheme and operating SCADA).
  - measurements of water quality parameters from the laboratory:
    - information on PW quality;
    - information on WW quality.
  - from expert technologists and expert technicians:
    - information on the planned demand for PW per shift;
    - information on the planned demand for ingredients per shift;
    - information on the state of technological equipment.
- **control factors**: requirements for the quality of PW, characteristics of the equipment (under which it operates according to its passport parameters), requirements for technological processes (under which the requirements for their regulatory flow are met), regulatory requirements for the quality of wastewater, cost of resources;
- **mechanisms**: electrical equipment that ensures the operation of the information system as a whole;
- **results**:
  - dynamic balance of water use (based on planned shift production tasks, as well as with the function of calculating financial costs for the period from the beginning of the shift to the current point in time - potentially a forecast is also necessary based on the current state, for example, until the end of the shift);
  - operational forecast of the quality of WW (based on planned shift tasks, as well as with a forecast correction function based on real indicators of production and quality of PW and WW at the current time (with various forecast projections);
  - dynamic recommendations for organizational, technical and technological actions in order to reduce financial costs for water use (based on the current situation and forecast);
  - dynamic costs for minimizing waste pollution (based on the current situation and forecast).

Performing a decomposition of the first level diagram (see Fig. 1) allows you to detail the tasks of digital modeling (Fig. III).

Analysis of the diagram of the first level of decomposition (see Fig. III):

- here the key will be the multi-level decomposition of the block "Systematization of data on water disposal parameters", it will include production subsystems and units that use "process water", polluting it and transforming it into "waste water" - with the main module "Intelligent information and reference system" (IIRS), which will analyze the situation and "consult" process and technical specialists: what operating modes they should choose initially, how to relate to the state of this or that equipment (the flow of production processes), what variants of the assortment task will lead to what resource costs and environmental risks, regarding the resource cost of a specific assortment task, regarding the acceptability of the functioning of a specific unit or the use of an ingredient, etc. (based on a retrospective production analysis);
- at the same time, the "Intelligent Information and Reference System" must, of course, work with the integrating technology support block "Analysis and forecasting of the efficiency of use of water resources of an enterprise."

The "intelligent information and reference system", in fact, should become an adaptive (interactive) technological regulation for water use of an enterprise, largely ensuring the interoperability of the entire system as a whole, with potential transformation along the chain: "decision support system – automated process control system for water use — digital MES (MIS, LIMS, EMI) resources — ERP."

Then the intelligent information and reference system for supporting specialists of milk processing enterprises in the water use segment is a software product (SP) intended for storage in a structured electronic form and prompt provision to other SPs and specialists of various technological information accumulated both in basic regulatory documents (BC ,SS R, BAT) and in specially created databases and knowledge. IISS will make it possible to create a unified information space of technological knowledge for prompt consultation on issues of interest to specialized specialists (managers, chief engineers, WW technologists, instrumentation and control engineers, designers) using data from regulatory documents and advanced solutions obtained, for example,

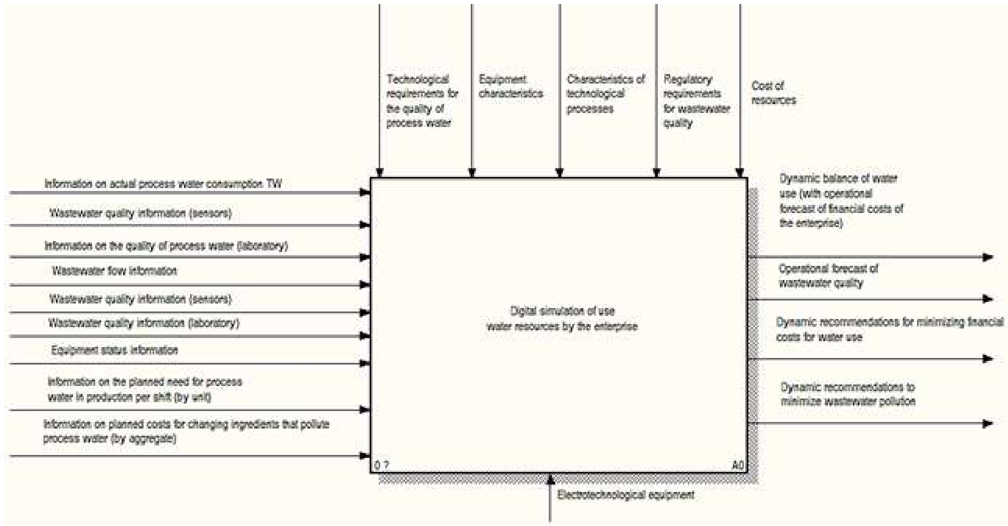


Figure 1. Context diagram for modeling water use of a milk processing plant

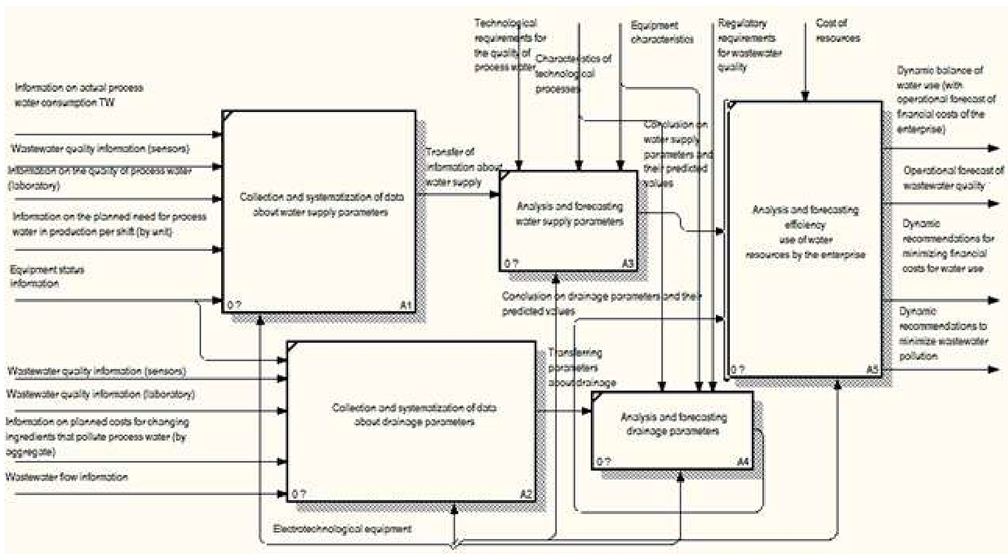


Figure 2. First level of context diagram decomposition

based on benchmarking and expert opinions.

Main planned products:

- software (SP), which can be used by all enterprises, including holding companies;
- educational and methodological materials for continuous improvement of qualifications and retraining of specialists in the field of digital technologies using the created product.

At the same time, it should be noted that water use processes, including the functioning of local treatment facilities, are characterized by nonlinearity, nonstationarity, multifactorial, multiprocess nature, constant changes in the structure of internal relationships, the presence of significant hidden mutual influences between technological parameters, the use of separately functioning ones when solving a single industrial problems of information systems (for example, 5-6 industrial SCADA) [4]–[9].

Accordingly, the proposed (reasonable) transformation ("intelligent decision support system – automated process control system for water use — digital MES (MIS, LIMS, EMI) resources — ERP") requires a specialized methodological apparatus of a new generation.

Such solutions include OSTIS Technology [10]. New generation intelligent computer systems developed on its basis are called OSTIS systems. The OSTIS Technology is based on a universal method of semantic representation (coding) of information in the memory of intelligent computer systems, called the SC code. SC code texts (sc-texts, sc-constructions) are unified semantic networks with a basic set-theoretic interpretation. Elements of such semantic networks are called sc-elements (sc-nodes and sc-connectors, which, in turn, depending on their orientation, can be sc-arcs or sc-edges). The universality and unification of the SC code makes it possible to describe on its basis any type of knowledge and any methods for solving problems, which, in turn, greatly simplifies their integration both within one system and within a group of such systems.

The basis of the knowledge base developed using the OSTIS Technology is a hierarchical system of semantic models of subject areas and ontologies, among which stands out the universal Core of semantic models of knowledge bases and the methodology for developing semantic models of knowledge bases, ensuring semantic compatibility of the developed knowledge bases [10]. The basis of the OSTIS -system problem solver is a set of agents that interact exclusively through the specification of the information processes they perform in semantic memory (sc-agents). All of these principles together make it possible to ensure semantic compatibility and simplify the integration of both various components of computer systems and such systems themselves. Within the framework of the OSTIS Technology, several universal options for visualizing SC code structures are proposed. Within the framework of this work, examples will be used in SCg-code - a graphical version of visualization of SC-code constructions and SCn-code — a structured hypertext version of visualization of SC-code constructions.

IV. Practical implementation of IIRS based on OSTIS Technology (using the example of a unit for post-treatment of wastewater from dairies — biological ponds)

In biological ponds for post-treatment, contaminants are removed by aerobic microorganisms, which enter them with purified liquid from secondary settling tanks (after aeration tanks and/or biological filters), and also develop directly in the ponds themselves. Higher aquatic vegetation (algae) plays an important role. The rate of the biochemical process of extraction and oxidation of organic pollutants in ponds with natural aeration is limited by the low rates of atmospheric reaeration and mass transfer processes. In ponds with artificial aeration, as a result of supplying the required amount of air with intensive mixing of the liquid, the speed of the biochemical process is 5–7 times greater.

The feasibility of their use at dairy processing enterprises is largely due to the fact that many harmful organic substances are not completely oxidized at biological wastewater treatment facilities; they remain stable in water for a long time and can have a toxic effect on living organisms. In addition, intensive technologies have significant disadvantages: high energy costs for aeration and problems associated with the processing and disposal of large amounts of excess sludge formed, its swelling and foaming.

The characteristics of the life cycles of biological ponds fully include: nonlinearity, nonstationarity, multifactoriality, multiprocessivity, constant changes in the structure of internal relationships, the presence of significant hidden mutual influences between technological parameters. Accordingly, it is justified to apply OSTIS Technology to the creation of a knowledge base of their processes with further synthesis of an intelligent information and reference system.

The formation of the IIRS knowledge base was based on the national regulatory document - BS 4.04.02-2019 “Building standards of the Republic of Belarus Sewerage. External networks and structures.” Let’s consider a fragment of the ontology that describes the concept of “biological pond” and the parameters specified on it in the SCn code [10].

biological pond
⇒ *explanation**:
  [Used for purification and post-treatment of municipal, domestic, industrial and surface wastewater containing organic substances.]
⇒ *subdividing**:
  {• *biopond with artificial aeration*
   • *biopond with natural aeration*
  }
⇒ *parameters**:
  {• *waste water flow*
    ∈ *measurable parameter*
    ⇒ *designation**:
      [Qw]
    ⇒ *measurement unit**:
      1 m^3*/day*
  }

waste water
:= *explanation**:
  [surface, domestic and industrial waters discharged into sewers]
⇒ *parameters**:
  {• *TBOD*
    ∈ *biochemical indicator*
   • *average monthly temperature in winter*
   • *average monthly temperature in summer*
  }

TBOD
:= [total biological oxygen demand]
⇒ *explanation**:
  [reflects the amount of oxygen required for the biochemical oxidation of organic wastewater contaminants in 20 days]
∈ *measurable parameter*

⇒ *designation**:
  [L]
⇒ *measurement unit**:
  1 mg/l

Consider for an example of using the developed ontology to describe a specific technological task. A biological pond with a flow rate of Q_w = 5640 m^3/day was taken as a technological prototype. Biochemical indicators of the processed WW: biological oxygen demand (TBOD) L_{en} = 18 mg/l; The technological task is to provide a TBOD of treated wastewater L_{ex} = 5 mg/l. At the same time, the average monthly temperature of PW for the summer period is T_w = 22 °C; average monthly temperature for the winter period T_w = 15 °C. The indicated parameters correspond to the use of a biological pond after a fairly well-functioning biological and physico-chemical wastewater treatment of a large milk processing plant.

Figures IV–IV show an example of a problem condition and related parameters, formalized in the SCg-code [10].

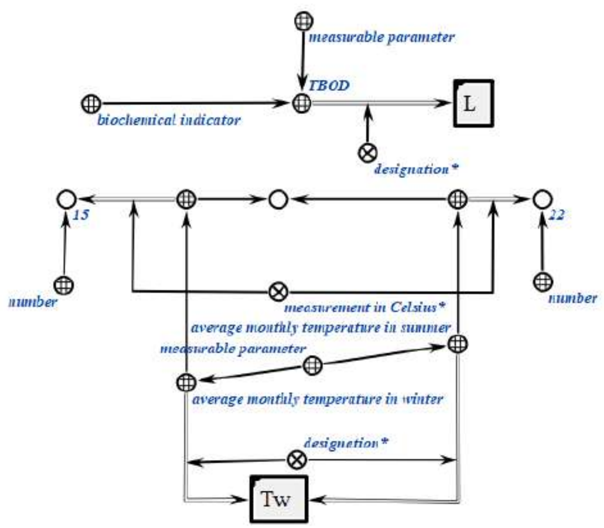


Figure 4. Formulation of the problem conditions (fragment 2)

V. Conclusion

The use of OSTIS Technology and the development of the corresponding ISRS in the analysis and formation of new production knowledge in the water use segment of dairy processing enterprises will allow:

- increase the efficiency and versatility of decision-making, including cases of complex production situations;
- improve administration flexibility, even when expanding technological lines;
- increase the level of qualifications of employees, since the OSTIS Technology represents a knowledge system in natural language and corresponds to modern ideas about the organization of human long-term memory.

Taken together, the creation of such products based on the OSTIS Technology will form an ontological basis for digital modeling of resource use processes (water, electricity, heat, steam, reagents) in dairies.

Acknowledgment

The authors would like to thank the scientific team of the department of intelligent information technologies of the Belarusian State University of Informatics and Radioelectronics for their assistance in the work and valuable comments.

References

[1] Mittal S., Khan M. A., Romero D., Wuest T. Smart manufacturing: Characteristics, technologies and enabling factors. Proc. of the Institution of Mechanical Engineers, Part B: Journal of Engineering Manufacture, 2019, Vol. 233(5), pp. 1342–1361. DOI:10.1177/0954405417736547.

[2] Grieves M., Vickers J. Digital Twin: Mitigating Unpredictable, Undesirable Emergent Behavior in Complex Systems. Kahlen F. J., Flumerfelt S., Alves A. (Eds). Transdisciplinary Perspectives on Complex Systems. Cham: Springer, 2017, pp. 85-113. DOI:10.1007/978-3-319-38756- 7_4.

[3] Tao F., Zhang H., Liu A., Nee A. Y. Digital twin in industry: State-of-the-art. IEEE Trans. on Industrial Informatics, 2019, Vol. 15, pp. 2405-2415.

[4] Shtepa, V. N. Funkcional'no-staticheskij analiz sistemy kontrolja vodootvedenija i ocenka podhodov k ejo cifrovomu modelirovaniju [Functional-static analysis of the water disposal control system and assessment of approaches to its digital modeling]. Computer science and cybernetics: scientific journal, 2023, No. 3(33), pp. 35-42.

[5] Shtepa V. N., Zolotykh N. Yu., Kireev S. Yu. Obosnovanie i shemy ispol'zovanija ranzhirujushhih izmeritel'nyh sistem jekologicheskogo monitoringa i intellektual'nogo analiza rezhimov vodootvedenija [Justification and schemes for using ranking measurement systems of environmental monitoring and intelligent analysis of water disposal regimes]. Bulletin of Polotsk State University. Series F. Construction. Applied sciences: scientific journal, 2023, No. 1, pp. 94–103.

[6] Alekseevsky D. Enhancing Ecological Efficiency in Biological Wastewater Treatment: A Case Study on Quality Control Information System. Water, 2023, Vol. 15, Iss. 21, P. 3744.

[7] Shifrin S. M., Ivanov G. V., Mishukov B. G., Feofanov Yu. A. Ochistka stochnyh vod predprijatij mjasnoj i molochnoj promyshlennosti [Wastewater treatment of meat and dairy industry enterprises]. Light and food industry, 1981, P. 272

[8] Garzanov A. L., Usov A. V., Kushniruk M. Yu., Barabash V. P., Kamalyan O. A. Modernizacija ochistnyh sooruzhenij fabriki morozhenogo [Modernization of treatment facilities at an ice cream factory]. "Morozhenoe i zamorozhennye produkty" ["Ice cream and frozen foods"], 2004, No. 6, pp. 26-27.

[9] Voitov I.V. O zadachah cifrovizacii sistem vodootvedenija kommunal'no-promyshlennyh ob'ektov [On the tasks of digitalization of water disposal systems of municipal and industrial facilities]. Petroleum and gas chemistry – 2023: materials of the VI International Scientific and Technical Forum on Chemical Technologies and Oil and Gas Processing, Minsk, November 1-3, 2023, BSTU, 2023, pp. 147-151.

[10] V. Golenkov, Ed., Tehnologija kompleksnoj podderzhki zhiznennogo cikla semanticheski sovmestimyh intellektual'nyhkomp'juternyh sistem novogo pokolenija [Technology of complex life cycle support of semantically compatible intelligent computersystems of new generation]. Bestprint, 2023.

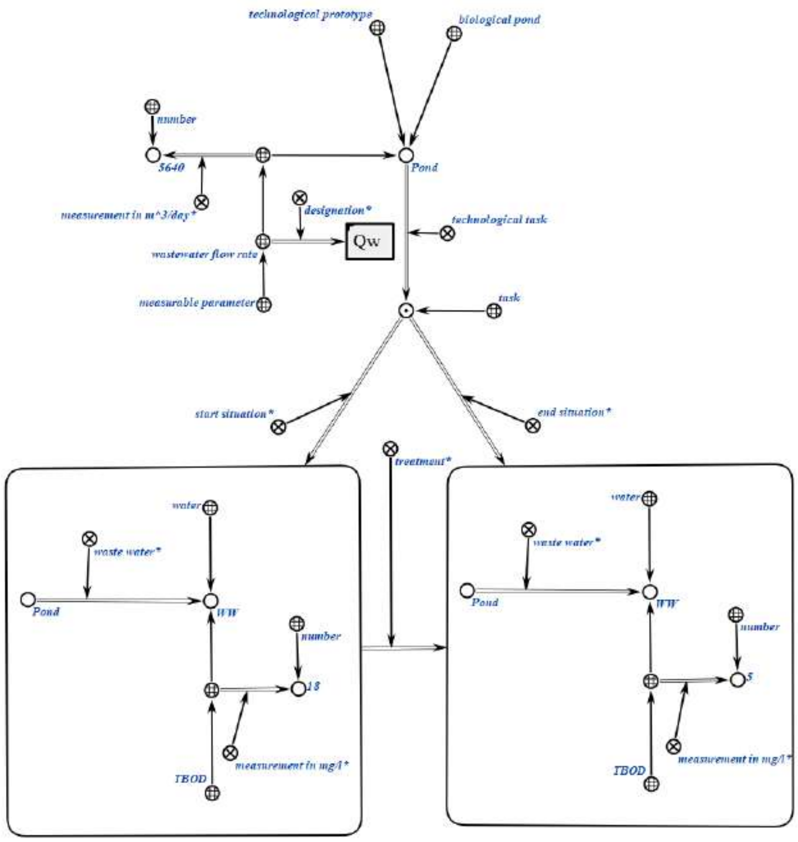


Figure 3. Formulation of the problem conditions (fragment 1)

ПРИКЛАДНЫЕ АСПЕКТЫ ИСПОЛЬЗОВАНИЯ ТЕХНОЛОГИИ OSTIS ПРИ ИНФОРМАЦИОННОМ ОБЕСПЕЧЕНИИ ЦИФРОВИЗАЦИИ ПРОЦЕССОВ ВОДОПОЛЬЗОВАНИЯ МОЛОКОПЕРЕРАБАТЫВАЮЩИХ ПРЕДПРИЯТИЙ

Штепа В. Н., Муслимов Э. Н.

Оценены концептуальные подходы цифровизации производств на основе идеологии e-Manufacturing, которые предложено использовать для моделирования процессов водопользования предприятий по переработке молока, проанализированы организационно-технологические процессы формирования загрязнителей их сточных вод. В методологии IDEF0 выполнено функциональное моделирование водопользования таких производств, что позволило выявить сложность и многонаправленность взаимосвязей параметров и обосновать использование технологии OSTIS для задач формирования интеллектуальной информационно-справочной системы. На примере биологического пруда, как узла очистки сточных вод, реализован элемент предложенного подхода.

Received 13.03.2024

NLP and LLM Based Approach to Enterprise Knowledge Base Construction

Valery Taberko and Dzmitry Ivaniuk
JSC «Savushkin Product»
Brest, Belarus
Email: tab@pda.savushkin.by, id@pda.savushkin.by

Viktor Smorodin and Vladislav Prokhorenko
Department of Mathematical Problems of Control and Informatics
Francisk Skorina Gomel State University
Gomel, Belarus
Email: smorodin@gsu.by, shnysct@proton.me

***Abstract*—An approach to the automated formation of a knowledge base about an enterprise is proposed based on the analysis of existing technical documentation using large language models. This approach makes it possible to integrate the proposed solution with other developments and enterprise software to ensure the construction of intelligent automated control systems, recommendation systems and decision support systems, and information support systems for enterprise personnel.**

***Keywords*—Industry 4.0, standard, large language model, NLP, knowledge extraction**

I. Introduction

Effective implementation of methods for automating control and decision making requires ensuring semantic compatibility of heterogeneous components of intelligent systems [1]. The OSTIS technology uses a multi-agent approach to the development of intelligent systems, when agents operate on a single knowledge base.

Thus, in the course of building intelligent systems that provide solutions to a class of problems of automated management and information support of an enterprise, the task arises of combining all information about the enterprise into a single information space - a knowledge base, which is stored in the semantic memory of the intelligent system. The sources of such knowledge can be existing descriptions of the work of enterprises within the framework of accepted international standards.

Information extraction (IE) is a fundamental challenge in the field of natural language processing (NLP), especially in the context of the modern digital era, where data volumes are rapidly increasing, as is the need for rapid processing [2]. The importance of IE is evident in a variety of applications, ranging from automating text analysis and extracting structured data from various sources to supporting decision making based on large volumes of information. Based on the principles of machine learning and deep learning, modern IE methods provide the ability to automatically extract information from texts and documents, which is an important step in the digital transformation of organizations and society as a whole.

In the framework of Industry 4.0, where digitalization and automation play a key role in industrial transformation, the importance of automating the extraction of information from technical documents becomes critical to building a digital twin of the enterprise. A digital twin of an enterprise is a virtual model of its physical processes, equipment and resources based on real-time data. Automated information extraction systems allow you to quickly and efficiently process technical documents, highlighting key parameters, equipment characteristics, production process structure and other important data necessary to create and maintain a digital twin.

Using a digital twin of an enterprise provides organizations with the opportunity to conduct virtual modeling and analysis of production processes, optimize their efficiency, predict possible problems and take preventive measures. This helps increase flexibility, respond to changes in real time and ensure more efficient use of enterprise resources. Thus, automation of information extraction from technical documents is an important link in the process of creating and maintaining a digital twin of an enterprise, which ultimately helps to increase its competitiveness and sustainability in the market.

Information support for employees of an industrial enterprise also plays a key role in ensuring competitiveness and production efficiency. With the development of digital technologies and the penetration of the Internet of things into production processes, employees are faced with a huge amount of data that requires analysis and decision-making. Information support not only allows you to effectively manage data, but also provides access to up-to-date information in real time, which is necessary for a quick response to changes in the production environment.

At the present stage of development of complex technological systems, the volume of information contained in technical documentation describing the relevant subject areas can be large and also difficult to analyze due to the heterogeneity and ambiguity of the interpretation of some provisions of the standards [3].

This paper proposes an approach to solving this problem: automation of knowledge extraction from documentation based on the use of modern neural network technologies - large language models.

II. Problems of working with standards when designing new generation intelligent systems

Existing international standards for describing production systems make it possible to release an ontological approach to solving problems of production automation, building recommender systems and enterprise information support [4] . Thus, we can highlight the correspondence of the concepts of the ontology of the subject area "probabilistic technological production processes" to the concepts from the ISA 5.1, ISA-88 and ISA-95 [1], [3], [5]–[7] standards.

At the same time, a number of problems arising in connection with the use of standards can be identified. One of the key difficulties is the heterogeneity and fragmentation of standards in various areas, which makes it difficult for them to interact and integrate into the overall system. These standards are often developed by different organizations and committees, resulting in a variety of formats, syntax and semantics, which makes them difficult to understand and interoperate within a single system. This creates compatibility problems between various system components, and also increases the complexity of the development and support process [8]–[10].

Another major challenge is the rapid development of technology, which leads to constant updating and modification of existing standards. With the development of new technologies, standards also evolve, but on the other hand, periodic changes and updates may lead to incompatibility between different versions of standards and technologies. This complicates the integration and support of intelligent systems and can also cause compatibility issues between different components and devices, making it difficult to communicate and communicate between them. Thus, successful implementation and maintenance of intelligent systems requires continuous updating and adaptation to changing standards and technologies.

Overcoming these difficulties is critical to the development of effective intelligent systems that can analyze, predict and optimize various aspects of enterprise functioning. This includes not only managing production processes, but also optimizing logistics schemes, forecasting demand, analyzing market trends and many other aspects of business. In this context, the use of modern methods of data processing, machine learning and artificial intelligence becomes a necessity to create adaptive and intelligent systems that can effectively respond to dynamically changing market conditions and the competitive environment.

III. Large language models

The historical path of natural language processing (NLP) before the advent of large language models is characterized by the gradual development of methods and approaches to text analysis. Since early research in linguistics and artificial intelligence, such as the creation of the first grammars and automatic translation systems, progress in NLP has been associated with finding ways for computers to efficiently understand and generate natural language. In the early periods of NLP development, the emphasis was on rules and symbolic approaches to text analysis, which often limited its applicability due to the difficulty of creating complete and accurate rules for different languages and contexts.

However, the true breakthrough came with the development of machine learning methods and deep learning in particular. The emergence of large language models in recent years has opened up wide opportunities for working with text information in various areas of human activity. These models, such as GPT [11] and BERT [12], have a unique ability to understand the semantics of natural language and can efficiently analyze and process large volumes of text data.

Research shows that the use of such models leads to significant improvements in natural language processing tasks, including machine translation, sentiment analysis, information extraction, and more [13]. Thanks to their capabilities, working with text information becomes more efficient and accurate, which opens up new prospects for the development of automated data processing systems and artificial intelligence.

In the context of the problem under consideration, it is important, in particular, to note significant progress in solving the problem of information extraction [14]. Integrating large language models into the process of extracting information from documents according to a given ontology is an important and promising direction of research in the field of natural language processing (NLP). These models have a unique ability to understand the semantics of the text and can effectively highlight key entities and relationships between them

At the moment, we can note existing proposals for integrating LLM with OSTIS Technology [15], [16], which use third-party services (ChatGPT [17])

It is also necessary to emphasize that despite the modern successes of using LLM in various fields of activity, it is also necessary to note their disadvantages, such as inconsistency and the possibility of errors. Such models can be highly sensitive to noise and anomalies in the input data, which can lead to unpredictable results and potential errors in text interpretation [18]. There is a well-known phenomenon of hallucination in large language models, which is a phenomenon in which the model generates content that may be unpredictable, unrelated to the context, or even fictitious [19]. This phenomenon is a consequence of the huge amount of data on which models are trained, as well as their complex architecture, which includes billions of parameters. Under certain conditions, models can produce text that appears consistent and plausible, but is actually a random combination of words and phrases [20].

Moreover, the interpretability of large language models poses a significant challenge. Due to their utilization of intricate deep learning architectures, the internal workings of these models often remain opaque and convoluted, hindering human comprehension. This lack of transparency raises concerns regarding the justification of the model's decisions and undermines the confidence in relying on its conclusions for critical decision-making processes. Without a clear understanding of how the model arrives at its outputs, stakeholders may hesitate to fully trust its recommendations or insights, thereby impeding the integration of these advanced AI systems into real-world applications where trust and accountability are paramount.

This phenomena highlight the need to critically analyze the results obtained from large language models and emphasizes the importance of developing methods to control the quality and reliability of their use.

IV. An approach to automated construction of a knowledge base about an enterprise from a standardized description

The main idea of the proposed approach is to implement a component of the OSTIS ecosystem that implements functionality for automated extraction of information from standardized technical documentation of an enterprise.

The importance of prompt engineering for large language models is emphasized in the context of ensuring accurate and efficient interaction with the model. Prompts, being the key interface between the user and the model, determine the formulation of the problem and the context of the request, which directly affects the quality and relevance of the answers. Prompt engineering allows you to structure the input data for the model, taking into account the specifics of the problem and the requirements of the application, which provides more accurate and relevant results [21]. Automatic generation of prompts for large language models is a key area of research aimed at improving their performance and adaptability to a variety of tasks and contexts.

To successfully solve the problem of extracting information from standardized documents in accordance with a given domain ontology, it is necessary to generate a prompt that will explicitly define the required information and provide the model with an understanding of the context of the document and the key concepts of probabilistic technological processes.

In the prompt, the target information should be explicitly defined, namely, a set of concepts for work, and the connections between them. In addition, it is necessary to take into account the specific structure and format of documents of a particular standard. Including examples of the target input and output in the prompt provides not only additional context for the model, but also a more explicit representation of the desired output. This approach allows the model to better capture user intent, recognize key query elements, and correlate them with corresponding output. In this way, the model becomes capable of generating more accurate and appropriate responses or actions, which significantly improves its functionality and meets the users' needs for better service.

Thus, during the operation of the automated knowledge extraction component, it is necessary to solve the following tasks:

- determine the standard of the document and carry out its preliminary preparation;
- generate a prompt containing a description of a specific standard and domain ontology, as well as example of the expected output;
- send a prompt to a large language model;
- receive the model's response and verify its correctness and compliance with the ontology;
- use the extracted knowledge from the answer in the knowledge base.

The general scheme of prompt formation is shown in Fig. 1.

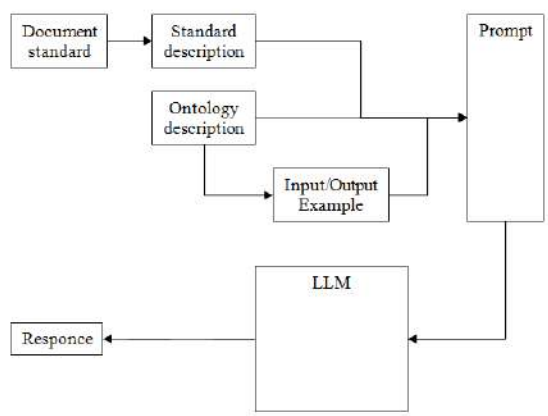


Figure 1. Scheme for generating a prompt for LLM

Within the framework of the OSTIS Technology, a multi-agent approach to problem solving [1] is used. This approach involves defining a hierarchy of agents that perform atomic functions for preparing and converting information. The proposed composite solver within this approach has the structure presented in Figure 2. Each agent specializes in certain aspects of information processing.

abstract non-atomic sc-agent of automated knowledge extraction from technical documentation
⇒ *decomposition of abstract sc-agent**:
{• *abstract sc-agent of document preparation*
• *abstract sc-agent of standard description*
• *abstract sc-agent of ontology description*
• *abstract sc-agent of LLM communication*
• *abstract sc-agent of LLM responce parsing*
}

Figure 2. Decomposition of the abstract non-atomic sc-agent of automated knowledge extraction

The sc-agent of document preparation is responsible for the preliminary preparation of the document and determining the type of standard to which it belongs. The document is converted into text form. This process may be accompanied by the use of image analysis tools (if images are present in the document).

The sc-agent of standard description is capable of returning a description of the standard to which the document in question was defined. It conveys information about the specific organization of text in documents of a particular standard.

The sc-agent of ontology description agent is capable of returning a description of the ontology of the subject area (for example, "probabilistic technological production processes"): a set of concepts, their description and a description of the connections between them.

The sc-agent of LLM communication is responsible for generating a prompt from information collected from previous agents and sending it as a request to the LLM. After sending, the agent expects to receive a response. The agent is not tied to a specific implementation or location of a large language model.

The sc-agent of LLM response parsing is the key element responsible for checking the correctness of the response received from the large language model (LLM). This agent checks the LLM response for compliance with the data format and consistency with the domain ontology. Its functionality is necessary to account for possible situations of "hallucination" or errors that may occur in the operation of the knowledge extractor-LLM. If the answer is successfully verified, the extracted information can be used to populate the knowledge base. Given the complex nature of the task that this agent needs to solve it may operate also using an LLM.

V. An example of extracting information from documentation using LLM

An example of the information extraction approach can be demonstrated using the project documentation for the PLCnext testbed [22] (Fig. 3, 4, 5).

The documentation contains extensive text on software settings for working with the stand (Fig. 6). Using the described prompt engineering, you can use LLM (ChatGPT 3.5 [17]) to structured knowledge extraction for each configuration step in JSON format, where each instruction step is correctly separated into a separate element (Fig. 7 and 8).

Figure 3. General view of the stand

Figure 4. Controller and input/output node of the stand

VI. Conclusion

This paper proposes an approach to solving an important practical problem - automated knowledge extraction for constructing enterprise knowledge bases, based on the application of an agent-based approach within the framework of OSTIS Technology and the use of large language models. The logic of operation of the corresponding component of the OSTIS Ecosystem is presented, which allows you to systematize and retrieve information from various data sources. The described approach can be applied not only in the field of enterprises, but also

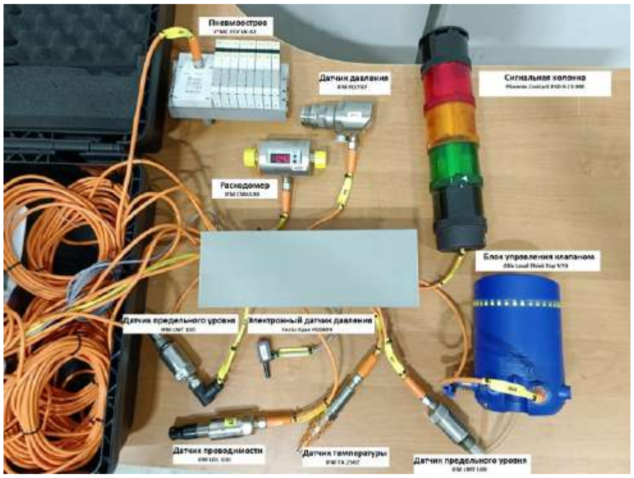


Figure 5. Sensors and actuators

```
# Тестовый стенд **"T1_PLCNext_Demo"**

| Тегов | Строк кода | Аппаратов | Агрегатов | Устройств | IO-Link
I/O | Узлов | Модулей |
|---|---|---|---|---|---|---|---|
|![](statistics/tags_total.svg) |![](statistics/lines_total.svg)
|![](statistics/units_total.svg)
|![](statistics/agregates_total.svg)
|![](statistics/devices_total.svg)
|![](statistics/io_link_usage.svg)
|![](statistics/io_couplers_total.svg)
|![](statistics/io_modules_total.svg)

## Внешний вид демостенда

![общий вид](images/demo_stand.jpg)
<p align="center"> Рисунок 1 - Демостенд</p>

## Содержание

Запуск в режиме эмуляции (эмулятор "на столе").

1. [Клонирование проекта](#клонирование-проекта)
2. [Установка Visual Studio Code](#установка-visual-studio-code)
3. [Установка Git LFS](#установка-git-lfs)
4. [Настройка DbEditorXML](#настройка-dbeditorxml)
5. [Установка и настройка EasyServer](#установка-и-настройка-
easyserver)
6. [Установка и настройка Monitor](#установка-и-настройка-monitor)
7. [Запуск и сборка проекта](#запуск-и-сборка-проекта)
8. [Взаимодействие с физическим устройством Bus Coupler]
(#взаимодействие-с-физическим-устройством-bus-coupler)

Запуск на контроллере.
```

Figure 6. Original view of the document text

in other subject areas where work with standardized documentation is required as a source for populating knowledge bases.

```
{
    "Запуск проекта через консоль": [
        "Установить Toolchain нужной версии (подробная инструкция по
установке здесь).",
        "Подключиться к контроллеру через клиентское программное
обеспечение WinSCP.",
        "Перейти в каталог для своего контроллера
ptusa_main\\bin\\install\\AXCF3152\\Release\\bin\\ и оттуда
скопировать исполняемый файл ptusa_main в свой каталог на
контроллере, например, /opt/main/.",
        "Далее в этот же каталог скопировать файлы с расширением lua
и plua, а также директории spec и sys из проекта T1-PLCnext-Demo.",
        "После чего открыть Kitty, перейти в каталог со всеми
скопированными файлами:",
        "cd /opt/main/",
        "И запустить исполняемый файл ptusa_main:",
        "./ptusa_main main.plua sys_path ./sys/",
        "В случае возникновения ошибки Permission denied, добавить
файлу флаг -x, чтобы сделать его исполняемым. Это можно реализовать
либо через интерфейс WinSCP, либо через консоль с помощью команды
chmod.",
        "chmod +x ptusa_main"
    ]
}
```

Figure 7. Structured Information Retrieval

```
{
    "Подключение к контроллеру в программе PLCnext Engineer": [
        "Переходим на вкладку с настройками контроллера.",
        "Далее нажимаем на кнопку подключения.",
        "Далее вводим имя пользователя (admin) и пароль (напечатан на
корпусе контроллера). После успешного подключения отображаются текущие
параметры контроллера."
    ],
    "Обновление библиотеки": [
        "Скачать файл последней версии библиотеки (*.PCWLX).",
        "Скопировать данный файл с заменой в папку PLCnext необходимого
проекта.",
        "Откройте клиентское программное обеспечение SFTP (например
WinSCP).",
        "Авторизуйтесь как администратор (admin - обратитесь к
администратору).",
        "Копировать файл с заменой файла библиотеки из
/AXCF2152_22.0.4.144/Release/lib/libPtusaPLCnextEngineer.so в каталог
/opt/plcnext/projects/PCWE/Libs/Ptusa, где AXCF2152_22.0.4.144 -
соответствующая версия библиотеки.",
        "Откройте командную оболочку с помощью инструмента командной строки
(например, KiTTY).",
        "Авторизуйтесь как администратор (admin - обратитесь к
администратору).",
        "Перезапустите службу plcnext (команда: sudo /etc/init.d/plcnext
restart)."
    ]
}
```

Figure 8. Structured Information Retrieval

References

[1] V. Golenkov, Ed., Tehnologija kompleksnoj podderzhki zhiznennogo cikla semanticheski sovmestimyh intellektual'nyh komp'juternyh sistem novogo pokolenija [Technology of complex life cycle support of semantically compatible intelligent computer systems of new generation]. Bestprint, 2023.

[2] C. D. Manning, H.Schütze Foundations of statistical natural language processing, 1999, MIT press.

[3] V. Taberko, D. Ivaniuk, V. Smorodin, V. Prokhorenko Adaptive Control System for Technological Process within OSTIS Ecosystem, Otkrytye semanticheskie tekhnologii proektirovaniya intellektual'nykh system [Open semantic technologies for intelligent systems], 2023, pp. 291-298.

[4] V. Taberko, D. Ivaniuk, N. Zotov, M. Orlov, O. Pupena, N. Lutska Principles Of Building A System For Automating The Activities Of A Process Engineer Based On An Ontological Approach Within The Framework Of The Industry 4.0 Concept, Otkrytye semanticheskie tekhnologii proektirovaniya intellektual'nykh system [Open semantic technologies for intelligent systems], 2021, pp. 209-218.

[5] "ISA5.1 Standard," https://www.isa.org/standards-and-publications/isa-standards/isa-standards-committees/isa5-1/, (accessed 2024, Jan)

[6] "ISA-88 standard," Available at: https://www.isa.org/isa88/, (accessed 2024, Jan).

[7] "ISA-95 standard," https://www.isa.org/standards-and-publications/isa-standards/isa-standards-committees/isa95/, (accessed 2024, Jan)

[8] J. Lee, B. Choi A survey of standardization efforts for service-oriented architecture, Expert Systems with Applications, 2009, 36(4), pp.7844-7855.

[9] M. Salama, S.K. Mostefaoui, A. Nour A survey on IoT interoperability and standardisation issues, International Journal of Ad Hoc and Ubiquitous Computing, 13(3), 145-155.

[10] V. Taberko, D. Ivaniuk, D. Shunkevich, O. Pupena Principles For Enhancing The Development And Use Of StandardsWithin Industry 4.0, Otkrytye semanticheskie tekhnologii proektirovaniya

intellektual'nykh system [Open semantic technologies for intelligent systems], 2020, pp. 167-174.
[11] A. Radford, K.Narasimhan, Improving Language Understanding by Generative Pre-Training, 2018
[12] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, 2019 North American Chapter of the Association for Computational Linguistics.
[13] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, Dario Amodei Language models are few-shot learners. arXiv preprint arXiv:2005.14165, 2020
[14] D. Xu, Wei C., W. Peng, C. Zhang, T. Xu, X. Zhao, X. Wu, Y. Zheng, E. Chen Large Language Models for Generative Information Extraction: A Survey, 2020, arXiv preprint arXiv:2005.14165.
[15] A. Zagorskiy Integration Of Third-Party Functional Services On A Unified Semantic Basis, Otkrytye semanticheskie tekhnologii proektirovaniya intellektual'nykh system [Open semantic technologies for intelligent systems], 2023, pp. 207-212.
[16] A. Cherkas, A. Kupo Ostis Technology Integration With Third-Party NLP Services, Otkrytye semanticheskie tekhnologii proektirovaniya intellektual'nykh system [Open semantic technologies for intelligent systems], 2023, pp. 207-212.
[17] "Introducing ChatGPT", Available at: https://openai.com/blog/chatgpt, (accessed 2023, January).
[18] P.Kohli, R.Gupta, S.Saxena, Robustness Challenges in Language Models: A Survey. arXiv preprint arXiv:2103.01021, 2021
[19] J. Dodge, S.Gururangan, D.Card, R.Schwartz, N. A.Smith, S. R.Bowman Fine-tuning language models from human preferences. arXiv preprint arXiv:1909.08593, 2020
[20] A. Holtzman, J. Buys, J. Du, M. Forbes, Y.Choi, The curious case of neural text degeneration. arXiv preprint arXiv:1904.09751, 2020
[21] Y. Liu, M.Ott, N. Goyal, J. Du, M.Joshi, D.Chen, V. Stoyanov, Roberta: A robustly optimized BERT pretraining approach, arXiv preprint arXiv:1907.11692, 2019
[22] "T1-PLCnext-Demo", Available at: https://github.com/savushkin-r-d/T1-PLCnext-Demo, (accessed 2023, January).

NLP ПОДХОД К ПОСТРОЕНИЮ БАЗЫ ЗНАНИЙ ПРЕДПРИЯТИЯ С ИСПОЛЬЗОВАНИЕМ БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ

Таберко В. В., Иванюк Д. С.,
Смородин В. С., Прохоренко В. А.

Предложен подход к автоматизированному формированию базы знаний о предприятии на основании анализа существующей технической документации с применением больших языковых моделей. Данный подход позволяет обеспечить возможность интеграции предлагаемого решения с другими разработками, программными средствами предприятия для обеспечения построения интеллектуальных систем автоматизированного управления, рекомендательных систем и систем поддержки принятия решений, систем информационного обеспечения персонала предприятия.

Received 13.03.2024

Basic Priciples of the Ontology of the Transportation Process in Railway Transport

Aleksandr Erofeev
Belarusian State University of Transport
Gomel, Belarus
Email: erofeev_aa@bsut.by

Ilya Erofeev
Belarussian State University of Informatics and Radioelectronics
Minsk, Belarus
Email: ilerofv@gmail.com

***Abstract*—The relevance of developing an ontology for an intelligent transportation process management system is determined. The structure of the theory of ontology construction is given. The existing ontologies of railway systems are described. The problems of existing ontologies are established. It is proposed to use a process-object approach in the formation of ontology. Examples of the use of ontology are given. Indicated that the OSTIS technology is an effective tool for describing the process-object ontology of the transportation process.**

***Keywords*—Intelligent transportation process management system, ontology, knowledge base, process-object approach, OSTIS technology**

I. Introduction

Digitalization of technological processes is one of the most effective ways to improve the efficiency of both individual enterprises and entire industries. This fully applies to railway transport.

For these purposes, the Belarusian Railway has implemented a number of projects on the informatization of key technological processes. The Information and Analytical Management Decision Support System for Freight Transportation (IAMDSS FT) provides specialists and managers at all levels with prompt and reliable information about freight transportation that are being performed and were performed, the condition and location of the wagon and locomotive fleets. The Unified Corporate Integrated Financial and Resource Management System (UC IFRMS) based on SAP SE products is designed for effective information support of planning, modeling and operational management of financial and economic activities of the BR divisions.

The creation of the Transportation Management Center and its equipping with automated dispatch control systems using "Neman" software and hardware has been completed, the development of train schedules using "Grafist" automated control systems has been automated, and separate planning procedures have been automated. At the line level, there are automated management systems for stations and linear areas (AMSS, AMSLA), a system for automating the preparation and processing of station and commercial reporting documents (SAPPD).

The implementation of the above-mentioned and a number of other systems made it possible to create a fairly detailed and complete database on the transportation process and the railway as a whole. At the same time, information technologies operating in railway transport make it possible to implement only certain functions of the management cycle, entrusting the procedures for forming, evaluating and making managerial decisions (MD) entirely to a human.

The greatest problems arise with significant deviations from pre-developed technologies (delay of a passenger train, shortage of capacity of the station's fleet, lack of empty wagons for loading, etc.). It is at those moments, when a person needs help in making managerial decisions, that modern automated systems are not effective enough. In such conditions, a person, as a rule, when searching for rational MD, resorts to his experience based on partially unformalized knowledge and empirical experience, and to a lesser extent relies on the help of information systems.

Next logical stage in the development of information technologies in railway transport is the transition from information and reference systems to information and analytical and information management systems. In this regard, as foundation for the digital transformation of railway transport is proposed to consider an intelligent transportation process management system (ITPMS).

The development of the ITPMS presupposes the existence of unified ways to describe the system and the processes occurring in it — ontologies. [1]

II. The concept and basis of ontology

The use of a single "metaplatform" contributes to the improvement of transport processes, simplification of document flow and acceleration of information exchange between participants in multimodal transportation. This approach leads to a synergistic effect on three levels: within each mode of transport, between different modes of transport and in the transport complex as a whole.

However, the development of such systems faces a huge amount of information that needs to be integrated and coordinated within a single digital platform. Data sources are often disparate even within a single company, and the information contained in them can be duplicated, diverse, and have complex relationships such as synonymy and mero-holonymy. The complexity of the task increases exponentially when scaled to the level of the entire transport system or the transport complex as a whole.

In the field of multimodal transportation, the active use of knowledge brings significant benefits to all participants. For example, successful innovative solutions that were once used and accelerated the movement of goods in the logistics chain or brought additional profits are becoming available to everyone. The ability to generate new knowledge can help to find such solutions, even if they have not been used before. Ontologies, as well as frame models of knowledge representation, are the basis of such solutions [2].

Ontologies can be classified into four levels depending on their goals:

1) Representation Ontology: They define the domain of knowledge representation and serve as the basis for creating a specification language for lower-level ontologies.
2) Top-level Ontologies: They describe abstract interdisciplinary concepts and the connections between them. Such ontologies capture general knowledge applicable in several subject areas.
3) Domain Ontology: They describe and summarize concepts specific to a particular subject area.
4) Applied Ontologies: They describe the conceptual model of a specific task or application.

The following approaches are distinguished within these levels:

- Formal taxonomies: They define a "class-subclass" relationship and ensure the transitivity of this relationship.
- Formal instances: They define a "class-instance" relationship and add instances to the class hierarchy.
- Disjunctive classes, inverse properties: They allow you to declare two classes disjoint and allow you to make not only direct inference, but also reverse.
- Ontologies with restrictions on the scope of property values: Property values are selected from a predefined set or subset of concepts.
- Frame-based properties: Each class can also have slots containing information about its properties, also they can be inherited by classes of higher levels.
- Arbitrary logical constraints: Allow you to define arbitrary axioms.

Ontologies can also be classified by content into:

- General Ontologies: They describe the most general concepts, independent of specific problems or areas.
- Task ontologies: They reflect the specifics of particular applications, tasks or programs.
- Subject ontologies: They describe real objects used in a certain activity.

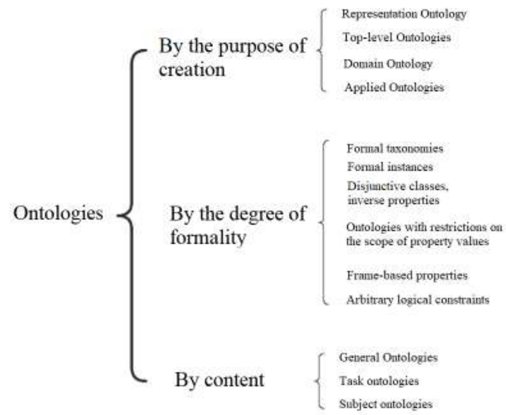


Figure 1. Ontology classification

According to Gruber's definition [4], an ontology is an explicit specification of a conceptualization, a formalized representation of concepts and the connections between them. Formally, it has the form

$< C, R, P, A >$,

where

C is a set of concepts of a specific subject area,

R is the set of relations between concepts,

P is the set of attributes of concepts,

A is a set of axioms.

III. Current state of ontologies in railway transportation

The process of creating ontologies is complex and non-trivial. When developing them, as a rule, experts from different fields are involved (in the subject area, linguists, programmers, etc.), each of whom applies their own methods of work. But there are still no universal methods, and the correctness of the result obtained is determined specifically for each particular task. [5], [6] Ontological analysis determines the functional structure (connections between objects) and determines the behavior of the system (scenarios).

The existing IDEF5 [3] standard describes the construction of ontologies in a very simplified way. Due to the lack of a theoretical methodology for determining the structure of information, the task of finding the fundamental foundations for creating a developed transport ontology is urgent.

At the moment, ontologies are used in intelligent railway transport management systems (IRTMS) both in JSC "Russian Railways" and on the railways of the European Union. However, the analysis of these ontologies shows that they are rather intended for storing and consolidating data on the state of various objects of the transportation process. They are applied and are not currently focused on expanding to the entire subject area.

The ontologies used on the railways of the European Union are even more narrowly oriented and do not include the more abstract, technological level of knowledge that is present in the ontologies of this field. One of the reasons for the inapplicability of Western ontologies is the structural difference of the railway system. The terminology sets describing the Belarusian Railway and European railways differ significantly. Despite the presence of common concepts at a high level, such as "Station" and "Wagon", at lower levels, such as classification of wagons, tracks, document management and others, the terms are fundamentally different.

Nevertheless, the ontology being formed should be interdisciplinary, since various modes of transport with their own technology and specifics of work, regulatory documents, loading and unloading regulations, etc. participate in transportation. To ensure internal consistency of ontologies, it is necessary to determine a common "point" around which they can be combined. In this case, such a unifying "point" may be cargo, and then the subject area will be "cargo transportation".

Thus, the use of foreign ontologies is not possible, and in the case of integration with European systems, the ontology will have to contain universal concepts describing those concepts in a form understandable to both sides.

It should also be noted that the railway infrastructure is a strategically important facility. Thus, this area should be import-independent. And one of the problems of the existing ontology used in IRTMS is the dependence on Western instruments.

Despite the fact that, at first glance, it is sufficient to consider the ITPMS as an ontology of the subject area, it must be borne in mind that the ITPMS is the first stage in building an Intelligent Transport Management System and must have all the features of top-level ontologies. This task is inextricably linked with the formation of a digital model of the transportation process (DMTP). The current mechanisms for the formation of the DMTP should allow for online modeling of the state of the transportation process. This requires the unification of requirements for the content and form of presentation of information about the parameters of the functioning of objects. DMTP may include:

- models of objects (including resources);
- process models — description of processes occurring both in the ITPMS and in the external object environment;
- models of the external object environment describing the external impact on the objects of the transportation process;
- models of predicting situations — study of options for the development of the transport situation with abnormal changes in the state of elements of the transport system, the external environment, with changes in the characteristics of information flows;
- models for the formation of managerial decisions that provide an analysis of the operational situation and the formation of effective managerial decision;
- evaluation models that provide an assessment of the effectiveness of implemented MD, the condition of facilities and the parameters of the transportation process.

The basis for the formation of the ontology under consideration can be the subsystem of the normative-reference information of IAMDSS FT (SNRI). This subsystem contains many tables describing both static and dynamic objects. However, in its pure form, SNRI does not allow solving operational tasks, but is used only as reference information. Thus, the scientific problem of developing tools for the transformation of normative-reference information into an ontological description of both the subject area and interdisciplinary concepts should be solved.

IV. Our proposal on the direction of research in the field of development of railway transport ontology.

In our work we propose to use the OSTIS technology and platform, which is being developed in the Republic of Belarus. Within the framework of this technology, tools for creating ontologies are implemented, as well as creating agents to work with it. [7]–[10]

Based on the above justifications of relevance, the following tasks can be distinguished for ontology:

- Fully describe the objects and processes of the transportation process.
- Converting information from existing standards to a knowledge base and back.
- Set of agents for data transformation.
- Storage and analysis of external information.
- Storage of managing decision models.
- Set of agents for converting incoming information into a managing decision.

Accordingly, the Knowledge Base can be divided into subdomains.

1) Description of the main objects involved in the transportation process;
2) Description of the processes occurring in the ITPMS, as well as external processes affecting the ITPMS;
3) Description of external objects that affect the transportation process;
4) Description of possible emergency situations, models for predicting their development;
5) Description of possible managing decisions;
6) Description of criteria for evaluating managing decisions and the result of their impact on the system.

Some examples of subject areas that can be included in the ontology of the transportation process:

1) Railway lines and stations: This subject area describes information about various railway lines, including their geographical location, length, speed limits, track types, and accessibility. It also includes information about stations, including their location, capacity, and cargo and passenger handling capabilities.
2) Cargoes and containers: This subject area describes various types of cargoes and containers, their characteristics (weight, volume, type of packaging), processing and storage requirements. It also includes information about special requirements for certain types of goods, for example, goods requiring special transportation conditions (temperature conditions, vibration protection, etc.).
3) Schedule and Train Timetable: This subject area describes information about the train timetable, including departure and arrival times, intermediate stops, travel speeds and possible overlaps on the track. It also includes information about the regularity and repeatability of the train schedule.
4) Resources: This subject area describes information about available resources such as locomotives, wagons, personnel, and technical equipment. It includes the characteristics of resources (load capacity, technical parameters), their availability and maintenance schedule.
5) Laws and Regulations: This subject area describes the laws, rules and regulations governing railway logistics. It includes safety rules, standards for the transportation of certain goods, requirements for staff working hours, rules for the priority of train traffic and other regulatory information.
6) Monitoring and Data collection: This subject area describes the monitoring and data collection systems used to obtain information about the current state of railway logistics. It includes data on train movement, cargo status, information about delays and other factors affecting planning and management.

Various problem solvers can be used to automate transportation planning and management. Here are some typical problem solvers that can be applied in this field:

1) The routing problem solver: This solver allows you to optimize train routes, taking into account various factors such as schedules, availability of railway lines, cargo requirements and restrictions on infrastructure and transportation conditions. It helps to choose the best routes, minimizing the time and cost of transportation.
2) Train timetable development solver: This solver allows to optimize the train timetable, taking into account the timetable, passenger and cargo requirements, infrastructure constraints and other factors. It helps to allocate trains by time and resources, ensuring efficient use of the railway network.
3) Loading planning problem solver: This solver helps to optimize the loading of wagons, taking into account the characteristics of goods and limitations of wagons, such as load capacity, dimensions and special requirements. It helps to maximize the loading of wagons, minimizing empty runs and improving the use of the wagon fleet.
4) The solver of the optimal resource planning problem: This solver allows you to optimize the allocation of resources such as locomotives, wagons, personnel and technical equipment to ensure the efficient operation of the railway system. It takes into account transportation requirements, traffic schedules and resource constraints, helping to achieve optimal use of resources and reduce costs.
5) Demand Forecasting and Planning Solver: This solver is used to predict the demand for rail transportation, taking into account various external factors such as economic conditions, seasonality, trends and others. It helps to plan capacity and resources to meet demand and prevent congestion or lack of transportation capacity.
6) Monitoring and Management Problem Solver: This solver is used for continuous monitoring and management of railway transportation. It analyzes data on train movements, traffic status, delays and other events, allowing you to make operational decisions and respond to changes in real time.

These are just some examples of problem solvers that can be used to automate the planning and management of railway logistics. The specific set of solvers will depend on the specific requirements and characteristics of the railway logistics system, as well as on the goals and constraints set.

Using ontology for these problem solvers in the field of railway logistics can bring several advantages:

1) Knowledge structuring: Ontology allows you to structure knowledge about the main subject areas of railway logistics, such as infrastructure, resources, cargo and schedules. This facilitates the understanding and organization of information, simplifies its accessibility and exchange between different systems and participants.
2) Data unification and standardization: An ontology defines common semantics and standards for data representation in railway logistics. This allows different systems and applications to use the same terms and data formats, which ensures consistency and compatibility of information. This is especially important when integrating different systems and exchanging data between them.
3) Improvement of planning and optimization: Ontology provides a formalized domain model on the basis of which planning and optimization algorithms

can be developed. The use of ontology makes it possible to take into account various limitations, requirements and dependencies between different aspects of railway logistics. This contributes to more efficient use of resources, route optimization and improved service quality.

4) Decision Making Improvement: Ontology can be used in decision support systems, providing a semantic model and context for data analysis and recommendation generation. It allows systems to identify dependencies and relationships between various factors, to carry out forecasting and scenario analysis, which helps to make reasonable and informed decisions.
5) Ease of expansion and modification: Ontology provides a flexible and extensible domain model that can be easily modified and supplemented as needed. This allows you to adapt the ontology t new requirements, changing conditions and expar its functionality. This flexibility makes it easier t support different types of tasks and to develop tl system in the future.

In general, the use of ontology for data solvers in tl field of railway logistics contributes to improving da organization, information compatibility and consistenc optimizing planning and management processes, as we as making informed decisions based on data analysis.

The ontology for these problem solvers in the fie of transportation process management should be flexib and modular in order to take into account various facto and adapt to changing conditions.

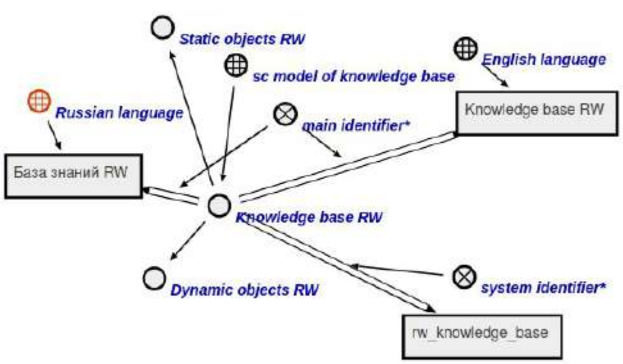


Figure 2. Semantic neighborhood of knowlegebase RW

Within the framework of this work, a top-level ontology is implemented describing the main objects of the transportation process, examples of objects of which are presented in Figures 2-4 in the form of an SCg code [10].

V. Conclusion

Thus, within the framework of this work, the relevance of developing an ontology for an intelligent transportation process management system has been determined.

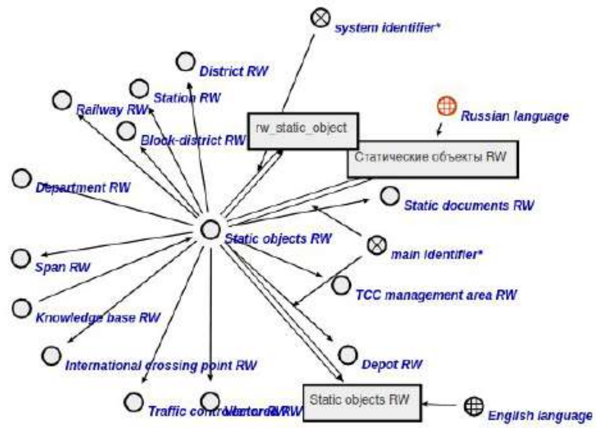


Figure 3. Semantic neighborhood of static objects RW

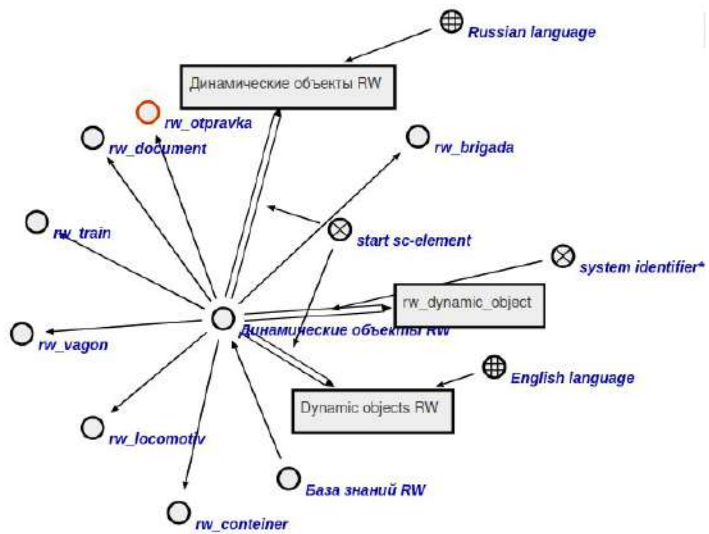


Figure 4. Semantic neighborhood of dynamic objects RW

The structure of the theory of ontology construction is given. The existing ontologies of railway systems are described. The problems of existing ontologies are established. It is proposed to use a process-object approach in the formation of ontology. Examples of the use of ontology are given. It is indicated that the OSTIS technology is an effective tool for describing the process-object ontology of the transportation process. The top-level ontology for the Belarusian Railway has been implemented.

References

[1] A. A. Erofeev Intelectualnaya sistema upravlenia perevozochnim processom [Intelligent control system for the transportation process in railway transport]. The Ministry of Education of the Republic. Belarus, Belarusian state University of Transport. Gomel, BSUT, 2022, P. 407
[2] A. P. Badeckii, O. A. Medved Ontologicheskii podhod k razrabotke edinoi basi znanii multimodalnih perevozok [An ontological approach to the development of a unified knowledge base

for multimodal transportation]. Proceedings of the St. Petersburg University of Railway Engineering, 2023, № 1, pp. 182–193

[3] Ontology Description Capture IDEF5. https://www.idef.com/idef5-ontology-description-capture-method/ (Accessed 10.03.2024)

[4] T. Gruber Toward Principles for the Design of Ontologies Used for Knowledge Sharing. Int. Journal Human-Computer Studies, Vol. 43, pp. 907–928.

[5] A. D. Obuhov Ontology neshtatnih situacii na sortirovochnoi stancii po nepriemu poezda [Ontology of emergency situations at the marshalling yard because of unreception of trains]. Materials of the IV International Scientific and Practical Conference "Modern concepts of scientific researchers", N. Novgorod, 2015, pp. 281-284.

[6] A. V. Borisov, A. V.Bosov, D. V.Zukov, A. V. Ivanov, D. V. Sushko, Informationnie aspekti obespechenia bezopasnosti na transporte: ontologiya predmetnoy oblasti, modeli i varianti ispolzovania [Information aspects of transport safety: domain ontology, models and use cases], Systems and Means of Informatics, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences: 30:1 (2020), pp. 126–134

[7] V. Golenkov and N. Guliakina, “Principles of building mass semantic technology component design of intelligent systems,” in Open semantic technologiesfor intelligent systems. Minsk, BSUIR, 2011, pp. 21–58 (in Russian)

[8] V. Golenkov and N. Gulyakina, “Semanticheskaya tekhnologiya komponentnogo proektirovaniya sistem, upravlyaemyh znaniyami [Semantic technology of component design of knowledge-driven systems],” in Open semantic technologiesfor intelligent systems, Minsk, BSUIR, 2015, pp. 57–78, (in Russian)

[9] V. V. Golenkov, “Ontology-based design of intelligent systems,” in Open semantic technologies for intelligent systems, ser. Iss. 1. Minsk : BSUIR, 2017, pp. 37–56.

[10] V. V.Golenkov, Tehnologija kompleksnoj podderzhki zhiznennogo cikla semanticheski sovmestimyh intellektual’nyh komp’juternyh sistem novogo pokolenija [Technology of complex life cycle support of semantically compatible intelligent computer systems of new generation] – Minsk:Bestprint, 2023. P. 1063 (in Russian)

ОСНОВЫ ОНТОЛОГИИ ПЕРЕВОЗОЧНОГО ПРОЦЕССА НА ЖЕЛЕЗНОДОРОЖНОМ ТРАНСПОРТЕ

Ерофеев А. А., Ерофеев И. А.

Определена актуальность разработки онтологии для интеллектуальной системы управления перевозочным процессом. Приведена структура теории построения онтологии. Описаны существующие онтологии железнодорожных систем. Установлены проблемы существующих онтологий. Предложено при формировании онтологии использовать процессно-объектный подход. Приведены примеры использования онтологии. Указано, что эффективным инструментом описания процессно-объектной онтологии перевозочного процесса является технология OSTIS.

Received 13.03.2024

Examples of Integration of Intelligent Computing Modules and the System GeoBazaDannych

Valery B. Taranchuk
Department of Computer Applications and Systems
Belarusian State University
Minsk, Republic of Belarus
taranchuk@bsu.by

***Abstract*— Methodological and technical solutions for the integration of modules of intelligent computing of the Wolfram Mathematica system and tools of the GeoBazaDannych software complex in the tasks of creation, interpretation, processing, visualization of digital fields in computer modeling of objects of geology, geoecology are discussed.**

***Keywords*—system GeoBazaDannych, intelligent adaptation of digital fields, methods of computer model adaptation, clustering**

I. Introduction

Computer modeling includes the development of mathematical methods and algorithms; software development, and computational experiments. The appropriate software should always provide downloading from different sources and preprocessing of data, correlation, formation of digital cubes of object characteristics, interactive data analysis, visualization using graphics. A modification of the typical clustering method is proposed, and the advantages are confirmed by calculations based on representative data.

One of the most important components in the development and implementation, in particular, of constantly operating computer-based geological and geoecological models is the task of assessing the adequacy and accuracy of the proposed digital descriptions. The key issues are automation of the creation, adaptation of models taking into account the constantly incoming additional data, as well as revision of the results of processing the initial information using new interpretation methods [1].

One of the main goals of data mining is the detection of previously unknown, non-trivial, but understandable interpreted knowledge in "raw" (primary) arrays of information.

At the same time, following [2], “data mining does not exclude human participation in processing and analysis, but significantly simplifies the process of finding the necessary data from raw data, making it available to a wide range of analysts who are not specialists in statistics, mathematics or programming. Human participation is expressed in the cognitive aspects of participation and the application of informational cognitive models".

Geodata mining tools are the same as for usual data; the basis is the theory, methods, and algorithms of applied statistics, databases, artificial intelligence, and image recognition.

A number of issues related to the analysis and evaluation of spatial data quality can be solved using the computer system GeoBazaDannych [3].

Possible options, methodological solutions, and software tools that allow you to confirm the validity of interpretations, visualize and obtain numerical values of errors calculated by different methods of intellectual data processing results included and used in computer geological models are discussed below. For illustrations, the key task of forming and processing digital fields used in computer models is selected. In particular, the methods proposed and tested in solving various applied problems are discussed, as well as specialized algorithms for calculating approximating digital fields implemented in the interactive computer complex GeoBazaDannych.

The GeoBazaDannych is the interactive computer complex of intelligent computer subsystems, mathematical, algorithmic and software for filling, maintaining and visualizing databases, input data for simulation and mathematical models, tools for conducting computational experiments, algorithmic tools and software for creating continuously updated computer models. GeoBazaDannych subsystems allow you to calculate and perform expert assessments of local and integral characteristics of ecosystems in different approximations, calculate distributions of concentrations and mass balances of pollutants; create permanent models of oil production facilities; generate and display thematic maps on hard copies [3], [4].

The main components of the GeoBazaDannych: the data generator Gen_DATv; the generator and editor of thematic maps and digital fields Gen_MAPw; the software package Geo_mdl — mathematical, algorithmic and software tools for building geological models of soil layers, multi-layer reservoirs; the Generator of the geological model of a deposit (GGMD) — the integrated software complex of the composer of digital geological and geoecological models.

We note the main additions to the GeoBazaDannych, including methodological aspects and software components implemented using artificial intelligence tools. At the same time, in order to understand the stages of development and the relationship of the components, we will give examples of the use of updated components of the complex and recall fragments of the results that were discussed and published in the proceedings of OSTIS.

In recent years, the activity of using artificial intelligence tools in solving problems of geology and geoecology has been rapidly increasing. In particular, dozens of articles are published every year on the algorithms and methods of cluster analysis considered in this paper. There are publications that provide solutions to practical problems, methods of preprocessing and interpretation of geophysical data, analysis of results, expert opinions of conclusions and recommendations (for example, [5], [6]).

A number of features of data preparation for computer-based geological and geoecological models from the perspective of the feasibility of using artificial intelligence tools have been regularly discussed at OSTIS conferences since 2019. We will note only the typical difficulties that arise when developing tools and conducting computational experiments for specific practical tasks, which were identified and solutions are presented in subsequent results.

Thus, in the materials of the OSTIS-2019 conference ("Examples of the use of artificial neural networks in the analysis of geodata"), methodological and technical solutions, software tools, results and examples of data processing typical for geophysical methods of studying geological objects, in particular, on observation profiles, were discussed and published. The effectiveness of the use of artificial neural networks to eliminate noise and errors in the measurement results, perform the necessary data preprocessing for mathematical models by smoothing in order to prepare regular digital distributions is illustrated.

In the materials of the OSTIS-2020 conference ("Examples of intelligent adaptation of digital fields by means of the GeoBazaDannych system" and "Interactive Adaptation of Digital Fields in the GeoBazaDannych System"), examples of interactive formation of digital models of geological and geoecological objects in computational experiments that meet the intuitive requirements of an expert are considered and given. Methodological and algorithmic solutions effective in processing remote environmental monitoring data, special tools of the GeoBazaDannych system are noted, the results of interactive adaptation and comparison with standard reference solutions in the complex "Generator of the geological model of the deposit" are presented.

The examples show that this way you can significantly improve the quality and adequacy of the digital description. But you need to understand that at that stage of the state of the algorithmic and software complex of GeoBazaDannych, the allocation itself is implemented according to the intuitive suggestions of the expert.

The issues of automatic identification of sites of the "highlighted" type using cluster analysis tools integrated into the GeoBazaDannych system were discussed at the OSTIS-2022 conference; some positive results were published in the proceedings of the conference and in [7]. In particular, the results illustrating the effects of choosing and confirming the best clustering algorithms are presented, and the options for using different clustering methods are compared, moreover, for different ways of setting the metric distance. In the above-mentioned published materials of OSTIS-2022, along with the positive ones, the disadvantages of the described software implementations and the settings used were noted.

A number of additions to the GeoBazaDannych complex, other variations of the settings of the Wolfram Mathematica system tools for data clustering have been tested in calculations, visualized and described below. Nevertheless, it should be understood that further research, improvement of algorithms and modification of software tools are needed.

II. Initial data, a reference distribution for computational experiments

The results below cannot be strictly mathematically justified, but are indicative and adequate due to the rules of their preparation. On the one hand, they are being formed by random number generators and refined by an expert in order to give them the character of comparable data from field observations in the practice of geophysical methods for studying geological, geoecological objects On the other hand, accepted mathematical expressions are used for measurement values in observations (in calculations, this is an imitation). In fact, for each particular processing algorithm, it is known that the basic (reference) digital distribution can be calculated with the necessary accuracy — comparing calculations with the standard, one can judge the advantages and disadvantages of the method used. Below, the reference distribution (in the accepted GeoBazaDannych terminology, the surface) differs slightly from that considered in [4], the location of the fragments of disturbances and their shape and dimensions have been changed for some. Generally speaking, there was no need for changes, because the set of fragments of typical relief elements in [4] was quite representative, but calculations were performed specifically for a reference surface of a different shape to ensure stable reproduction of the qualitative properties of the results obtained. All the calculation variants described below were performed for both reference distributions, they confirmed that the results are qualitatively the same; they do not change with variations in the size, position, orientation of perturbations.

The results presented in this paper are obtained using the numerical values of the level marks of the (reference) surface according to the formula (1):

$$
\begin{gathered}
zSurfH(x,y) = fOriginF(x,y)+ \\
+400 \cdot fHill6(0.005 \cdot (x-250), 0.007 \cdot (y-400))+ \\
+600 \cdot fHill3(0.01 \cdot (x-150), 0.01 \cdot (y-150))- \\
-200 \cdot fHill(0.01 \cdot (x-880), 0.015 \cdot (y-500))- \\
-150 \cdot fHill(0.02 \cdot (x-920), 0.004 \cdot (y-100))+ \\
+200 \cdot fHill5(0.006 \cdot (x-450), 0.001 \cdot (y-150)), \\
fOriginF(x,y) = zBasicF(x).
\end{gathered} \tag{1}
$$

The visualization of the zSurfH reference surface is shown in Figures 1 and 2, where 3D views are shown in the surface and volume variants; Figure 2 shows a map of isolines. Additionally, the numbers of fragments of disturbances are added to the image on the contour map (in parentheses in blue).

The corresponding scheme of their placement is shown in Figure 3, where the isolines of the reference surface and the one reconstructed in Wolfram Mathematica are also given (the Interpolation method, InterpolationOrder = 1).

The initial data for demonstrating methods and algorithms of mining and clustering were obtained using a random number generator, in which the following were set: the number of observation profiles, points on each profile, and coordinates of the beginning and end of the profile were generated in the specified ranges of values. The values in the points were calculated using the formula (1) — simulation of measurements of the level of the surface being restored. Note that in fact we have a scattered set of points.

III. Illustrations, comparison of the results of the refined clustering methods

Cluster analysis allows for many different types of clustering techniques/algorithms to determine the final result [8], [9]. We note the development, a new way of processing data, and present the results of the comparison Without going into the details of the algorithmic, software implementation, we recall that in [7] clustering was actually performed according to two parameters, namely, grouping was carried out according to the criterion of proximity of points with measurements. In the presented results, grouping is performed according to a combination of three values, namely, for each point, their coordinates (x,y) and the z value (surface level) were taken into account. Here are several interpretations

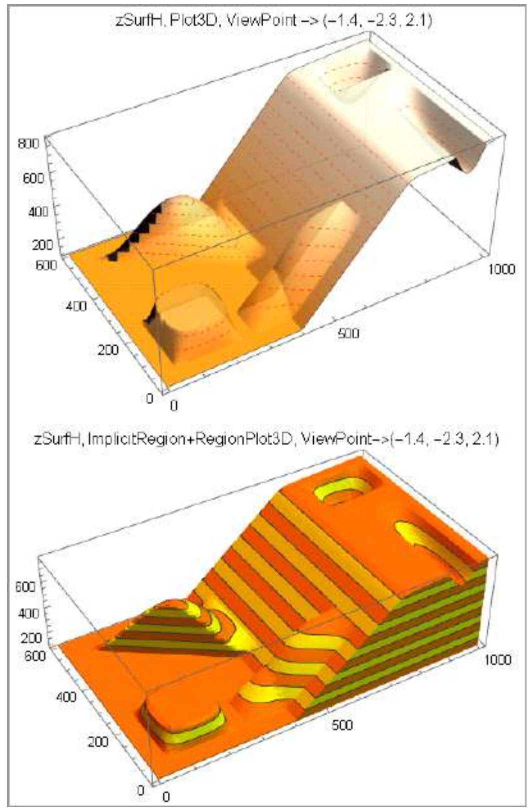


Figure 1. 3D views of the zSurfH reference surface.

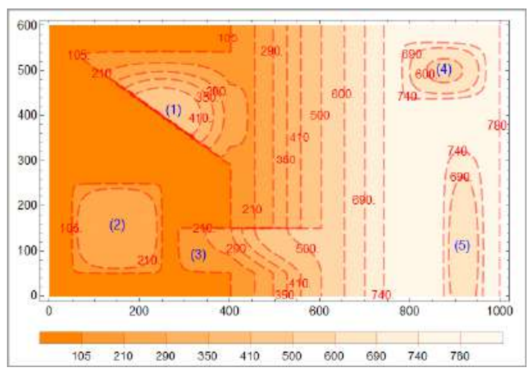


Figure 2. Contour map of the reference surface zSurfH.

of representative calculation results using the Wolfram Mathematica FindClusters function with different criteria of Criterion Function, we will explain the illustrations.

The schemes in Figure 4 repeat the graphic layers from the maps above, they are useful for interpretation and explanation. The schema at the top details the reference, clearly marked sections of fragments-perturbations. The schema below clearly shows areas where there is no reproduction of the standard, which is explained by the lack of measurements. The contour map with data points at the bottom is useful for understanding that in some areas the field cannot be restored to the standard

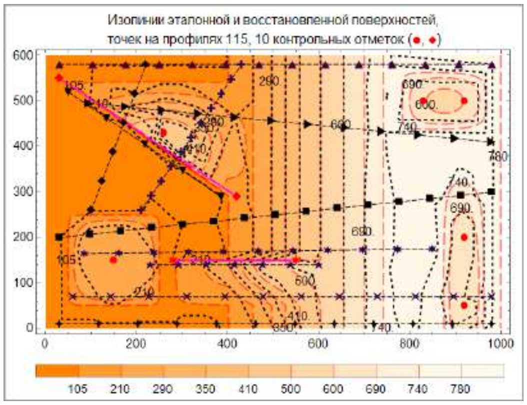


Figure 3. A scheme of the measurement points of the levels, a map of the isolines of the reference and reconstructed surfaces.

because there are no measurements. It is clear that in parts of the area where the digital field differs from the reference, classification / binding to a fragment-disturbance is unlikely.

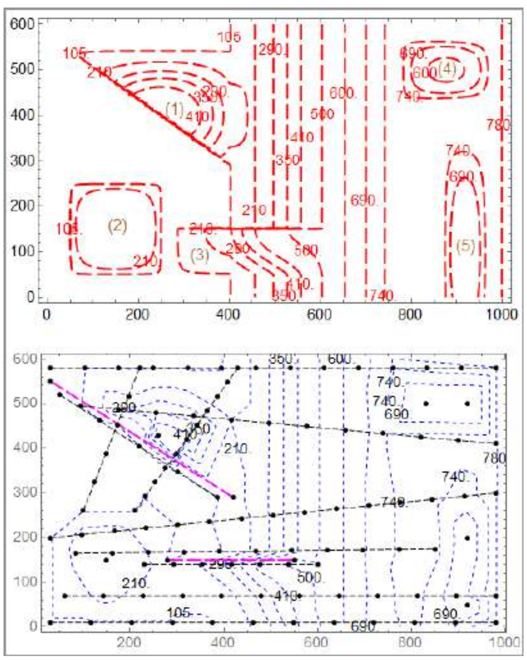


Figure 4. Schemas for understanding the differences between the processed data and the reference.

Figure 5 shows the results of clustering by two parameters (rProfXY) in the upper part, and by three (rProfXYZ) at the bottom. The upper illustration shows the results of calculations with the settings of the KMeans method, as in [7]. Below is the current implementation for the same method; in both versions, the metric is "default", the number of clusters is 6.

The proposed clustering method for three parameters is clearly preferable to the method for two, in particular, in terms of localization of fragments 1, 4 and 5. The

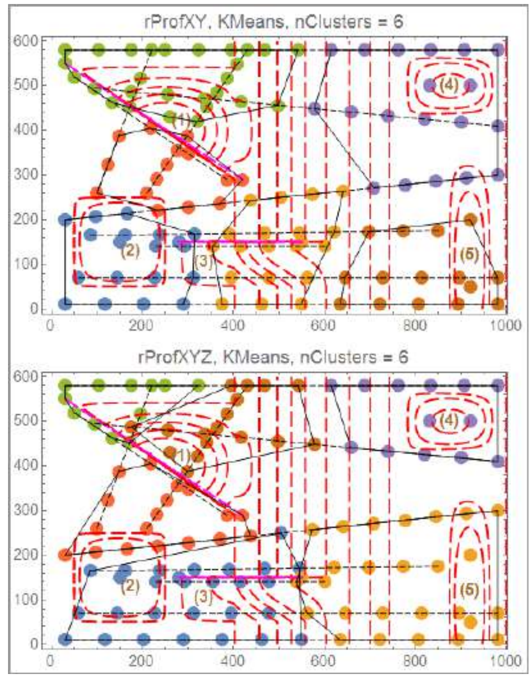


Figure 5. Clustering results for two (rProfXY) and three (rProfXYZ) parameters.

results near fragments 2 and 3 are not indicative due to the discontinuous distribution on different sides of the horizontal dotted magenta line (projection onto the horizontal plane of the fragment section-perturbation 3 by the vertical plane).

Such situations (gaps, offset) are not identified at all in conventional digital field restoration systems without special a priori additional conditions. Note that in the GeoBazaDannych, such conditions can be set interactively by correcting the initial information on the map [3].

The figure shows the results of the variant when the number of clusters is set to 6. Why. Determining the number of clusters is one of the most important segmentation problems. In a broader sense, this is the problem of initializing the algorithm. The results of the variants for the number of clusters from 4 to 9 were calculated and compared. According to the results of the comparison, the variant of 6 clusters seems preferable, and the explanation for this may probably be that the distribution is reproduced when there is a basic continuous surface-a ribbon and 5 fragments-perturbations, i. e. 6 different shapes.

A. The impact of the clustering method

The variants of clustering results using different methods and metrics are illustrated in Figures 6 and 7, the names of the methods and metrics are written in the headings of the diagrams The corresponding software application included in the GeoBazaDannych from the Wolfram Mathematica system allows variants of the clustering method (Criterion Function): Automatic, Agglomerate, DBSCAN, GaussianMixture, JarvisPatrick, KMeans, KMedoids, MeanShift, NeighborhoodContraction, Optimize, SpanningTree, Spectral [10]. What segmentation methods are used in the calculations are written in the headings of the diagrams. Representative clustering options are shown, namely K Means (k-means clustering algorithm), k-medoids (partitioning around medoids), Optimal (Wolfram Mathematica method). The effects of the accepted clustering method (Possible settings for Method) are illustrated by the schemes in Figure 6. Clustering in the examples of this series was considered for three parameters, the FindClusters function was used, the norm in the examples of the series in Figure 6 was not set, but was determined by the default calculation module. These results are quite indicative. At the same time, taking into account the reference and the digital field of the original, we can consider the clustering options by the KMeans and Optimal methods as preferable.

B. The impact of the metric

The issues of measuring the proximity of objects have to be solved with any interpretation of clusters and various classification methods, moreover, there is an ambiguity in choosing the method of normalization and determining the distance between objects. The influence of the metric (DistanceFunction) is illustrated by the diagrams in Figure 7. The results presented in this series are obtained by means of the corresponding software application included in the GeoBazaDannych from the Wolfram Mathematica, which allows different options for setting DistanceFunction.

The Wolfram Language provides built-in functions for many standard distance measures, as well as the capability to give a symbolic definition for an arbitrary measure. In particular, the following metric variant sare available for analyzing digital data [10]: EuclideanDistance, SquaredEuclideanDistance, NormalizedSquaredEuclideanDistance, ManhattanDistance, ChessboardDistance, BrayCurtisDistance, CanberraDistance, CosineDistance, CorrelationDistance, BinaryDistance, WarpingDistance, CanonicalWarpingDistance. What methods of DistanceFunction are used in calculations is recorded in the headers of the schemes. Representative variants are shown, namely: EuclideanDistance (the length of a line segment between the two points), ChessboardDistance, SquaredEuclideanDistance, BrayCurtisDistance, ChebyshevDistance (a metric defined on a vector space where the distance between two vectors is the greatest of their differences along any coordinate dimension), ManhattanDistance. It follows from the above results that for the considered configuration of data points, taking into account the digital field of the original, clustering options using Spectral EuclideanDistance methods can be considered preferable.

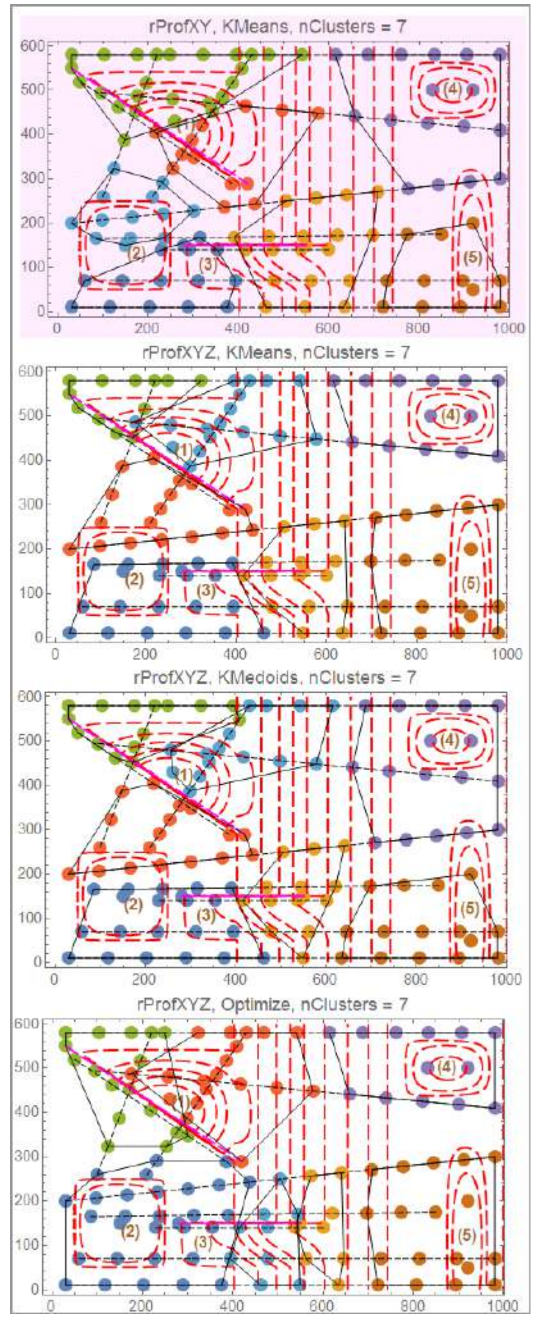


Figure 6. The effects of the accepted clustering method.

IV. Conclusion

The issues of instrumental filling and use of the interactive computer system GeoBazaDannych, expansion of its functionality through integration with the Wolfram Mathematica computer algebra system are considered. A modification of the typical clustering method is proposed, and computational experiments have confirmed the advantages in comparison with traditional methods.

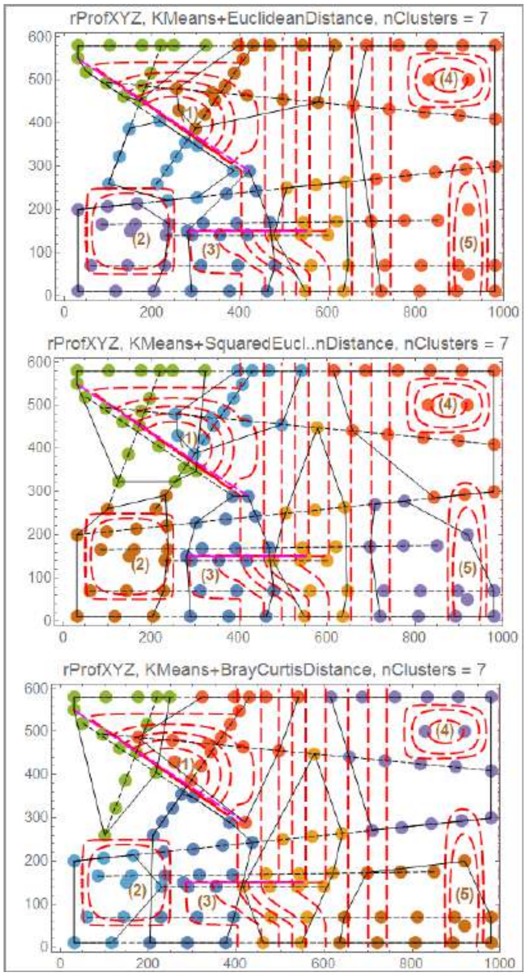


Figure 7. The effects of the accepted metric (EuclideanDistance, SquaredEuclideanDistance,BrayCurtisDistance).

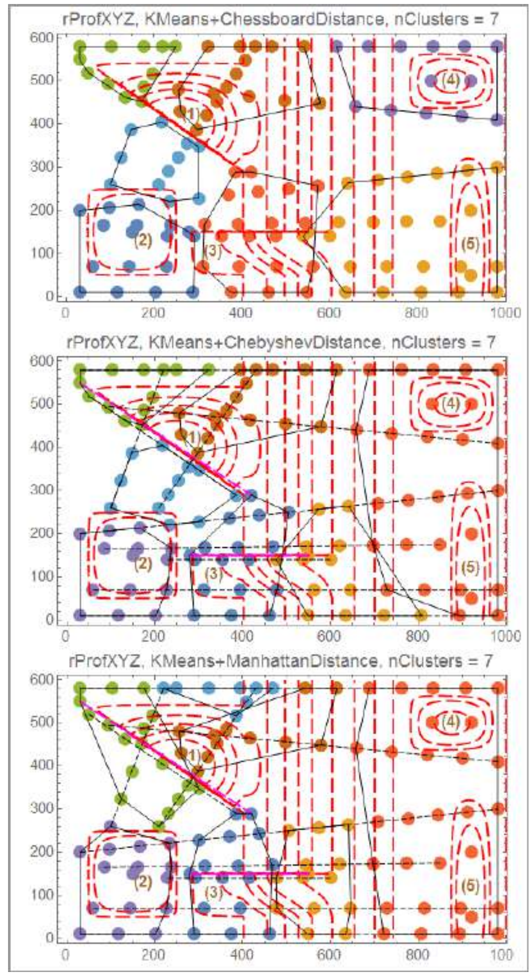


Figure 8. The effects of the accepted metric (ChessboardDistance, ChebyshevDistance, ManhattanDistance).

References

[1] V. P. Savinyh, V. Ya. Cvetkov, “Geodannye kak sistemnyi informacionnyi resurs”, Vestnik Rossiiskoi akademii nauk, vol. 84, no. 9, pp. 826–829, 2014, (in Russian).

[2] V. Golenkov, N. Guliakina, and D. Shunkevich, Otkrytaja tehnologija ontologicheskogo proektirovanija, proizvodstva i jekspluatacii semanticheski sovmestimyh gibridnyh intellektual’nyh komp’juternyh sistem [Open technology of ontological design, production and operation of semantically compatible hybrid intelligent computer systems], V. Golenkov, Ed. Minsk: Bestprint [Bestprint], 2021. 690 p.

[3] V. Taranchuk, Komp’yuternye modeli podzemnoi gidrodinamiki. Minsk: BGU, 2020. 235 p., (in Russian)

[4] V. Taranchuk, “Interactive Adaptation of Digital Fields in the System GeoBazaDannych”, Communications in Computer and Information Science. Book series Springer, vol. 1282, –2020, P. 222–233.

[5] M. Abdideh, A. Ameri, “Cluster Analysis of Petrophysical and Geological Parameters for Separating the Electrofacies of a Gas Carbonate Reservoir Sequence. Nat Resour Res 29, 1843–1856 (2020). https://doi.org/10.1007/s11053-019-09533-1

[6] M. A. Tkachenko, Y. V. Karelina “A cluster analysis as a method of identifying potential chromite mineralisation within the Voykar-Synyinsky massif”. International Research Journal. — 2023. — 10 (136). DOI: 10.23670/IRJ.2023.136.63

[7] V. Taranchuk, “Integration of computer algebra tools into OSTIS applications”. Otkrytye semanticheskie tekhnologii proektirovaniya intellektual’nykh system [Open semantic technologies for intelligent systems], 2022, no 6, pp. 369-374.

[8] D. Tupper Charles, “Concepts of Clustering, Indexing, and Structures”, Data Architecture, 2011, pp. 241–253, https://doi.org/10.1016/B978-0-12-385126-0.00013-9.

[9] B. S. Everitt, S. Landau, M. Leese, D. Stahl. Cluster Analysis. 5th Edition, John Wiley & Sons, 2011, 360 p.

[10] Distance and Similarity Measures. https://reference.wolfram.com/language/ guide/ DistanceAndSimilarityMeasures.html/ (accessed 2024, Feb).

ПРИМЕРЫ ИНТЕГРАЦИИ МОДУЛЕЙ ИНТЕЛЛЕКТУАЛЬНЫХ ВЫЧИСЛЕНИЙ И СИСТЕМЫ ГЕОБАЗАДАННЫХ

Таранчук В.Б.

Обсуждаются методические и технические решения интеграции модулей интеллектуальных вычислений системы Wolfram Mathematica и инструментов программного комплекса ГеоБазаДанных в задачах формирования, интерпретации, обработки, визуализации цифровых полей при компьютерном моделировании объектов геологии, геоэкологии. Предложена модификация типового способа кластеризации, расчетами на представительных данных подтверждены преимущества.

Received 04.04.2024

Designing Intelligent Systems with Integrated Spatially Referenced Data

Sergei Samodumkin
Belarusian State University of Informatics and Radioelectronics
Minsk, Belarus
Email: Samodumkin@bsuir.by

***Abstract*—The paper is devoted to the issues of representation, integration and processing of spatially referenced data in intelligent systems built on the principles of ostis-systems.**

***Keywords*—OSTIS, intelligent system with integrated spatially referenced data, semantic model, question language, design process automation**

I. Introduction

Large sets of spatially referenced data accumulated by mankind, their representation and storage through the created cartographic services [1]–[3], the development of remote sensing technologies [4] have contributed to the creation and development of applied geoinformation systems for various purposes.

Modern geographic information systems are computer systems that provide input, manipulation, analysis and output of spatially referenced data about the territory, social and natural phenomena in solving tasks related to inventory, analysis, modeling, forecasting and management of the environment and territorial organization of society

Since the above tasks are intelligent, such systems belong to the class of intelligent systems with integrated spatially referenced data (ISRD).

The currently offered ISRD development tools are not sufficiently interoperable due to the lack of unification of knowledge of subject areas for the benefit of which application systems are designed, ontologies of terrain objects and phenomena, and temporal components.

It is obvious that for a fixed territory the same spatially related data are used in different application areas: epidemiology, construction, environmental protection and nature management, land relations, creation of digital twins of enterprises, mobile robotics systems, etc., which determines that it is necessary to harmonize the ontologies of subject areas with the objects of terrain and phenomena inherent in a given territory, thus ensuring a vertical (subject-oriented) level of design. On the other hand, when designing ISRD for a new territory, the basic functional requirements are preserved and it is necessary to take into account not only the previous experience of system design, but also to use previously designed functional components, i. e. we are talking about the horizontal level of ISRD design when the territorial area is expanded and systems are designed for new territories.

The third aspect is the temporal component, relevant for retrospective analysis and modeling, thus ensuring the creation of dynamic ISRD that can deal with terrain objects and phenomena within a specific time period.

Therefore, the constant evolution of models and tools for ontological description of subject areas using spatial and temporal components, heterogeneity of spatial components and ambiguity of temporal components, poses new challenges in terms of interaction, integration and interoperability of different types of knowledge used in ISRD by integrating subject area ontologies (vertical level), extending systems at the horizontal level to new territories and time intervals, re-using

The necessity of solving the indicated problems determines the demand for intelligent systems with integrated spatially referenced data and indicates the existence of scientific and technical problem of intellectualization of systems with integrated spatially referenced data, and also becomes relevant to create the development tools themselves, which ultimately provides information support and automation of the activities of developers of application systems.

Within this article, fragments of structured texts in the SCn-code [5] will often be used, which are simultaneously fragments of the source texts of the knowledge base, understandable to both human and machine. This allows making the text more structured and formalized, while maintaining its readability. The symbol ":=" in such texts indicates alternative (synonymous) names of the described entity, revealing in more detail certain of its features.

II. Knowledge-based approach in the tasks of representation, integration and processing of spatially related data

The importance and necessity of analyzing data in space and time is primarily to discover hidden connections and patterns that may not be obvious at first glance. Representation of spatial data, correlating them with objects on the ground and time intervals provides opportunities to determine cause-and-effect relationships, to identify groups of similar data, and to predict future events.

On the one hand, the representation, integration and processing of spatially referenced data is the task of a corresponding class of systems called geographic information systems (GIS). On the other hand, the focus on using spatially referenced data to establish semantic links between spatially referenced data and the knowledge of the subject areas for which the GIS is being developed indicates the need to use artificial intelligence technologies and design intelligent systems.

It is noted in the work [6] that at the current level of development geographic information systems have become practically the main tool for modeling natural, economic, social processes and situations, tracing their relationships, interactions, predicting further development in space and time, and most importantly a means of providing (supporting) decision-making management. Modeling in geographic information systems is based on databases and knowledge bases. The former integrate digital cartographic, aerospace, statistical and other data reflecting the spatial position, state and attitude of objects, and the latter contain sets of logical rules, information, concepts necessary for modeling and decision-making. At the same time, GIS is a special technology based on computer complexes and software tools.

Consequently, from the very definition of GIS follows the need to implement intelligent tasks: analysis, modeling, forecasting and environmental management, because all these tasks are intelligent and require decision support in their implementation, and systems that use spatially referenced data belong to the class of intelligent systems with integrated spatially referenced data (ISRD).

In the past decade, remote sensing data have been the main source of new data about the Earth, which necessitates the creation of an information system with specialized services that allow scientists and specialists to perform thematic processing of remotely sensed data by changing the data processing parameters in a certain way and to analyse the obtained information independently [4]. At the same time, large crowdsourced geographic datasets have been generated about the Earth today as a result of the observed web phenomenon known as Volunteered Geographic Information (VGI) [7] through the development of spatial information systems and web mapping projects, the main ones being:

- Yandex search and information mapping service [2];
- a non-commercial web-mapping project to create a detailed free and free geographical map of the world OpenStreetMap (OSM) by the community of participants — Internet users [1];
- Google Maps is a set of applications built on top of the mapping service provided by Google [3].

The [8] argues that the growth of web services and applications for geographic information systems has made large archives of spatial data available over the Internet. Significant advances in GIS web service development technologies have resulted in several examples of mapping and graphics services that conform to web service standards and provide geospatial data and digital maps to enterprise developers. Thus, both government surveying and mapping services and private sector enterprises have recently experienced a surge in the development of web services and web-based applications for GIS, making large archives of spatial data available over the Internet.

In this regard, the role of the map as an image-sign geoinformation model of reality for quick and adequate perception of information is acquired. Creation of maps in electronic form, using GIS–technologies, is the most important task of modern society, because it is the map that becomes the tool with which a person can make a decision, from the simplest to the most complex, even in emergency situations. Accordingly, the society makes more and more demands to maps, the user, referring to the map, wants to receive reliable information and from a huge array of data to choose only the information that would be more suitable for making the right decision [9].

In addition, new and more sophisticated data collection technologies (knowledge bases based on wiki technologies, classifiers, natural language parsing, etc.) are now available. The large amount of accumulated geospatial data generated by Earth observation satellites as well as ground-based devices and sensors offers enormous potential to address global social issues related to natural disasters, health, transportation, energy and food security [10], [11]. Interoperability is particularly important as the level of cooperation between information sources at national, regional and local levels increases, requiring new methods to develop interoperable geographic systems [12]. Therefore, the use of terrain objects as integrating elements in information systems is essentially interdisciplinary in nature, as they integrate research in economics, ecology, climate forecasting, terrain development, formation of optimal routes, and more.

In the industry of geoinformation systems development nowadays there is a need in their intellectualization, i. e. in solving problems traditionally related to geoinformatics with the use of artificial intelligence methods. First of all, these are the tasks of intelligent search. Existing instrumental GIS, which are the means of development of applied GIS, do not solve the problems of intelligent search for a number of reasons, among which we will emphasize the following:

- practically all of them are based on internal (closed) formats of spatial data representation, and exchange open formats serve only as a means of map data exchange between different GIS tools;
- thematic data are mapped to specific spatial objects

and exclude the possibility of establishing links and relationships between such data;
- implementation of applied tasks of geoinformatics is carried out in internal programming languages, thus only simplifying access to spatial data, and the map serves as a means of visualization.

In the field of GIS development it is necessary to emphasize the problem of formation of cartographic images from information resources, for the solution of which the methods of dynamic representation of spatial data in GIS [13] are proposed, as well as the unsolved to date problem of information integration.

Thus, a group of international geographic and environmental scientists from government, industry, and academia brought together by the Vespucci Initiative for the Advancement of Geographic Information Science, and the Joint Research Centre of the European Commission [14], argue that despite significant progress, the ability to integrate geographic information from multiple sources is very limited and in order to facilitate such integration, an understanding of the statistical challenges of integration at different scales is needed, as well as the study of linguistic services

A mathematical model is proposed to facilitate the integration of spatial information and attribute data, which enabled the researcher to reduce the time to obtain data for management decision making in municipal services [15].

It should be noted that the need for information integration requires semantic geo-interoperability and harmonized understanding of the semantics of geodata [16]. Interoperability is an indicator of effective communication between systems [17]–[19].

On the other hand, known technologies of designing intelligent systems use cartographic materials, as a rule, in the form of raster images, i. e. there is no possibility to consider a map as a set of geographical objects with specified topological and subject-oriented (depending on the type of map) relations, while it is argued in the paper [9] that we need new maps, the content of which is supplemented with spatial knowledge, corresponding to the subject area for the preparation of spatial maps.

Besides, for a fixed territory the same objects of terrain and phenomena are used in different application areas: epidemiology, construction, environmental protection and nature management, land relations, etc., which determines the necessity to harmonize the ontology of subject areas with the objects of terrain and phenomena inherent in a given territory, thus providing a vertical (subject-oriented) level of GIS design.

Note that when designing a GIS for a new territory, the basic functional requirements are preserved and it is necessary to take into account not only the previous experience of GIS design, but also to use previously designed functional components, i. e. we are talking about the horizontal level of GIS design, when the territorial area is expanded and systems are designed for new territories.

The third aspect is the temporal component, relevant for retrospective analysis and modeling, thus providing a dynamic GIS that can deal with terrain objects and phenomena within a specific time period.

Currently proposed GIS tools have «weak» compatibility due to the lack of unification of subject knowledge with ontologies of terrain objects and phenomena, as well as with temporal components.

Known research on the integration of spatial data and domain knowledge to ensure semantic interoperability has been conducted for systems based on the Semantic Web technology stack RDF, RDFS, OWL and the Web Ontology Language OWL provides advanced capabilities for describing the subject areas of interacting systems and provides machine-interpretable definitions of fundamental concepts in the subject area and the relationships between such concepts in the ontology.

Recently, due to the development of Semantic Web technology, the key element of which is ontologies, it has become possible in GIS to emphasize the semantics of subject knowledge, to integrate and merge different datasets in related fields, to establish subject rules and their recording using RDF (Resource Description Framework) [20]–[22]. This capability certainly enhances the capabilities of GIS technologies. However, in order to do so, several important tasks must be solved. These are, first, justifying the use of tools to integrate spatial data and subject knowledge [23], and second, computing the similarity between geospatial objects that belong to different data sources [24]–[26].

For example, the paper [27] states that there are research problems related to the integration of different types of geographic information. The authors propose to base the GIS architecture on ontologies acting as a system integrator in order to ensure smooth and flexible integration of geographic information based on its semantic value. In this approach, the ontology system is a component, such as a database or knowledge base (in general case, an information component), interacting to achieve the goals of the geographic information system, and viewing the ontology, allows the user to obtain information about the existing (formalized) knowledge in the system. The use of several ontologies eventually allows to extract information at different stages of classification, i. e. for different types of information used for the purposes and in the interests of GIS. These ideas are developed in the works of [28]–[30].

The process of ontology development is called ontology engineering and according to the concept of ontology engineering, ontologies must be developed before they can be used in a GIS. Thus, a GIS is based on a subject area described initially by an ontology model, with ontologies acting as a tool for knowledge generation [31].

At present, scientific areas are developing so-called Smart-systems aimed at qualitative improvement of technical and economic indicators within the subject area. The application of geoinformation technologies for scientific research in the subject areas in conjunction with traditional tools, methods and models of artificial intelligence allow obtaining qualitatively new scientific results, as well as aimed at reducing the time of searching for acceptable solutions for the set tasks. At the same time, the authors pay special attention to the integration of terrain objects and data and knowledge in system research of a particular subject area.

Thus, Massel L. V. et al. proposed a methodical approach to the integration of remotely sensed earth observation data based on the methods of data and knowledge integration in energy system research [32], [33]. For this purpose, the authors developed a theoretical model of hybrid data based on the fractal stratified model (FS-model) of information space.

The hybrid data model is based on the development of a system of ontologies of the remote sensing information space, including a metaontology describing the layers of the FS model and ontologies of individual layers (subject areas).

As a result of ontological modeling, an ontological space including a set of ontologies is created, which should allow working not only with data, but also with knowledge, including descriptions of scenarios of various situations, models and software complexes, and integrate them into the IT infrastructure of interdisciplinary research.

The Open Geospatial Consortium (OGC) GeoSPARQL standard supports representing and querying geospatial data on the Semantic Web [34], [35]. GeoSPARQL defines a vocabulary for representing geospatial data in RDF, and it defines an extension to the SPARQL query language for processing geospatial data. In addition, GeoSPARQL is designed to accommodate systems based on qualitative spatial reasoning and systems based on quantitative spatial computations.

Thanks to Semantic Web technology and ontology engineering, as well as standardization processes for ontology development in the web ontology language, the problem of declarative knowledge representation has been solved, which contributes to the understanding of map objects and allows querying spatial data explicitly represented in spatial data storage formats [36], [37].

However, subject domain formalization and ontology engineering is only one step in intelligent systems design technology and by itself is not sufficient for knowledge-based inference, since ontology engineering allows for the description of declarative knowledge of subject domains, whereas procedural knowledge allows for the design of problem solvers and knowledge-based inference.

The above-mentioned possibilities of the technology based on the semantic web have certainly contributed to the development of geographic information systems with the ability to process colossal volumes of crowdsourced data. At the same time, decision making in problem domains of human activity requires obtaining an intelligent reference, i. e. actually solving a problem when the answer is not available in the datasets themselves or represented knowledge in the current version of the knowledge base or in the repository. A way of expressing such a need is the question [38]–[40]. In the process of communication there is always a context, which determines additional information that contributes to the correct understanding of the meaning of the message. Systems that are able to provide background information on the user's question belong to the class of intelligent help systems.

In intelligent reference systems, the problem is formulated in the form of a question, and the answer to the question requires specialized knowledge in science, technology, art, craft or other fields of activity, which is represented in knowledge bases. In other words, within the framework of the considered technologies it is necessary to first generate knowledge of the problem domain necessary for giving an answer. At the same time, the capabilities of knowledge bases of intelligent systems allow not only to represent and structure knowledge about the surrounding world, but also to quickly obtain and form this knowledge about it, thus satisfying the information need of the user [41].

One of the key features of an intelligent system is that the user has the ability to formulate his/her information need. The peculiarity of information representation in the knowledge bases of intelligent systems simplifies the formation of the user's information need, since the presented information in the knowledge bases is already structured and the relations defined on a certain concept, in respect of which the question-problem situation is solved, are known. In the work [42] it is shown that the question-problem situation cannot be solved within the framework of formal logic and the nature of the question can be understood in the system of subject-object relations. In connection with the fact that at formation of knowledge bases of intellectual systems the formation of subject-object relations within the given subject area takes place, thereby simplifying the expression of information need by the user by means of knowledge representation languages.

The proposed approaches to optimize information retrieval currently lie in the development of question-answer systems (QAS) in which user questions are matched with the required information. Such systems carry out a dialog between the user and the system in the form of the procedure "QUESTION-Answer" in the mode when the user asks a question and the system answers [43], [44]. A clear advantage of question-and-answer systems is the possibility of linguistic processing of user questions [40], [45]. At the same time, semantic classification of question-answer texts contributes to the isolation of specific types of relations, question types and answer classes [45]–[48]. The conceptual basis for formalizing questions in QAS is the question language and erotetic logic, which allows us to specify question-answer relations [39]. Currently known question-answer systems, which are capable of parsing a question and matching the answer with the help of a natural language analyzer, are Mulder [49], AllQuest (http://www.allquests.com) and AskNet Global Search (http://www.asknet.ru).

However, such systems are focused only on analyzing and detecting semantic relations between subject domain objects in indexed texts. This fact imposes the following limitations [50]:

- there is no possibility to strictly formally establish semantic relations between objects in the text;
- it is impossible to generate a response to the user when such a response does not exist in the indexed texts or in the current information state of the system;
- does not support questions to identify correspondences and analogies between objects and concepts.

The elimination of these limitations requires the creation of the next generation of question-and-answer systems — intelligent reference systems (IRS), or intelligent question-and-answer systems (IQAS). In such systems, the emphasis shifts from textual representation of information to the formation and use of knowledge spaces. The combination of work on the representation of knowledge in the IQAS knowledge base, the processing of this knowledge by special operations of the knowledge processing machine, and the interaction of the end user with the IQAS requires the coordination of all three stages of work. Thus, for the mass development of IQASs in various subject areas, a design technology for intelligent question-answering systems is needed in which, first, all design stages are coordinated, second, the languages of knowledge representation are compatible with the languages of knowledge processing and the languages of user communication with the IQAS

Despite the successes in the development of geoinformation services and their standardization, creation of ontologies of subject areas, knowledge about terrain objects and phenomena as integrating elements of subject areas, due attention has not been paid and not investigated semantic compatibility of GIS components and applied GIS, procedures for integration of spatial knowledge with knowledge of subject areas have not been established. In this regard, the actual task is:

- designing spatial ontologies and based on them solving the problem of integration of knowledge of subject areas and spatial relations, as well as solving the problem of metadata management and improving search, access and exchange in the conditions of growing volumes of spatial information and services provided by multiple sources of geoinformation;
- realizing knowledge inference using spatial and thematic information and meeting the information needs of users using question language;
- development of cartographic interface as a natural way for human to present information about terrain objects and phenomena, based on the formal description of the syntax of the map language, and providing both understanding of the semantics of terrain objects and phenomena immersed in knowledge bases, and providing changes in the state of knowledge bases.

Thus, the constant evolution of models and tools for ontological description of subject areas using spatial and temporal components, the heterogeneity of spatial components and the ambiguity of temporal components, poses new challenges in terms of interaction, integration and compatibility of different types of knowledge used in GIS by integrating subject area ontologies (vertical level), extending GIS subsystems at the horizontal level and utilizing new territories and time intervals

Analysis of human activities and works in the field of geoinformatics shows that further development in this area lies in the field of intellectualization of geographic information systems [51]–[58].

intelligence of systems with integrated spatially referenced data
⇒ *subdividing**:
{
- *use of digital cartographic material and Earth remote sensing data in problem-oriented areas, creation of systems for pattern and image recognition from Earth remote sensing data [4], [59], [60]*
- *planning of actions in a dynamically changing situation under incomplete or fuzzy data using expert knowledge [61]–[63]*
- *Analysis of emergency situations and preparation of materials for decision-making on prevention or elimination of their consequences, creation of expert systems for forecasting the occurrence and development on the ground of man-made and natural situations: floods, earthquakes, extreme weather conditions (precipitation, temperature), epidemics, spread of radionuclides, chemical emissions, meteorological forecast, etc. [64]*

- *creation of decision support systems for applied geoinformation systems of territorial planning and management [65]–[68]*
- *development of diagnostic expert systems for geological exploration and underground hydrodynamics [69]–[71]*
- *control systems of transportation and transportation processes [72]–[74]*
- *logistic planning, creation of expert systems and software tools for enterprise and building management [75]–[77]*
- *creation of monitoring, control and navigation systems [78]*
- *creation of expert systems and software tools for geodata analysis [79]*
- *resolution of land disputes [80]*
- *medico-geographical assessment of environmental impact on human health [81]*
- *retrospective analysis of events and inventory of cultural heritage [82]–[86]*
- *creation of digital cartographic information banks with remote access to them, spatial information management based on spatial data portals, creation of information retrieval systems on Earth sciences and geoinformatics [87]–[90]*
- *development of support systems for pedagogical, educational and training activities, as well as training systems using spatial information [91]–[97]*

}

The intellectualisation of geographic information systems implies [51], [98]:

- the possibility of end-user communication with the system in the language of questions [99];
- the use of various interoperable problem solvers with the possibility of explaining the obtained solutions;
- use of cartographic interface for visualisation of initial data and results.

Realising the capabilities of intelligent reference geographic information systems can be done by:

- knowledge base management systems,
- intelligent search,
- interoperable problem solvers,
- intelligent map interfaces,
- expert systems in various fields of human activity,
- decision support systems,
- intelligent assistance systems

Full solution of the above tasks requires the use of open systems standards, the use of ontologies of terrain objects as integrating elements of different subject areas, communication of users with the system in the mode of question-and-answer system using the language of questions.

The technology that satisfies these requirements is the open complex technology for developing intelligent systems based on semantic networks [100] (OSTIS technology — Open Semantic Technology for Intelligent Systems), the main provisions and principles of which are described in the work [101], the principles of creation and unified design models in the works [102], [103].

OSTIS technology is based on the following principles [104]:

- orientation on semantic unambiguous representation of knowledge in the form of semantic networks having basic theoretical-multiple interpretation, which provides problems of diversity of forms of representation of the same meaning, and problems of ambiguity of semantic interpretation of information constructions;
- the use of associative graph-dynamic model of memory;
- application of agent-based model of knowledge processing;
- realisation of OSTIS technology in the form of intellectual Metasystem IMS [5], which itself is built on OSTIS technology and provides design support for computer systems developed on OSTIS technology;
- ensuring in the designed systems a high level of flexibility, stratification, reflexivity, hybridity, interoperability and, as a consequence, learnability.

Systems built on this technology are called ostis-systems, and the universal abstract language of semantic networks (SC-code) or semantic code is used as a language tool for knowledge representation. In this case, knowledge bases of ostis-systems have a semantic representation, and the knowledge and skills interpreter is a collection of agents that process the knowledge base and manage situations and events in this knowledge base [104].

At the same time, the systems developed using this technology do not have the disadvantages of systems based on generative models (systems like ChatGPT [105]), because it is not the generation of new data, which are similar to the training data, but the relationship between the actual data and knowledge of the subject area is established, which ensures the reliability of conclusions based on knowledge.

In order to realise these possibilities, reduce the labour intensity of building and modifying intelligent systems with integrated spatially related data, it is proposed to build a semantic model of ISRD based on spatial ontologies, to ensure communication between users of ISRD in a formal language of questions and to develop means of automation and information support of the design process of this class of systems, including the formation of components of the core of intelligent systems with integrated spatially related data.

An important point that reduces the development time of intelligent systems with integrated spatially referenced data on the one hand, and increases their functionality on the other hand, is the availability of tools for designing such systems. At the same time, the technology of designing intelligent systems with integrated spatially referenced data should be oriented to multiple use of functional components of the system in order to reduce the design and development time of application systems. Thus, this study is about the creation of automation and information support for the design of intelligent systems with integrated spatially referenced data.

The structure of the study is presented in Figure 1. The first (upper) level corresponds to the data and knowledge level, where the integration of subject matter knowledge with spatial data and knowledge takes place; the second (middle) level is the system level, which corresponds to the principles underlying the proposed approach; the third (lower) level is the application level and corresponds to the application systems developed on the basis of the models and tools proposed in the paper.

The models, software tools and means of automation and information support of design proposed in this paper are proposed to be developed as a part of the open complex technology for the development of intelligent systems based on semantic networks OSTIS.

Using the notation adopted in the design of systems using OSTIS technology, let us clarify the concept of an intelligent system with integrated spatially referenced data (synonymous with intelligent geographic information system).

intelligent system with integrated spatially referenced data
:= [intelligent geographic information system]
:= [an information system designed to provide answers to a user's question, the main object of study of which is knowledge and data on terrain objects, acting as an integration basis for solving applied tasks in various subject areas]
⊃ *intelligent ostis-system with integrated spatially referenced data*
 ⊂ *ostis-system*
  := [an intelligent system with integrated spatially referenced data, developed according to the principles of OSTIS technology]
  ⇒ *subdividing**:
  {• *knowledge base of intelligent ostis-system with integrated spatially-referenced data*
  • *problem solver for intelligent ostis-system with integrated spatially referenced data*
  • *map interface of intelligent ostis-system with integrated spatially referenced data*
  }

III. Semantic model of an intelligent system with integrated spatially referenced data

This work is based on the method of knowledge representation in the form of a homogeneous semantic network with basic theoretical-multiple interpretation, which allows not only to describe declarative knowledge of subject areas, but also procedures of their processing, i. e. we are talking about procedural knowledge, the interpretation of which is carried out in a special component — semantic problem solver.

Therefore, in this paper, the concept of an interconnected system of coordinated semantic models of intelligent systems proposed in the works [102], [103], [106]–[109] is further developed in a more complete realisation of a special class of intelligent systems - intelligent systems with integrated spatially referenced data. For this purpose it is necessary to form the core ISRD, which is based on the semantic model of this class of systems. Thus, it is necessary to describe the semantic model of the knowledge base of an intelligent system with inegrated spatial data, which includes the construction of geontology and necessitates the development of a model of integration of subject areas with spatial components of geoinformation systems, the semantic model of the ISRD problem solver and the semantic model of the ISRD user interface [98], [110]. In this case, the user interface is not a viewer of cartographic (spatial) data, but a component interfaced with the knowledge base, i. e. knowledge about space understood by the system with simultaneous updating of knowledge both in the knowledge base of the system and in the knowledge base of the user interface.

In order to ensure semantic compatibility, an ontology of spatial objects has been developed, the task of which is to clearly and unambiguously define the semantics of terrain objects and phenomena.

The ontology of spatial objects consists of two sections. The first section includes directly the ontology of terrain objects and phenomena. The second section contains semantic characteristics of such objects — special elements specifying spatial and semantic characteristics of terrain objects and phenomena.

In the ontology of terrain objects and phenomena the objects of classification are terrain objects and phenomena to which map objects correspond, as well as features (characteristics) of these objects.

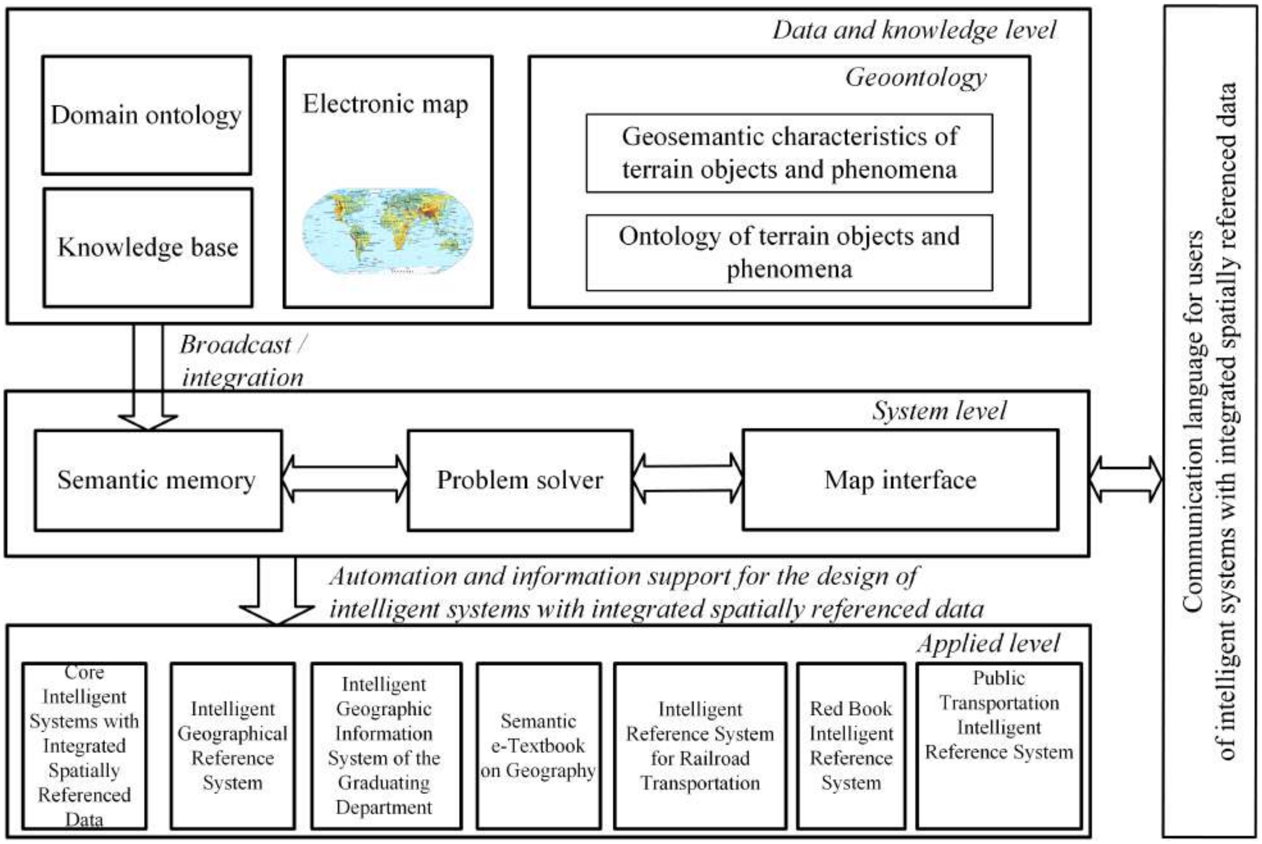


Figure 1. Structure of the study.

This section of the spatial object ontology presents classes of geospatial concepts of natural or artificial origin, natural phenomena that have common attributes (semantic attributes) characteristic of a certain *class of terrain objects* and describe the internal characteristics of the concept. Thus, *terrain objects* and *classes of terrain objects* are intended for different purposes. *Terrain objects* as physical elements themselves are formed in the knowledge base according to the specifications given in the ontology of terrain objects and phenomena. Accordingly, the knowledge base will store directly knowledge about a particular object, while in the terrain object ontology the terrain object is a concept and properties and relations defined on all objects of this type are established for it.

The characteristics of terrain objects are defined by the following types of relationships:

- *coordinate location of a terrain object* — geographic position, the location of a terrain object or phenomenon that is given in a geodetic coordinate system;
- *spatial relations* — a class of relations specifying the semantic properties of a terrain object in relation to other terrain objects and includes: *topological spatial relations*, *spatial ordering relation*, *relationship of principal directions of terrain objects*, *metric spatial relation*;
- *dynamics of the state of the terrain object* — a relation characterising the change of the state of the object in time.

spatial relations
⇒ *subdividing**:
{• *topological spatial relations*
• *spatial ordering relations*
• *metric spatial relation*
}

In order to determine the mutual spatial location of terrain objects depending on their type of localisation, basic *topological spatial relations* are established between instances of *terrain objects*:

- *inclusion**,
- *boundary**,
- *intersection**
- *adjacent**

Figure 2 shows schematically the possible variants of the *inclusion** relation with respect to area objects of the terrain.

To determine the spatial ordering of terrain objects,

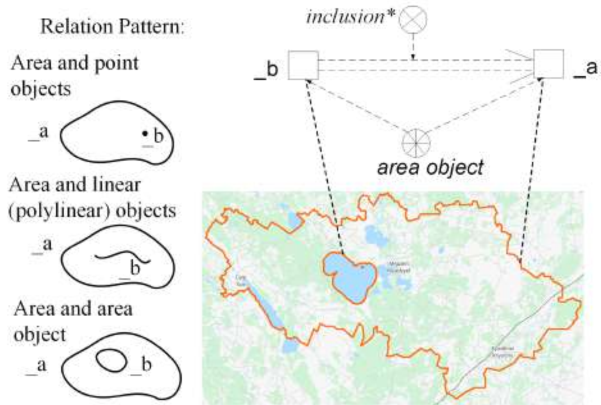


Figure 2. Establishing a topological relation "inclusion*".

relationship of location of terrain objects and *relationship of main directions of terrain objects* are introduced, which allow to determine the mutual location of terrain objects, as well as to determine the geographical direction of one terrain object relative to another object.

In order to convey the dimensions of the distances between *terrain objects*, a *metric spatial relation* is introduced, which characterises the distance information between *terrain objects* and can be measured in units (kilometre, metre, etc.) or have a *scale metric spatial relation* (0-100 m., 100-500 m., etc.).

Terrain objects, instances of which are stored in the knowledge base, are an integration framework for different subject areas, allowing the same terrain objects to be used for different application tasks as shown in the figure 3.

In order to integrate knowledge of subject areas with spatial components of intelligent systems, a *stratified model of information space of terrain objects* is proposed, which is defined by a family of the following sets according to equation (1).

$$S^{\mu} = \{S_{\mathrm{SA}\mu}, S_{\mathrm{TO}}, E_{\mathrm{TO}}\}, \mu \in I, \tag{1}$$

I – set of *subject areas*; $S_{\mathrm{SA}\mu}$ – ontology of the μ-th *subject area*; S_{TO} – ontology of *terrain objects*; E_{TO} – instances of *terrain objects*.

The layer of terrain object instances is an integrating layer with the knowledge of different subject areas in which the instantiated physical terrain objects and phenomena are already directly used. With this organisation of knowledge it is possible to repeatedly use the developed ontology of terrain objects and phenomena in different subject areas and, consequently, to solve different application problems.

The generalised structure of an intelligent system with integrated spatially referenced data is shown in Figure 4.

The problem solver is a collection of agents, which includes the following main agents: *computing geometric characteristics of terrain objects*; *determining the type of localisation of a terrain object*; *interfacing with various map systems and services, measurement systems and time intervals*; *establishing topological relations between terrain objects*; *verification of terrain objects knowledge base* for completeness of filling of knowledge bases with necessary terrain objects, correctness of values of semantic attributes of terrain objects entered into the knowledge base, correctness of characteristics of terrain objects entered or stored in the knowledge base.

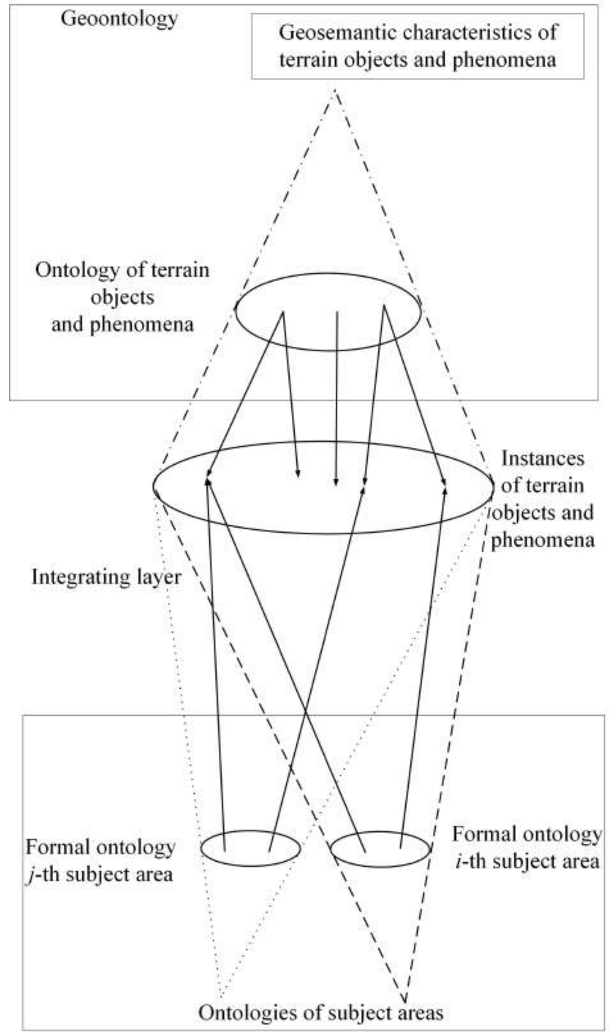


Figure 3. Integration of subject areas based on terrain objects.

The cartographic interface of an intelligent system with integrated spatially referenced data is considered as a particular type of user interface designed for visual display of terrain objects and phenomena in a formal map language. This language belongs to the family of semantic compatible languages and is intended for formal description of objects of terrain and phenomena, as well as relations between them in the systems of the intellectual system with integrated spatially related data.

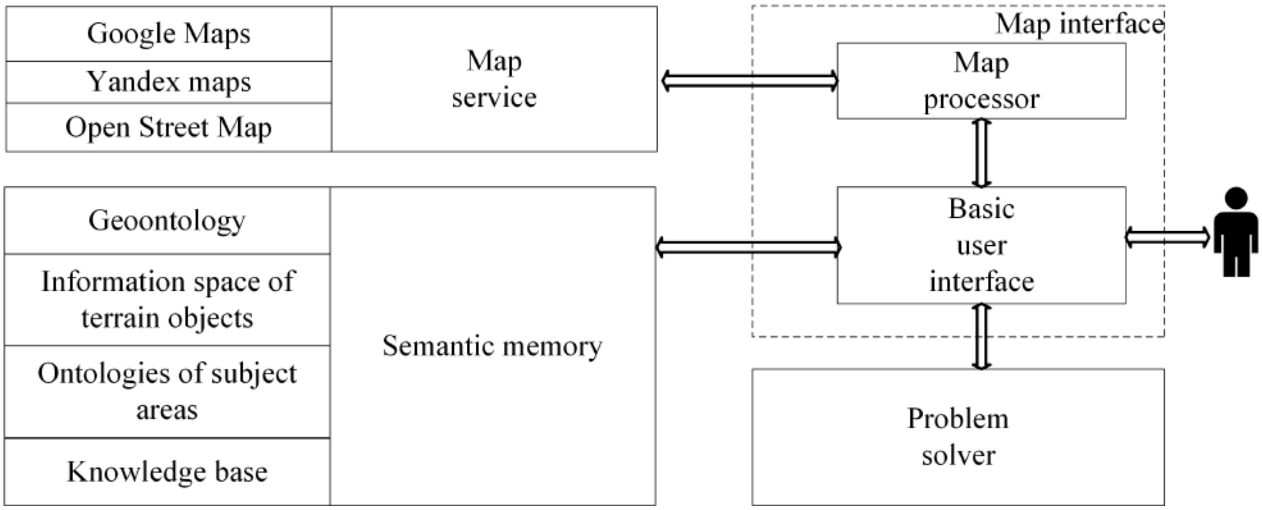


Figure 4. Structure of an intelligent system with integrated spatially referenced data.

It allows: to use minimum means for interpretation of given objects of terrain and phenomena on the map; to use the language of questions for intelligent systems; to reduce the search on the most part of given questions to the search of information in the current state of the knowledge base of the intelligent system with integrated spatially-referenced data.

IV. Language tools for communication between users of intelligent systems with integrated spatially referenced data

One of the key features of an intelligent system is that the user has the ability to articulate his or her information need. The way of expressing such a need is a question. The peculiarity of information representation in the knowledge bases of intelligent systems simplifies the formation of the user's information need, since the information presented in the knowledge bases is already structured and the relations defined on a certain concept, in relation to which the question-problem situation is resolved, are known. At formation of knowledge bases of intellectual systems the formation of subject-object relations within the given subject area takes place, thus simplifying the expression of information need by the user by means of semantic code [111].

The purpose of the question language and its subsequent development is to make it possible to understand the actions performed by the ISRD when forming an answer to a question. In the process of forming an answer to a question, the following options are possible:

1) the answer to the posed question exists in the knowledge base and a fragment of the knowledge base is localised in the context of the user's information need expressed by means of semantic code;
2) the answer is related to the resolution of some task situation, which is contained in the context of the question and the formation of the answer to the question is assigned to the problem solver.

In this regard, an intelligent system with integrated spatially referenced data is defined by the following set of systems according to equation (2).

$$S_{\text{IRSD}} = \{S_{\text{IIRS}}, S_{\text{AG}}\}, \tag{2}$$

S_{IRSD} – intelligent system with integrated spatially referenced data; S_{IIRS} – intelligent information retrieval system; S_{AG} – answer generation subsystem.

The intelligent information retrieval system is represented by a tuple of the following form (3).

$$S_{\text{IIRS}} : \{\{Q\}, \{A\}, \{F\}, \{UI\}\}, \tag{3}$$

S_{IIRS} – intelligent information retrieval system; $\{Q\}$ – set of questions; $\{A\}$ – the set of responses available in the current state of the system; $\{F\}$ – a set of operations of a problem solver that search for and generate answers to users' questions; $\{UI\}$ — a set of ways to visualise responses to the user.

The denotational semantics of a question language includes the classes of questions and the corresponding classes of answers needed to specify the wording of questions and their answers, as well as the classes of signs and relations that make up the structure of any question.

The following relations are introduced to systematise question types:

- *relationship within a given question* — a particular relationship between subject matter characters in the context of a question;
- *basic relation within a given question* — a class of relations uniting the relations in a given question, reflecting the same meaning and revealing a

certain feature of the signs of the subject area (state relation, action relation, composition relation, set-theoretic relation, temporal relation, spatial relation, quantitative relation, qualitative relation);
- *composition relation within a given question* — a stable combination of two action relations: *action directed to the question parameter* and *action directed to the answer to the question*.

Any question in the question language is a specification of an action to search for or synthesise knowledge that satisfies the information need of the user initiating the question. That is, a question is nothing but a task that expresses the user's need for some information, possibly stored or output in the knowledge base of an intelligent system. All the agents outputting answers to the question form a collective of agents and constitute a question language interpreter for intelligent reference systems with integrated spatially correlated data. Where each class of questions corresponds to certain agents realising the search or synthesis from the ISRD knowledge base of appropriate answers to the questions posed.

V. Design automation tools for intelligent systems with integrated spatially referenced data

We will refer to the means of automating the design of intelligent systems with integrated spatially referenced data as those that allow to improve the quality of designed systems and reduce the development time due to the repetition and compatibility of used components [112], [113]. In this case, we will consider as reusable components both fragments of knowledge bases, as well as problem solver agents and user interface components. The minimum set of components necessary for designing ISRD and based on them applied intelligent geoinformation reference systems will be called the ISRD core, which includes [114]:

1) Knowledge base components:
   - ontology of terrain objects and phenomena;
2) Problem solver components:
   - stack of mapping agents:
     – agents for calculating geometric features, localisation type agents,
     – agents for interfacing with different map systems and services, measurement systems and time intervals, topological linking agents;
   - search agents,
   - sophisticated search agents, Logical inference agents,
   - transaction calculation agents;
3) User interface components:
   - map interface:
     – work with point objects, linear and polylinear objects, area objects;
     – map viewer.

The formation of the knowledge base of the application ISDS is carried out in stages as shown in Figure 5.

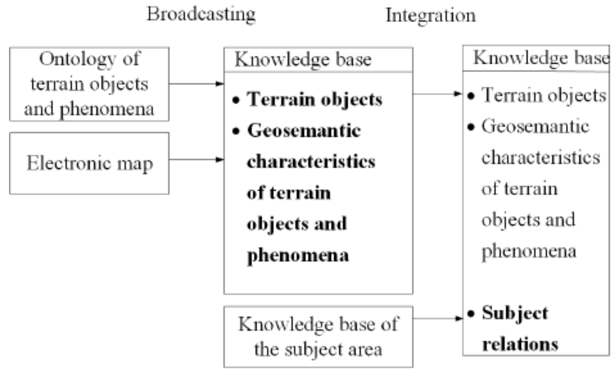


Figure 5. Formation of knowledge base of applied intelligent system with integrated spatially related data.

The first stage of formation of knowledge base of the applied intellectual system with integrated spatially related data is selection of data presented in the form of electronic map and translation into the knowledge base of terrain objects for a given territory. This stage is necessary for selection of terrain objects of a given territory in the interests of the application system. At this stage it is determined to which class the investigated terrain object belongs and, further, depending on the type of object, using the ontology of terrain objects and phenomena proposed in the paper, a fragment of the knowledge base corresponding to a particular physical terrain object in the semantic memory is formed. Thus, specific terrain objects are loaded into the knowledge base. At this stage, topological relationships between terrain objects are established in the knowledge base as illustrated in the example in Figure 2.

The second stage of knowledge base formation is integration with knowledge bases of subject areas. At this stage, knowledge of related subject areas is added, thus making it possible to set interdisciplinary links. At the same stage, homonymy in the names of geographical objects (toponymy) is removed, which makes it possible to avoid collisions in the knowledge base, when different localities belonging to the class of settlements correspond to the same sign in the knowledge base.

The means of automation and information support of ISRD design also include verification of knowledge bases of terrain objects and phenomena. Due to the specifics of development and operation of ISRD, it is necessary to ensure constant control of correctness and correctness of replenishment of the knowledge base with new terrain objects and updating of existing knowledge, i. e. it is necessary to ensure verification of completeness of filling the knowledge base with necessary terrain objects, correctness of values of semantic attributes of terrain objects entered into the knowledge base, as well as correctness of characteristics of terrain objects entered or stored in the knowledge base.

In order to develop ISRD for various application tasks, the design methodology of this class of systems is included in the means of automation and information support of design.

Stage 1: Deployment of basic software and information support for the design and operation of systems designed in accordance with the open integrated technology for the development of intelligent systems based on semantic networks.

Stage 2: Deployment of the components of the core of intellectual systems with integrated spatially correlated data. Installation of the developed components of the ontology of spatial objects, problem solver, including the main agents for processing spatially related data, question language interpreter, as well as software components of the mapping interface is carried out. The use of ISRD core components allows to create application systems with minimal functional purpose.

Stage 3: Formation of the subject area ontology. Formation of the ontology of the subject area in the interests of which the application system is being developed.

Stage 4: Broadcasting and loading of cartographic material into the knowledge base. Cartographic material is selected for a given area, translated and loaded into the knowledge base with the establishment of topological relations between the objects of the area using the ontology of spatial objects.

Stage 5: Formation of fragments of knowledge bases of the subject area. Filling of the knowledge base with knowledge of the subject area is carried out.

Stage 6: Integration of subject area knowledge bases with spatial knowledge. At this stage, the components of knowledge bases obtained at stages 4 and 5 are interlinked.

Stage 7: Development of problem solver components. If necessary, development of additional agents necessary for solving problems of the subject area is performed.

Stage 8. Verification of the developed components. At this stage the knowledge bases are verified by special verification agents.

Stage 9. Debugging of components. At this stage the components are debugged and errors are corrected.

At stages 3-9 of the design, the ISRD developer can decide to return to any previous stage, which corresponds to rapid prototyping technology, when a prototype of the system with minimal functionality is created and the functionality is subsequently increased. In this case, the prototype with minimal functionality is obtained after the 2nd design stage, i. e. after the deployment of the ISRD core components.

In accordance with the models and means of representation, integration and processing of spatially referenced data proposed in the work, means of automation of design of intelligent systems with integrated spatially referenced data, applied intelligent reference geoinformation systems on geography of the Republic of Belarus, graduating department, on public and railway transport, as well as semantic electronic textbook on geography have been developed.

The availability of developed application systems and analysis of the process of development of the mentioned intelligent reference geoinformation systems, software tools for their design, as well as means of design automation of the mentioned class of systems proposed in the work allow to unify and universalize to a high degree various components of intelligent systems with integrated spatially referenced data.

Thus, to develop a prototype of an intelligent reference geographic information system in the interests of a certain subject area, which allows simple navigation through the knowledge base, a ready-made ontology of terrain objects and phenomena, a set of information search agents and a stack of cartographic agents included in the corresponding libraries, as well as components of the cartographic interface - map viewer and map editor — can be used without any additions.

Based on the data presented in the studies [108], [109], time estimation for the development of the stack of map agents and map interface, and taking into account the number of used elements (fragments of knowledge bases, agents and ready-made user interface components) and time for their development the duration of application systems development is calculated by (4)

$$t_{pj} = \sum_{i=1}^{k} t_{fi} \cdot N_i + \sum_{i=1}^{l} t_{ai} \cdot P_i + \sum_{i=1}^{m} t_{ui} \cdot O_i, \quad (4)$$

t_{pj} - duration of development of the *j*th system (min); t_{fi} - duration of development of fragments of the *i*-type BR (min); t_{ai} - duration of development of *i*-type agents (min); t_{ui} - duration of development of the user interface of the *i*-type (min); *k* - number of types of database fragments included in the *j* system (pcs.); *l* - number of agent types included in the system *j* (pcs.); *m* - number of user interface types included in the system *j* (pcs.); N_i - number of fragments of the database of the *i*-type in the *j*-type system (pcs.); P_i - number of agents of the *i*th type in the *j*th system (pcs.); O_i - number of user interfaces of the *i*-type in the *j*-th system (pcs.).

Table I shows the indicators characterizing the stages of application systems development.

The average percentage of borrowing of knowledge base fragments, problem solver components and user mapping interface for existing systems without taking into account the complexity of development is calculated according to (5).

$$P_{\mathrm{F}} = \frac{1}{k} \sum_{i=1}^{k} p_{\mathrm{F}i}, \quad (5)$$

Table I
Indicators characterizing the stages of application systems development

Application system indicators	***System*** [a]	***System*** [b]	***System*** [c]	***System*** [d]	***System*** [e]	***System*** [f]	***System*** [g]
Time spent on formation of knowledge base fragments, min.	28224	56565	31896	42075	32391	42408	35991
Time spent on formation of borrowed fragments of knowledge bases, min.	0	28224	28224	28224	28224	28224	28224
Share of development time of borrowed knowledge base fragments to the total number of knowledge base fragments, %	0	50	88	67	87	67	78
Time spent on development of task solver agents, min.	11700	13500	13740	14160	11700	10260	12180
Time spent on development of borrowed task solver agents, min.	0	7860	10860	13500	10260	10140	10620
The proportion of development time of borrowed agents to the total number of problem solver agents, %	0	58	79	95	88	99	87
Time spent on the development of the map interface, min.	12600	12600	12600	12600	12600	12600	12600
Time spent on development of borrowed components of the map interface, min.	0	12600	12600	12600	12600	12600	12600
The proportion of development time for borrowed components of the mapping interface, %	0	100	100	100	100	100	100
Time spent on system development, min.	51924	82665	58236	68835	56691	65268	60771
Time spent on development of borrowed system components to total development time, min.	0	48084	51084	53724	50484	50364	50844
The proportion of the development time of borrowed components to the total system development time, %	0	58	88	78	89	77	84

[a] Core Intelligent Systems with Integrated Spatially Referenced Data.
[b] Intelligent Geographical Reference System.
[c] Intelligent Geographic Information System of the Graduating Department.
[d] Semantic e-Textbook on Geography.
[e] Intelligent Reference System for Railroad Transportation.
[f] Red Book Intelligent Reference System.
[g] Public Transportation Intelligent Reference System.

p_{Fi} – the share of development time of borrowed F-type components (knowledge base fragments, problem solver components, user map interface) to the total number of F-type fragments of the i-th application system (%); k – the number of application systems using borrowed components (pcs.).

Substituting into the formula (5) the obtained experimental data from the table I we obtain the average percentage of borrowing fragments of knowledge bases, problem solver agents and map interface for existing systems without taking into account the complexity of their development:

$$P_{\text{KB}} = \frac{1}{6}(50 + 88 + 67 + 87 + 67 + 78) = 72\ \%,$$
$$P_{\text{PS}} = \frac{1}{6}(58 + 79 + 95 + 88 + 99 + 87) = 84\ \%,$$
$$P_{\text{PS}} = \frac{1}{6}(100+100+100+100+100+100) = 100\ \%.$$

The obtained experimental data show that the development of each next system is significantly simplified due to the use of ready universal components. The number of components borrowed from the core of intelligent systems with integrated spatially referenced data shows that the current version of library filling allows to compose a significant part of the system components from ready-made components. In the future it is planned to actively replenish the libraries with new components.

As time passes and the production of intelligent systems with integrated spatially referenced data evolves, the number of such components will increase, consequently leading to an even higher borrowing rate.

A pessimistic approach was used to determine the reduction in time costs due to borrowing at this point in the library's development by estimating the least amount of borrowing in the systems in place today.

According to the results obtained, the availability of automation tools can reduce the duration of development of the class of systems in question by at least 58 % (min (58 %; 88 %; 78 %; 89 %; 77 %; 84 %)) due to the borrowing of previously developed components for developed application systems, which confirms the effectiveness of the methodology used. The average percentage of borrowing for existing systems without taking into account the complexity of the development of knowledge base fragments is 72 %, problem solver components — 84 %, user mapping interface – 100 %.

Conclusion

Let us list the main points of this paper:

- The problems solved by systems with integrated spatially referenced data have been analysed and the relevance of designing this class of systems has been substantiated. Based on the systematisation of tasks solved by ISRD, it is shown that one of the directions of increasing the efficiency of information-computing means use is the intellectualisation of systems with integrated spatially related data. It assumes: integration of knowledge of subject areas with spatial data and knowledge; possibility of communication of the end user with the system in the language of questions; use of various semantically compatible problem solvers with the possibility of explanation of the obtained solutions; use of cartographic interface for realisation of communication of the user with ISRD.

- Using ontological engineering the problem of semantic compatibility of knowledge of subject areas is solved and the model of integration of subject knowledge with objects of terrain and phenomena is proposed. Formal means have been developed to provide description of terrain objects and phenomena in knowledge bases of intelligent systems with integrated spatially related data taking into account semantics of links between terrain objects and phenomena, as well as knowledge of subject areas.
- The semantic model of ISRD including semantic memory, problem solver and cartographic interface is proposed, which unlike the known architectures of systems with integrated spatially related data allows integrating in the knowledge base the objects of terrain and phenomena of a given territory, translated into the internal language of the knowledge base, and the knowledge of different subject areas. Based on the formal description of the map language syntax, the cartographic interface provides a natural way for humans to represent information about terrain objects and phenomena and allows understanding the semantics of terrain objects and phenomena loaded into knowledge bases, as well as recording changes in the state of the ISRD knowledge base.
- A semantic model of user communication has been developed, in which interaction is carried out in the language of questions, semantically compatible with the languages of knowledge representation and processing, and designed to formally describe the search prescription in order to meet the user's information need. This allowed: to unify the form of questions and knowledge representation, with the help of which answers to the questions are constructed; to reduce the formation of answers to most of the given questions to the search of information in the current state of the knowledge base; to reduce the time to search for an answer in the knowledge base by the time required to parse the user's information request. Unlike existing approaches based on natural language processors, it does not require semantic parsing of the question sentence.
- The means of automation and information support of the process of design of ISRD are offered and their efficiency in the development of applied intelligent reference geoinformation systems is estimated. It is shown that the availability of automation tools can reduce the duration of the development of intelligent reference geographic information system by at least 58 % due to the borrowing of previously developed components, which confirms the effectiveness of the methodology used. The average percentage of borrowing for existing systems without taking into account the complexity of development of knowledge base fragments is 72 %, problem solver components – 84 %, user mapping interface – 100 %.

Acknowledgment

The author would like to thank the scientific teams of the Department of Intelligent Information Technologies of the Belarusian State University of Informatics and Radioelectronics for their help and valuable comments.

References

[1] Open Street Map [electronic resource]. Mode of access: https://www.openstreetmap.org/. – Date of access: 26.01.2024.
[2] Yandeks karty [electronic resource]. Mode of access: https://yandex.by/maps/. – Date of access: 26.01.2024.
[3] Google Maps [Электронный ресурс]. Mode of access: https://www.google.by/maps/. – Date of access: 26.01.2024.
[4] S. A. Zolotoi, I. B. Strashko, D. S. Kotov, I. M. Nesterovich, V. V. Roubo, and K. I. Kostyuk, "Sovershenstvovanie instrumental'nykh sredstv obrabotki i analiza kosmicheskoi informatsii," *Informatika*, vol. 18, no. 3, pp. 106–114, 2021.
[5] Metasistema IMS.ostis [electronic resource]. Rezhim dostupa: http://ims.ostis.net. – Date of access: 27.01.2024.
[6] A. M. Berlyant, "Geoikonika – nauka o geoizobrazheniyakh," *Izvestiya Rossiiskoi akademii nauk. Seriya geograficheskaya*, vol. 1, pp. 121–127, 2005.
[7] A. Ballatore, M. Bertolotto, and D. Wilson, "Geographic Knowledge Extraction and Semantic Similarity in OpenStreetMap," *Knowledge and Information Systems*, vol. 37, pp. 61–81, 10 2013.
[8] S. Tu and M. Abdelguerfi, "Web Services for Geographic Information Systems," *Internet Computing, IEEE*, vol. 10, pp. 13 – 15, 10 2006, mode of access: https://www.computer.org/csdl/magazine/ic/2006/05/w5013/13rRUzpQPK8 – Date of access: 28.01.2024.
[9] S. S. Yankelevich and E. S. Antonov, "Kontseptsiya novogo vida kart, osnovannogo na znaniyakh," *Vestnik SGUGiT (Sibirskogo gosudarstvennogo universiteta geosistem i tekhnologii)*, vol. 24, no. 4, pp. 188–196, 2019.
[10] K. Janowicz, S. Gao, G. McKenzie, Y. Hu, and B. Bhaduri, "GeoAI: spatially explicit artificial intelligence techniques for geographic knowledge discovery and beyond," *International Journal of Geographical Information Science*, vol. 34, no. 4, pp. 625–636, 2020. [Online]. Available: https://www.tandfonline.com/doi/full/10.1080/13658816.2019.1684500
[11] B. Adams, M. Gahegan, P. Gupta, and R. Hosking, "Geographic Information Observatories for Supporting Science," *CEUR Workshop Proceedings*, vol. 1273, pp. 32–39, 01 2014.
[12] A. Buccella, A. Cechich, D. Gendarmi, F. Lanubile, G. Semeraro, and A. Colagrossi, "GeoMergeP: Geographic Information Integration through Enriched Ontology Matching," *New Generation Computing*, vol. 28, pp. 41–71, 01 2010.
[13] Orekhova, D. A., "Razrabotka i issledovanie metoda dinamicheskogo predstavleniya prostranstvennykh dannykh v geoinformatsionnykh sistemakh ," dis. ... k-ta tekhn. nauk : 05.25.05, Taganrog, 2013, 152 c.

[14] M. Craglia, M. Goodchild, A. Annoni, G. Câmara, G. Michael, W. Kuhn, D. Mark, I. Masser, M. David, L. Steve, and E. Parsons, "Next-generation Digital Earth. a position paper from the Vespucci Initiative for the Advancement of Geographic Information Science," *International Journal of Spatial Data Infrastructures Research*, vol. 3, pp. 146–167, 01 2008.

[15] M. S. Sokolov, "Model' i algoritmy integrirovannoi obrabotki i analiza prostranstvennoi i atributivnoi informatsii v munitsipal'nykh GIS dlya podderzhki prinyatiya upravlencheskikh reshenii ," dis. ... k-ta tekhn. nauk : 05.13.01, Murom, 2012, 154 c.

[16] S. K. Dulin, N. G. Dulina, and D. A. Nikishin, "Problemy obespecheniya semanticheskoi geointeroperabel'nosti i soglasovaniya ponimaniya semantiki geodannykh," *Sistemy i sredstva informatiki*, vol. 26, no. 1, pp. 86–108, 2016.

[17] D. D. Kulikov, E. I. Yablochnikov, and A. I. Vostropyatov, "Interoperabel'nost' sistem tekhnologicheskoi podgotovki proizvodstva [elektronnyi resurs]," in *Internet i sovremennoe obshchestvo (IMS- 2018) : Trudy XXI Mezhdunarodnoi ob"edinennoi nauchnoi konferentsii «Internet i sovremennoe obshchestvo», Sankt-Peterburg, 30 maya – 2 iyunya 2018 g.*, Universitet ITMO. Sankt-Peterburg, 2018, pp. 11–13.

[18] I. S. Mikhailov, "Matematicheskoe i programmnoe obespechenie strukturnoi i semanticheskoi interoperabel'nosti informatsionnykh sistem na osnove metamodelei ," dis. ... k-ta tekhn. nauk : 05.13.11, Moskva, 2008, 235 s.

[19] D. Bernini, D. Micucci, and F. Tisato, "A Space-Based Interoperability Model," *Lecture Notes in Business Information Processing*, vol. 62, pp. 75–89, 06 2010.

[20] M. Brodie, "Semantic Technologies: Realizing the Services Vision," *IEEE Intelligent Systems*, vol. 22, pp. 12–17, 09 2007.

[21] G. Lan and Q. Huang, "Ontology-based Method for Geospatial Web Services Discovery," *International Journal of Computational Intelligence Systems*, 10 2007, mode of access: https://www.semanticscholar.org/paper/Ontology-based-Method-for-Geospatial-Web-Services-Lan-Huang/e44ac33d2ba6b788c9e44b5393ac7a807a664081?p2df — Date of access: 01.02.2024.

[22] G. J. Nalepa and W. T. Furmańska, "Review of Semantic Web Technologies for GIS," *AUTOMATYKA*, vol. 2, pp. 485–492, jun 2013,
Mode of access: https://web.archive.org/web/20150701211302id_/http://journals.bg.agh.edu.pl:80/AUTOMATYKA/2009-02/Auto36.pdf – Date of access: 20.03.2024.

[23] L. M. Vilches-Blazquez and J. Saavedra, "A framework for connecting two interoperability universes: OGC Web Feature Services and Linked Data," *Transactions in GIS*, 11 2018, mode of access: https://onlinelibrary.wiley.com/doi/abs/10.1111/tgis.12496 — Date of access: 20.03.2024.

[24] Y. Zhang, P. Yang, C. Li, G. Zhang, C. Wang, H. He, X. Hu, and Z. Guan, "A Multi-Feature Based Automatic Approach to Geospatial Record Linking," *International Journal on Semantic Web and Information Systems*, vol. 14, pp. 73–91, 10 2018,
Mode of access: https://www.igi-global.com/article/a-multi-feature-based-automatic-approach-to-geospatial-record-linking/210653 — Date of access: 12.03.2024.

[25] Y. Zhang, C. Li, N. Chen, S. Liu, L. Du, Z. Wang, and M. Ma, *Semantic-Based Geospatial Data Integration With Unique Features: Concepts, Methodologies, Tools, and Applications*, 01 2019, pp. 254–277, mode of access: https://www.researchgate.net/publication/330919971_Semantic-Based_Geospatial_Data_Integration_With_Unique_Features_Concepts_Methodologies_Tools_and_Applications – Date of access: 20.03.2024.

[26] G. Mai, K. Janowicz, P. Sathya, and Y. Bo, "Visualizing The Semantic Similarity of Geographic Features," 2018, mode of access: https://people.geog.ucsb.edu/~gengchen_mai/papers/2018-AGILE_Semantic_Map.pdf — Date of access: 18.03.2024.

[27] F. T. Fonseca, M. J. Egenhofer, P. Agouris, and G. Câmara, "Using Ontologies for Integrated Geographic Information Systems," *Transactions in GIS*, vol. 6, no. 3, pp. 231–257, 2002.

[28] T. Bittner, "Formal ontology of space, time, and physical entities in Classical Mechanics," *Applied Ontology*, vol. 13, pp. 135–179, 05 2018.

[29] T. Bittner, M. Donnelly, and B. Smith, "A Spatio-Temporal Ontology for Geographic Information Integration," *International Journal of Geographical Information Science*, vol. 23, pp. 765–798, 06 2009.

[30] A. Diosteanu and L. Cotfas, "Agent Based Knowledge Management Solution Using Ontology, Semantic Web Services and GIS," *Informatica Economica*, vol. 13, no. 4, p. 90, 2009.

[31] V. Y. Tsvetkov, *Geoinformatika znanii. Geoinformatika*. LAP Lambert Academic Publishing GmbH & Co, 2020, 157 c.

[32] L. V. Massel', T. N. Vorozhtsova, and N. I. Pyatkova, "Ontologicheskii inzhiniring dlya podderzhki prinyatiya strategicheskikh reshenii v energetike," *Ontologiya proektirovaniya*, vol. 7, no. 1(23), pp. 66–76, 2017.

[33] L. V. Massel' and A. N. Kopaigorodskii, "Metodicheskii podkhod k integratsii dannykh distantsionnogo zondirovaniya zemli na osnove metodov integratsii dannykh i znanii v sistemnykh issledovaniyakh energeti," in *Obrabotka prostranstvennykh dannykh i distantsionnyi monitoring prirodnoi sredy i masshtabnykh antropogennykh protsessov (DRPS'2013) : trudy konf., Barnaul, 30 sent. – 04 okt. 2013 g.*, Ros. akad. nauk, Sib. otdelenie, In-t vodnykh i ekologicheskikh. problem ; redkol.: N.N. Dobretsov [i dr.]. Barnaul, 2013, pp. 4–13.

[34] R. Battle and D. Kolas, "GeoSPARQL: Enabling a Geospatial Semantic Web," *Semantic Web Journal*, vol. 3, no. 4, pp. 355–370, 2011.

[35] "A Geographic Query Language for RDF Data [electronic resource]," mode of access: https://www.ogc.org/standard/geosparql/. — Date of access: 12.01.2024.

[36] "Distantsionnoe zondirovanie Zemli iz kosmosa. Dannye distantsionnogo zondirovaniya Zemli iz kosmosa i produkty ikh obrabotki. Ontologiya. Chast' 1. Struktura : STB ISO/TS 19150-1-2021 – Vved. 10.02.21." Minsk: Gosstandart, 2021.

[37] "Distantsionnoe zondirovanie Zemli iz kosmosa. Dannye distantsionnogo zondirovaniya Zemli iz kosmosa i produkty ikh obrabotki. Ontologiya. Chast' 2. Pravila razrabotki ontologii na yazyke veb-ontologii : STB ISO/OP 19150-2-2022 – Vved. 01.12.22." Minsk: Gosstandart, 2023, 80 s.

[38] V. K. Finn, "K formal'nomu opredeleniyu ponyatiya informatsionno-poiskovoi sistemy," *NTI, ser 2*, pp. 5–15, 1981.

[39] N. Belnap and T. Stil, *Logika voprosov i otvetov*. Progress, 1981.

[40] V. K. Finn, *Logicheskie problemy informatsionnogo poiska*. Nauka, 1976.

[41] S. A. Samodumkin, N. V. Zotov, D. V. Shunkevich, and V. P. Ivashenko, "Yazyk voprosov dlya ostis-sistem," *Tekhnologiya kompleksnoi podderzhki zhiznennogo tsikla semanticheski sovmestimykh intellektual'nykh komp'yuternykh sistem novogo pokoleniya*, pp. 359–370, 2023.

[42] L. Y. Aver'yanov, "Pochemu lyudi zadayut voprosy," *M.: Sotsiolog*, p. 152, 1993.

[43] P. Sosnin, "Question-answer means for collaborative development of software intensive systems," in *Complex Systems Concurrent Engineering: Collaboration, Technology Innovation and Sustainability*. Springer, 2007, pp. 151–158.

[44] V. P. Zakharov, *Informatsionnye sistemy (dokumental'nyi poisk)*. Federal'noe gosudarstvennoe byudzhetnoe obrazovatel'noe uchrezhdenie vysshego, 2002, 188 s.

[45] D. Suleimanov, "Issledovanie bazovykh printsipov postroeniya semanticheskogo interpretatora voprosno-otvetnykh tekstov na estestvennom yazyke v aos," *Obrazovatel'nye tekhnologii i obshchestvo*, vol. 4, no. 3, pp. 178–192, 2001.

[46] ——, "Dvukhurovnevyi lingvisticheskii protsessor otvetnykh tekstov na estestvennom yazyke," in *Otkrytye semanticheskie tekhnologii proektirovaniya intellektual'nykh sistem (OSTIS-2011) : materialy Mezhdunar. nauch.-tekhn. konf., Minsk, 10–12 fevr. 2013 g.*, Belorus. gos. un-t informatiki i radioelektroniki ; redkol.: V. V. Golenkov (otv. red.) [i dr.]. Minsk, 2011, pp. 311–322.

[47] R. Bukharaev and D. Suleimanov, *Semanticheskii analiz v voprosno-otvetnykh sistemakh*. Kazan' : Izd-vo Kazan. un-ta, 1990.

[48] D. Suleimanov, "Sistema semanticheskogo analiza otvetnykh tekstov obuchaemogo na estestvennom yazyke," *Ontologiya proektirovaniya*, no. 1 (11), pp. 65–77, 2014.

[49] C. C. Kwok, O. Etzioni, and D. S. Weld, "Scaling question answering to the web," in *Proceedings of the 10th international conference on World Wide Web*, 2001, pp. 150–161.

[50] S. Samodumkin, "Semanticheskaya tekhnologiya proektirovaniya intellektual'nykh voprosno-otvetnykh sistem," *Doklady BGUIR*, no. 7(45), pp. 67–73, 2009.

[51] S. A. Samodumkin and N. V. Zotov, "Podsistema ekosistemy ostis, obespechivayushchaya podderzhku zhiznennogo tsikla intellektual'nykh geoinformatsionnykh sistem razlichnogo naznacheniya," *Tekhnologiya kompleksnoi podderzhki zhiznennogo tsikla semanticheski sovmestimykh intellektual'nykh komp'yuternykh sistem novogo pokoleniya*, pp. 831–848, 2023.

[52] S. L. Belyakov and I. N. Rozenberg, "Intellektual'nye geoinformatsionnye sistemy." *Zheleznodorozhnyi transport*, no. 4, pp. 32–37, 2011.

[53] A. A. Glotov, "Intellektualizatsiya geoinformatsionnykh sistem: podkhody i napravleniya," *GEOMATIKA*, no. 4, pp. 18–24, 2015.

[54] A. Berezko, A. Rybkina, A. Solov'ev, and R. Krasnoperov, "Intellektual'naya gis [elektronnyi resurs]," *Vestnik ONZ RAN*, vol. 1, 2003, rezhim dostupa: https://onznews.wdcb.ru/publications/v01/2009NZ000006.pdf. — Data dostupa: 24.04.2019.

[55] T. V. Safonova, "Obzor tekhnologii sozdaniya intellektual'nykh geoinformatsionnykh sistem," *Informatsionnye tekhnologii i sistemy: upravlenie, ekonomika, transport, pravo*, no. 3, pp. 18–27, 2020.

[56] Y. Shokin and V. Potapov, "Gis segodnya: sostoyanie, perspektivy, resheniya," *Vychislitel'nye tekhnologii*, no. 20, pp. 175–213, 2015.

[57] V. A. Buchkin, "Sostoyanie i razvitie intellektual'nykh gis," *Informatsiya i kosmos*, no. 3, pp. 119–123, 2020.

[58] M. Shpakov, "Razrabotka intellektual'nykh geoinformatsionnykh na osnove nastraivaemoi ob"ektnoi modeli predmetnoi oblasti ," dis. ... k-ta tekhn. nauk : 05.13.11, Sankt-Peterburg, 2004, 168 l.

[59] S. V. Ablameiko and D. M. Lagunovskii, *Obrabotka izobrazhenii: tekhnologiya, metody, primenenie*. Amalfeya, 2000.

[60] A. M. Chandra and S. K. Gosh, "Distantsionnoe zondirovanie zemli i geograficheskie informatsionnye sistemy," *Moskva : Tekhnosfera*, 2008, 312 c.

[61] Y. A. Ivakin, "Metody intellektualizatsii promyshlennykh geoinformatsionnykh sistem na osnove ontologii ," dis. ... d-ra tekhn. nauk : 05.13.06, Sankt-Peterburg, 2009, 371 s.

[62] Y. Y. Lyubarskii, "Operativnyi dispetcherskii analiz neshtatnykh situatsii v elektricheskikh setyakh promyshlennykh predpriyatii. ispol'zovanie komp'yuternoi podderzhki na osnove tekhnologii ekspertnykh sistem," *Promyshlennaya energetika*, no. 9, pp. 2–6, 2017.

[63] ——, "Dispetcherskii analiz neshtatnykh situatsii v vysokovol'tnykh elektricheskikh setyakh. komp'yuternaya podderzhka," *Energiya edinoi seti*, no. 6, pp. 58–64, 2017.

[64] L. F. Nozhenkova, "Tekhnologiya postroeniya ekspertnykh geoinformatsionnykh sistem podderzhki prinyatiya reshenii po preduprezhdeniyu i likvidatsii chrezvychainykh situatsii ," dis. ... d-ra tekhn. nauk : 05.13.14, Krasnoyarsk, 2000, 365 s.

[65] M. L. Belyakova, *Intellektual'nye geoinformatsionnye sistemy dlya upravleniya infrastrukturoi transportnykh kompleksov*. Taganrog : Izdatel'stvo Yuzhnogo federal'nogo universiteta, 2016, 190 l. — Rezhim dostupa: po podpiske. - https://hub.sfedu.ru/repository/material/800749244. — Data dostupa: 01.03.2024.

[66] A. A. Filipenko, N. G. Sinyak, and S. A. Shavrov, "Kompleksnyi monitoring v sisteme ustoichivogo upravleniya territoriyami," *Trudy BGTU. Seriya 5: Ekonomika i upravlenie*, no. 7, pp. 163–167, 2015.

[67] J.-Y. Kang, J. Choi, and T. Bittner, "An Ontology-based Geospatial Database for Identifying Geographical Change," *Advancing Geographic Information Science: The Past and Next Twenty Years*, pp. 293–299, 2016, editors: Harlan Onsrud, Werner Kuhn.

[68] A. G. Litvinov, "Tekhnologiya razrabotki intellektual'nykh geoinformatsionnykh sistem gornopromyshlennykh kompleksov ," dis. ... k-ta tekhn. nauk : 25.00.35, Moskva, 2006, 152 c.

[69] V. B. Taranchuk, "Komp'yuternye modeli podzemnoi gidrodinamiki." Minsk : BGU, 2020, 235 s.

[70] M. A. Zhuravkov and A.-M. K. Rashid, "Avtomatizirovannaya komp'yuternaya sistema sopryazhennogo geoekologicheskogo monitoringa dlya otvetstvennykh regionov (s intensivnym tekhnogennym vozdeistviem i/ili unikal'nymi prirodnymi usloviyami): Chast' 3. inzheneriya znanii. sistema baz znanii i dannykh geomonitoringovoi sistemy," *Geoinformatika*, no. 1, pp. 11–21, 2003.

[71] V. V. Vidyakin, M. A. Zhuravkov, and O. L. Konovalov, *GIS-tekhnologii pri dobyche poleznykh iskopaemykh. Spetsializirovannaya korporativnaya geoinformatsionnaya sistema "MapManager"*. Minsk : Izd. tsentr BGU, 2004, 208 s. – Pod obshch. red. M.A. Zhuravkova.

[72] A. A. Erofeev, *Intellektual'naya sistema upravleniya perevozochnym protsessom na zheleznodorozhnom transporte*, Gomel' , 2022, 407 s.

[73] T. I. Mikheeva, O. K. Golovnin, and A. A. Fedoseev, "Intellektual'naya geoinformatsionnaya platforma issledovaniya transportnykh protsessov," in *Sbornik trudov III mezhdunarodnoi konferentsii i molodezhnoi shkoly «Informatsionnye tekhnologii i nanotekhnologii»(ITNT-2017)-Samara: Novaya tekhnika*, 2017, pp. 753–761.

[74] I. I. Poznyak, A. A. Kholuyanov, and M. A. Losenkov, "Geoinformatsionnye sistemy i iskusstvennyi intellekt na transporte," *Slavyanskii forum*, no. 3, pp. 444–453, 2022.

[75] A. F. Tuzovskii, "Ontologo-semanticheskie modeli v korporativnykh sistemakh upravleniya znaniyami ," dis. ... d-ra tekhn. nauk : 05.13.01, Tomsk, 2007, 373 s.

[76] S. Shavrov, "Kadastrovaya model' zdaniya sovmestnogo domovladeniya," *Trudy BGTU. Seriya 5: Ekonomika i upravlenie*, no. 7 (189), pp. 182–186, 2016.

[77] A.-H. Hor, "A Semantic web platform for BIM-GIS integration [Electronic resource]," mode of access: https://www.researchgate.net/publication/305775768_A_semantic_Web_platform_for_BIM-GIS_integration. – Date of access: 20.03.2024.

[78] V. V. Popovich, S. N. Potapychev, A. V.Pan'kin, S. S. Shaida, and M. N.Voronin, "Intellektual'naya gis v sistemakh monitoringa," *Trudy SPIIRAN*, vol. 1, no. 3, pp. 172–184, 1 2006, Rezhim dostupa: http://proceedings.spiiras.nw.ru/index.php/sp/article/view/1226/1090. — Data dostupa: 11.03.2024.

[79] V. Taranchuk, "New computer technologies, analysis and interpretation of geodata," in *VI International Scientific Conference "Integration, Partnership and Innovation in Construction Science and Education" (IPICSE-2018)*, Moscow, Russia, November 14-16, 2018. https://doi.org/10.1051/matecconf/201825104059: 14 December 2018, nov 2018.

[80] "Kodeks Respubliki Belarus' o zemle," in *Kodeks Respubliki Belarus' ot 23.07.2008 N 425-Z (red. ot 18.07.2022)*.

[81] A. A. Glotov, "Primenenie geoinformatsionnykh tekhnologii dlya formirovaniya sistem podderzhki prinyatiya reshenii v oblasti zdravookhraneniya [elektronnyi resurs]," *GEOMATIKA*, no. 3, pp. 38–42, 2014, rezhim dostupa: https://sovzond.ru/upload/iblock/d34/3_2014_38-42.pdf. – Data dostupa: 15.02.2024.

[82] A. V. Gubarevich, O. L. Morosin, and D. V. Lande, "Ontologicheskoe proektirovanie intellektual'nykh sistem v oblasti istorii," in *Otkrytye semanticheskie tekhnologii proektirovaniya intellektual'nykh sistem*, ser. Vyp. 1. Minsk : BGUIR, 2017, pp. 245–250.

[83] A. V. Gubarevich, S. P. Vityaz', and R. B. Grigyanets, "Struktura baz znanii v intellektual'nykh sistemakh po istorii," in *Otkrytye semanticheskie tekhnologii proektirovaniya intellektual'nykh sistem*, ser. Vyp. 2. Minsk : BGUIR, 2018, pp. 347–350.

[84] R. N. Batalov and L. K. Radchenko, "Geoprostranstvennye informatsiya, dannye, znaniya v istoriko-kartograficheskikh

issledovaniyakh," *Regulirovanie zemel'no-imushchestvennykh otnoshenii v Rossii: pravovoe i geoprostranstvennoe obespechenie, otsenka nedvizhimosti, ekologiya, tekhnologicheskie resheniya*, no. 1, pp. 112–118, 2021.

[85] T. G. Nikiforova, R. L. Grigor'eva, and P. E. Grigor'ev, "Ispol'zovanie servisa «gugl-karta» dlya postroeniya boevogo puti uchastnikov velikoi otechestvennoi voiny," in *Informatsionnye tekhnologii. Avtomatizatsiya. Aktualizatsiya i reshenie problem podgotovki vysokokvalifitsirovannykh kadrov (ITAP-2020)*, 2020, pp. 577–582.

[86] T. Kauppinen, J. Väätäinen, and E. Hyvönen, "Creating and Using Geospatial Ontology Time Series in a Semantic Cultural Heritage Portal," in *The Semantic Web: Research and Applications: 5th European Semantic Web Conference, ESWC 2008, Tenerife, Canary Islands, Spain, June 1-5, 2008*, 06 2008, pp. 110–123, mode of access: https://link.springer.com/chapter/10.1007/978-3-540-68234-9_11 — Date of access: 20.03.2024.

[87] O. M. Ataeva, K. A. Kuznetsov, V. A. Serebryakov, and V. I. Filippov, "Sreda integratsii prostranstvennykh dannykh «GeoMeta»," in *Internet i sovremennoe obshchestvo (IMS-2011) : trudy XIV Vserossiiskoi ob"edinennoi konferentsii , Sankt-Peterburg, Rossiya, 2011.* Sankt-Peterburg, 2011, pp. 11–16, rezhim dostupa: http://ojs.itmo.ru/index.php/IMS/article/view/142/142 — Data dostupa: 12.03.2024.

[88] A. A. Bliskavitskii, "Semantika geoprostranstvennykh ob"ektov, funktsional'naya grammatika i intellektual'nye gis," in *Izvestiya vysshikh uchebnykh zavedenii. Geologiya i razvedka*, no. 2, 2014, pp. 62–69.

[89] L. Di, P. Zhao, W. Yang, and P. Yue, "Ontology-driven Automatic Geospatial-Processing Modeling based on Web-service Chaining," in *Proceedings of the sixth annual NASA earth science technology conference. College Park, MD, USA*, jun 2006, p. 7 p., mode of access: https://www.semanticscholar.org/paper/Ontology-driven-Automatic-Geospatial-Processing-on-Di-Zhao/5c25c4e32585540e3039771100e921efb86b9092?p2df – Date of access: 20.02.2024.

[90] A. Sayar, M. Pierce, and G. Fox, "Developing gis visualization web services for geophysical applications," in *ISPRS International Society for Photogrammetry and Remote Sensing Workshop Commission II WG/2 METU, Ankara, Turkey, November, 24-25*, 07 2005, mode of access: https://www.researchgate.net/publication/265167317_Developing_GIS_Visualization_Web_Services_for_Geophysical_Applications – Date of access: 20.03.2024.

[91] K. Meteshkin and M. Sokolova, "Osobennosti predstavleniya geodannykh i znanii na stranitsakh Web-saitov uchebnogo naznacheniya," *Vostochno-Evropeiskii zhurnal peredovykh tekhnologii*, vol. 2, no. 2 (56), pp. 63–65, 2012.

[92] V. B. Taranchuk, "Wolfram mathematica – sredstva i tekhnologii razrabotki intellektual'nykh obuchayushchikh sistem," in *Otkrytye semanticheskie tekhnologii proektirovaniya intellektual'nykh sistem (OSTIS-2015) : materialy V Mezhdunar. nauch.-tekhn. konf., Minsk, 19–21 fevr. 2015 g.*, Belorus. gos. un-t informatiki i radioelektroniki ; redkol.: V. V. Golenkov (otv. red.) [i dr.]. Minsk, 2015, pp. 339–346.

[93] S. A. Samodumkin, "Predstavlenie dannykh i znanii v geoinformatsionnoi sisteme kafedry," in *Novye informatsionnyn tekhnologii : materialy V Mezhdunar. nauch. konf., Minsk, 29-31 okt. 2002 g.*, Belorusskii gosudarstvennyi ekonomicheskii universitet. Minsk, 2002, pp. 312–316.

[94] N. A. Gulyakina, S. S. A., E. I. Zabella, and M. A. B., "Ispol'zovanie geoinformatsionnykh tekhnologii dlya organizatsii i kontrolya uchebnogo protsessa," in *Distantsionnoe obuchenie-obrazovatel'naya sreda KhKhI veka : materialy II Mezhdunar. nauch.-metod. konf., Minsk, 26-28 noyab. 2002 g.*, Belorusskii gosudarstvennyi universitet informatiki i radioelektroniki. Minsk, 2002, pp. 356–359.

[95] S. A. Samodumkin, "Geoinformatsionnaya sistema dlya organizatsii i kontrolya uchebnogo protsessa," in *Pervyi Belorusskii kosmicheskii kongress : materialy kongressa, Minsk, 28-30 oktyabrya 2003 g.*, Ob"edinennyi institut problem informatiki Natsional'noi akademii nauk Belarusi. Minsk, 2003, pp. 229–231.

[96] ——, "Podderzhka prinyatiya reshenii v geoinformatsionnoi sisteme kafedry," in *Distantsionnoe obuchenie-obrazovatel'naya sreda KhKhI veka : materialy III Mezhdunar. nauch.-metod. konf., 13-15 noyab. 2003 g.*, Belorusskii gosudarstvennyi universitet informatiki i radioelektroniki. Minsk, 2003, pp. 378–381.

[97] ——, "Semanticheskii elektronnyi uchebnik po geografii belarusi," in *Distantsionnoe obuchenie-obrazovatel'naya sreda KhKhI veka : materialy V Mezhdunar. nauch.-metod. konf., 10-11 noyab. 2005 g.*, Belorusskii gosudarstvennyi universitet informatiki i radioelektroniki. Minsk, 2003, pp. 181–184.

[98] S. Samodumkin, "Support of the life cycle of intelligent geoinformation systems for various purposes," *Open Semantic Technologies for Intelligent Systems*, no. (7), pp. 299–304, 2023.

[99] Y. Y. Lyubarskii, *Intellektual'nye informatsionnye sistemy*. Moskva, 1990, 232 s.

[100] "Tekhnologiya kompleksnoi podderzhki zhiznennogo tsikla semanticheski sovmestimykh intellektual'nykh komp'yuternykh sistem novogo pokoleniya / pod red. V.V. Golenkova — Minsk : Bestprint," 2023, 1064 s.

[101] V. V. Golenkov, N. A. Gulyakina, and D. V. Shunkevich, "Otkrytaya tekhnologiya ontologicheskogo proektirovaniya, proizvodstva i ekspluatatsii semanticheski sovmestimykh gibridnykh intellektual'nykh komp'yuternykh sistem," Minsk : Bestprint, 2021, 690 s.

[102] V. V. Golenkov and N. A. Gulyakina, "Proekt otkrytoi semanticheskoi tekhnologii komponentnogo proektirovaniya intellektual'nykh sistem. chast' 1: Printsipy sozdaniya," *Ontologiya proektirovaniya*, no. 1, pp. 42–64, 2014.

[103] ——, "Proekt otkrytoi semanticheskoi tekhnologii komponentnogo proektirovaniya intellektual'nykh sistem. chast' 2: Unifitsirovannye modeli proektirovaniya," *Ontologiya proektirovaniya*, no. 4, pp. 34–53, 2014.

[104] V. Golenkov, N. Gulyakina, I. Davydenko, and D. Shunkevich, "Semanticheskie tekhnologii proektirovaniya intellektual'nykh sistem i semanticheskie assotsiativnye komp'yutery," *Doklady Belorusskogo gosudarstvennogo universiteta informatiki i radioelektroniki*, no. 3 (121), pp. 42–50, 2019.

[105] ChatGpt [Electronic resource]. – Mode of access:https://openai.com/chatgpt/. — Date of access: 20.03.2024.

[106] V. V. Golenkov, "Grafodinamicheskie modeli i metody parallel'noi asinkhronnoi pererabotki informatsii v intellektual'nykh sistemakh ," dis. ... d-ra tekhn. nauk : 05.13.11 ; 05.13.17, Minsk, 1996, 396 l.

[107] V. V. Golenkov and N. A. Gulyakina, "Grafodinamicheskie modeli parallel'noi obrabotki znanii: printsipy postroeniya, realizatsii i proektirovaniya," in *Otkrytye semanticheskie tekhnologii proektirovaniya intellektual'nykh sistem (OSTIS-2012) : materialy II Mezhdunar. nauch.-tekhn. konf., Minsk, 16–18 fevr. 2012 g.*, Belorus. gos. un-t informatiki i radioelektroniki ; redkol.: V. V. Golenkov (otv. red.) [i dr.]. Minsk, 2012, pp. 23–52.

[108] D. Shunkevich, "Agentno-orientirovannye reshateli zadach intellektual'nykh sistem ," dis. ... k-ta tekhn. nauk : 05.13.17, Minsk, 2017, 254 l.

[109] I. Davydenko, "Modeli, metodika i sredstva razrabotki gibridnykh baz znanii na osnove semanticheskoi sovmestimosti mnogokratno ispol'zuemykh komponentov ," dis. ... k-ta tekhn. nauk : 05.13.17, Minsk, 2017, 284 l.

[110] S. Samodumkin, "Next-generation intelligent geoinformation systems," *Open Semantic Technologies for Intelligent Systems*, no. 6, pp. 411–416, 2022.

[111] ——, "Implementation of information needs of ostis-systems users," *Open Semantic Technologies for Intelligent Systems*, no. (7), pp. 141–146, 2023.

[112] A. N. Borisov, "Postroenie intellektual'nykh sistem, osnovannykh na znaniyakh, s povtornym ispol'zovaniem komponentov," in *Otkrytye semanticheskie tekhnologii proektirovaniya intellektual'nykh sistem (OSTIS-2014) : materialy IV Mezhdunar. nauch.-tekhn. konf., Minsk, 20–22 fevr. 2014 g.*, Belorus. gos. un-t informatiki i radioelektroniki

; redkol.: V. V. Golenkov (otv. red.) [i dr.]. Minsk, 2014, pp. 97–102.
[113] I. T. Davydenko, “Semanticheskaya model’ kollektivnogo proektirovaniya baz znanii,” in *Otkrytye semanticheskie tekhnologii proektirovaniya intellektual’nykh sistem (OSTIS-2016) : materialy VI Mezhdunar. nauch.-tekhn. konf., Minsk, 18–20 fevr. 2016 g.*, Belorus. gos. un-t informatiki i radioelektroniki ; redkol.: V. V. Golenkov (otv. red.) [i dr.]. Minsk, 2016, pp. 107–114.
[114] “Modelirovanie rassuzhdenii kognitivnogo agenta na osnove neklassicheskikh logik: otchet o nir (zaklyuch.) / belor. gos. univ. inf. i radioel.; ruk. temy n.a. gulyakina. — minsk, 2022. — 50 s. — № gr 20201308.”

ПРОЕКТИРОВАНИЕ ИНТЕЛЛЕКТУАЛЬНЫХ СИСТЕМ С ИНТЕГРИРОВАННЫМИ ПРОСТРАНСТВЕННО-СООТНЕСЕННЫМИ ДАННЫМИ

Самодумкин С. А.

Работа посвящена вопросам представления, интеграции и обработки пространственно соотнесенных данных в интеллектуальных системах и проектированию на основе предложенных моделей и средств прикладных интеллектуальных справочных геоинформационных систем, которые относятся к классу интеллектуальных систем с интегрированными пространственными данными (ИСПД). Основные положения данной работы:

- Проведен анализ задач, решаемых системами с интегрированными пространственно-соотнесенными данными, и обоснована актуальность проектирования данного класса систем. На основе систематизации задач, показано, что одним из направлений повышения эффективности использования информационно-вычислительных средств является интеллектуализация систем с интегрированными пространственно-соотнесенными данными, которая предполагает: интеграцию знаний предметных областей с пространственными данными и знаниями; возможность общения конечного пользователя с системой на языке вопросов; использование различных семантически совместимых решателей задач с возможностью объяснения полученных решений; использование картографического интерфейса для реализации общения пользователя с системой.
- С использованием онтологического инжиниринга решена проблема семантической совместимости знаний предметных областей и предложена модель интеграции предметных знаний с объектами местности и явлений. Разработаны формальные средства, обеспечивающие описание объектов местности и явлений в базах знаний ИСПД с учетом семантики связей между объектами местности и явлений, а также знаниями предметных областей.
- Предложена семантическая модель ИСПД, включающая семантическую память, решатель задач и картографический интерфейс, что в отличие от известных архитектур систем с интегрированными пространственно-соотнесенными данными позволяет интегрировать в базе знаний объекты местности и явления заданной территории, транслированные на внутренний язык базы знаний, и знания различных предметных областей. Основанный на формальном описании синтаксиса языка карт картографический интерфейс обеспечивает естественный для человека способ представления информации об объектах местности и явлениях и позволяет понимать семантику объектов местности и явлений, загруженных в базы знаний, а также фиксировать изменение состояния базы знаний ИСПД.
- Разработана семантическая модель общения пользователей ИСПД, в которой взаимодействие осуществляется на языке вопросов, семантически совместимом с языками представления и обработки знаний, и предназначенном для формального описания поискового предписания с целью удовлетворения информационной потребности пользователя, что позволило: унифицировать форму представления вопросов и знаний, с помощью которых строятся ответы на поставленные вопросы; свести формирование ответов на большую часть заданных вопросов к поиску информации в текущем состоянии базы знаний; сократить время на поиск ответа в базе знаний на время, необходимое для разбора информационного запроса пользователя, и в отличие от существующих подходов на основе естественно-языковых процессоров не требует семантического разбора вопросительного предложения.
- Предложены средства автоматизации и информационной поддержки процесса проектирования ИСПД и оценена их эффективность при разработке прикладных интеллектуальных справочных геоинформационных систем. Показано, что наличие средств автоматизации позволяет сократить продолжительность разработки интеллектуальной справочной геоинформационной системы как минимум на 58 % за счет заимствования разработанных ранее компонентов, что подтверждает эффективность применения использованной методики. Средний процент заимствования для существующих систем без учета сложности разработки фрагментов баз знаний составляет 72 %, компонентов решателя задач — 84 %, пользовательского картографического интерфейса — 100 %.

Received 01.03.2024

OSTIS Ecosystem Security Problems

Vasili Khoroshavin, Vladimir Zakharov
Belarusian State University of
Informatics and Radioelectronics
Minsk, Belarus
Email: vasya.khoroshavin@gmail.com, zakharov@bsuir.by

***Abstract*—In this article the threats and vulnerabilities relevant for ostis-systems are examined. The differentiation of access to the knowledge bases of ostis-systems, the implementation of mechanisms of configuration of a personal ostis-assistant and the safety of agents' source code are defined as the main directions for ensuring security of ostis-systems. Options for implementing the according security mechanisms in those directions are proposed.**

***Keywords*—security, knowledge base, ostis-system, OSTIS Ecosystem, knowledge processing.**

I. Introduction

In the era of digitalization and the rapid growth of information technologies, the safety issues of information systems become more and more relevant and significant. The next-generation intelligent computer systems, operating with global knowledge bases and capable of autonomous decision-making, open new horizons for various spheres of human activity. From the financial sector and health care to space research — the potential for the use of such systems is unlimited. Even the cybersecurity area is not an exception [1], [2], [3], [4]. However, it is worth noting that safety requirements in different areas are not the same, which is confirmed by existing standards in the field of information security (for example, an evaluation assurance level [5]). Consequently, with the expansion of areas of the use of intelligent computer systems, the criticality of the implementation of potential threats of security increases as well, which makes the task of protecting them especially relevant and complex.

Modern intelligent systems, in contrast to traditional computer systems and neural networks, are a combination of global knowledge bases and problem solvers built using multi-agent architecture, which puts special requirements for ensuring security. The problem is not only in protecting data from unauthorized access, but also in ensuring the protection of the decision-making process itself. The aspects associated with the addition of a new level of processing — knowledge processing — must also be considered. This entails the necessity to take into account the new corresponding vulnerabilities and threats. Also, one of the features of the next-generation intelligent computer systems that should be noted is their interoperability [6], which entails the need to unite them into collectives to jointly solve problems, whereby the OSTIS Ecosystem is formed [6].

Currently the most attention is paid to machine learning methods. Despite they show good quality in several cases vulnerabilities that might be exploited are created as well [7]. So this article is focused on semantic aspects of AI and in this work as its main goal was set to consider threats and vulnerabilities that are relevant for ostis-systems and ostis-communities, as well as propose ways for solving certain security problems at the level of both individual ostis-systems and their collectives.

II. Intelligent systems security vulnerabilities

Based on the ostis-systems architecture there are three main origins of vulnerabilities: devices running the system; the knowledge base; communication channels and communicative processes; incorrect actions of users.

Since nowadays there is no hardware implementation of semantic computers and only software implementation is used, the vulnerabilities related to devices aren't considered and are beyond the scope of this paper. It is also worth noting that in contradiction to the traditional systems two aspects of protection in the knowledge bases can be distinguished: data and knowledge. This means that such classic problem as access to fragments of knowledge bases is not the only one that must be solved. The correctness of their contents and the influence of the process of solving certain problems on the emergence of new data in the knowledge base, the presence of which in the public domain is unacceptable, must also be taken into account.

It is easy to notice that the only new issues are those related to knowledge processing. For this reason, when compiling a vulnerabilities hierarchy, CWE (Common Weakness Enumeration), namely the CWE-1000 view [8], was taken as basis.

Taking into account the aforementioned necessity to consider the features of knowledge processing, an analysis of the specifics of working with them was carried out. According to [9], knowledge has the following traits: connectivity, complex structure, (internal) interpretability, activity and presence of semantic metrics. The activity of knowledge is the source of activity in the knowledge-driven system [10]. So one of the main challenges is handling non-factors of knowledge. In practice, in almost all cases, knowledge possesses certain non-factors. According to the classification given in [11], it is possible to decompose the set of non-factors N into two types: N_1 and N_2 ($N = N_1 \cup N_2$). The first one contains those, which might be received from experts automatically. These are fuzziness, uncertainty, inaccuracy and under-determination. non-factors of the second type, respectively, include those that arise in other ways (incompleteness, inconsistency, incorrectness, non-monotonicity). The second-type factors are characterized by their dynamic nature. For example, incorrectness of knowledge might easily flow into inconsistency, incompleteness can cause non-monotonicity and so on. The "cause of occurrence" dependency of second-type factors on the other factors might be described with the relation C (xCy means that x is the cause of y occurrence).

$$C \subset \{< x, y > |(x \in N) \wedge (y \in N_2) \wedge (x \neq y)\}$$

Notice that some of the mentioned factors can be detected automatically and their negative influence might be limited [12], [13]. Partly the solution of this problem depends on the knowledge engineers who must design knowledge bases taking into account the following requirements:

- origins of knowledge must be easily detected;
- uncertainty, inaccuracy and under-determination rates must be explicitly saved.

This thorough design and appropriate processing mechanisms will drastically increase knowledge bases security.

The main non-factors are the following:

1) Uncertainty is the factor preceding incompleteness. It is determined by the fact that this or that knowledge is set by a certain degree of confidence, which can have a complex nature.
2) Inaccuracy is a factor associated with the impossibility of accurately obtaining a particular value. For example, due to the error of measurement devices [14].
3) Under-determination is a factor that, unlike inaccuracy, is associated with the possibility of clarifying the value, but this is not necessary within the framework of a specific task [14].
4) Inconsistency is a state in which knowledge base contains fragments contradicting to each other.
5) Incompleteness is the absence of the elements in the knowledge base. The criteria for completeness in the knowledge base is defined by a set of formal statements about completeness [15].

So there are the following ostis-systems vulnerabilities related to the knowledge processing:

1) processing of uncertain knowledge as certain knowledge;
2) improper usage of inaccurate or under-determined knowledge;
3) certainty check mechanism failure (or its absence);
4) completeness check mechanism failure (or its absence);
5) inconsistency search mechanism failure (or its absence);
6) reliance on knowledge from untrusted sources;
7) knowledge verification violation during logical inference.

Top-level of formalized ontology of threats and vulnerabilities contains 11 basic classes of vulnerabilities. Their subclasses are more accurate vulnerabilities and provide the corresponding descriptions, so they might be used in practical evaluation of the systems.

ostis-system vulnerability
$\supset$ *improper access control*
$\supset$ *improper interaction between multiple correctly-behaving entities*
$\supset$ *improper control of a resource through its lifetime*
$\supset$ *incorrect calculation*
$\supset$ *insufficient control flow management*
$\supset$ *protection mechanism failure*
$\supset$ *incorrect comparison*
$\supset$ *improper check or handling of exceptional conditions*
$\supset$ *improper neutralization*
$\supset$ *improper adherence to coding standards*
$\supset$ *improper processing of knowledge non-factors*

According to [6] the OSTIS Ecosystem is a collective of interacting:

1) ostis-systems;
2) users of these ostis-systems (both end users and developers);
3) computer systems that are not ostis-systems, but considered by them as additional information resources or services.

Since systems that are not ostis-systems can also be actors in the Ecosystem, in the case of consideration of the entire Ecosystem, the CWE-1000 view with additions related to knowledge processing introduced into it will be completely inherited.

III. Security threats for the intelligent systems

Taking the goals of security assurance in the traditional systems given in [16] the following goals of security assurance in the next-generation intelligent systems might be defined:

- ensuring the safety of semantic compatibility of information;
- protection of the reliability and integrity of information;
- ensuring the availability of information at different levels of the intelligent system;
- minimizing damage from events that pose a threat to information security.

Given the specifics of the operation of individual ostis-systems and ostis-communities that exist within the OSTIS Ecosystem [6] and knowledge-driven systems in general [17], the following threats can be distinguished.

threat in the ostis-system
⊃ *violation of the confidentiality of information*
  := [unauthorized reading of information]
⊃ *violation of the integrity of information*
  := [unauthorized or erroneous changes, distortion or destruction of information, as well as unauthorized impact on technical and software tools for processing information]
⊃ *violation of accessibility*
  := [blocking access to the system, its individual components, functions or information, as well as the impossibility of timely obtaining information (unacceptable delays in obtaining information)]
⊃ *violation of semantic compatibility*
  := [violation of the commonality of concepts and commonality of basic knowledge]
⊃ *destruction of the semantics of knowledge bases*
  := [semantic virus]
  := [replacement or removal of nodes and connections between them in the knowledge base]
⊃ *excessive volume of incoming information*
⊃ *violation of non-repudiation*
  := [the issuance of unauthorized actions for legitimate, as well as concealment or substitution of information about the actions of the subjects]
⊃ *violation of accountability*
  := [unauthorized or erroneous change, distortion or destruction of information about the performance of actions by the subject]
⊃ *violation of authenticity*
  := [performing actions in the system on behalf of another person or issuing unreliable resources (including data) for genuine]
⊃ *violation of reliability*
  := [deliberate or unintentional provision and use of erroneous (incorrect) or irrelevant (at a particular moment in time) information, as well as the implementation of procedures in violation of the order (protocol)]

Thus, one can distinguish several main areas for the implementation of protection measures in the OSTIS Ecosystem and its components:

- providing mechanisms for access differentiation to knowledge bases;
- organization of safe mechanisms of communication and interaction of ostis-systems within the ostis-communities;
- implementation of mechanisms for protecting the interaction process of a personal ostis assistant and user;
- implementation of the mechanisms of automated verification of the source code of agents of ostis-systems.

IV. Potential solutions

A. Differentiation of access to knowledge base

When solving the problem of differentiation of access to ostis-system knowledge base there are two main sub-problems: access differentiation to the entire knowledge base and access differentiation to specific fragments of the knowledge base.

The main difficulty arising in the design of the ontology of the subject domain of users is that, firstly, it is necessary to guarantee quick access to access rights for a particular sc-element, and, secondly, minimize the number of additionally created sc-elements in the knowledge base.

The formation of structures in the knowledge base for which a special access policy will be configured entails the creation of additional sc-connectors. Those will not be isomorphic to the search template for access rights for a particular sc-element. For example, on the Fig. 1 those would be edges *edge 1*, *edge 2* and *edge 3*.

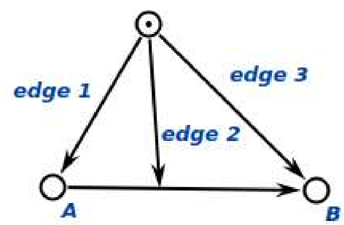


Figure 1: Example of sc-connectors that are not isomorphic to the search template for access rights for an sc-element.

Taking this into account it is necessary to implement a special mechanism of access rights detection for them. In order to do so the special bit in the sc-address of an element will be used. This bit set to 1 will indicate that the sc-element belongs to the class of sc-elements that can be edited only with administrator rights.

The current implementation of the ostis-platform uses linked lists for storing elements incident with a specific sc-element, so in order to decrease access time to the sc-node denoting the fragment of the knowledge base all those access edges must be placed in the beginning of the linked list. To represent and process access rights, the following concepts in the knowledge base are defined: ***user***, ***group of users***, ***user action class within sc-structure****, ***knowledge base reading***, ***knowledge base editing***, ***interpretation of an scp-program***.

The proposed scheme of access rights processing is described by the attribute based access control (ABAC) model. Users and groups of users act as subjects of impact in ostis-systems. As a result of the analysis of the classes of operations on sc-memory, we define 3 main classes of actions which will build the decomposition of the set of operations. These are **reading** (as a result of search in knowledge base), **writing** and **execution** of a program, represented in sc-code. An example of a fragment of the knowledge base, recorded in the SCg language, with the access rights specified for it is presented in the Fig. 2. The construction with the access right for *user 2* to edit all knowledge base is presented in the Fig. 3. The construction with the access right for *group 2* to edit a fragment of the knowledge base is presented in the Fig. 4.

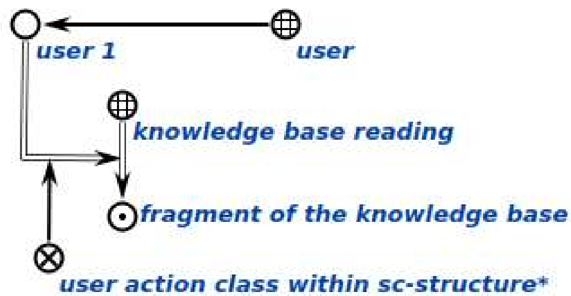


Figure 2: Formal representation of a user's right to edit the fragment of the knowledge base.

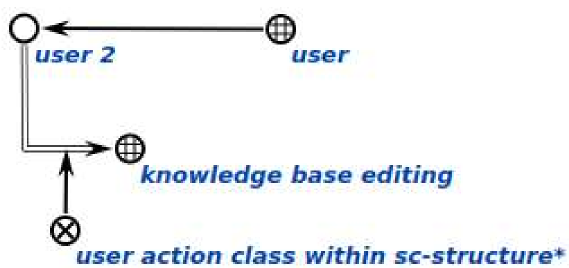


Figure 3: Formal representation of a user's right to edit all knowledge base.

B. Options for solving security problems of the OSTIS Ecosystem

One of the fundamental advantages of the ostis-systems is their high level of interoperability, which eases their ability to form collectives (ostis-communities) for cooperative solving of problems, as well as formation of the digital ecosystem, referred to as the OSTIS Ecosystem, on the basis of these communities [6]. So provision of secure functioning of the entire OSTIS Ecosystem is one of the priority directions for research.

Among the existing communication protocols, special attention should be paid to the Matrix protocol. It is a set

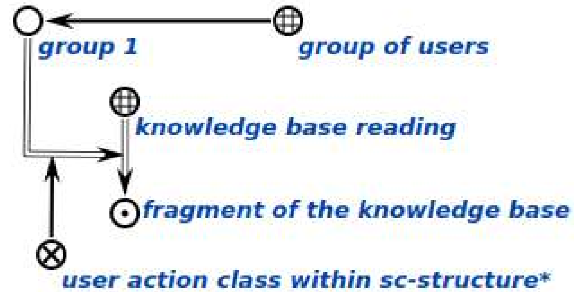


Figure 4: Formal representation of a group of users' right to edit the fragment of the knowledge base.

of open APIs for decentralised communication, suitable for securely publishing, persisting and subscribing to data over a global open federation of servers with no single point of control [18]. For communication in Matrix protocol virtual "rooms" are used. The local copies of their descriptions are stored on homeservers and are automatically synchronized between each other. Fig. 5 shows a schematic example of a room.

Nodes *@alice:alice.com*, *@bob:bob.com* and *@charlie:charlie.com* represent the clients of the end users and *matrix.alice.com*, *matrix.bob.com* и *matrix.charlie.com* represent the homeservers.

Based on the specification of this protocol, the following structural elements of the Ecosystem can be distinguished: ***ostis-community***, ***room***, ***storage***, ***homeserver***. The Fig. 6 shows a fragment of the description of the OSTIS Ecosystem.

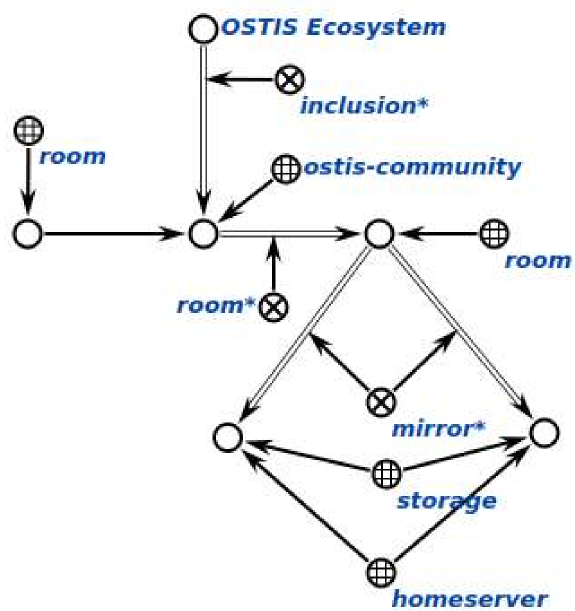


Figure 6: A fragment of the OSTIS Ecosystem structure.

Every community in the Ecosystem will create a

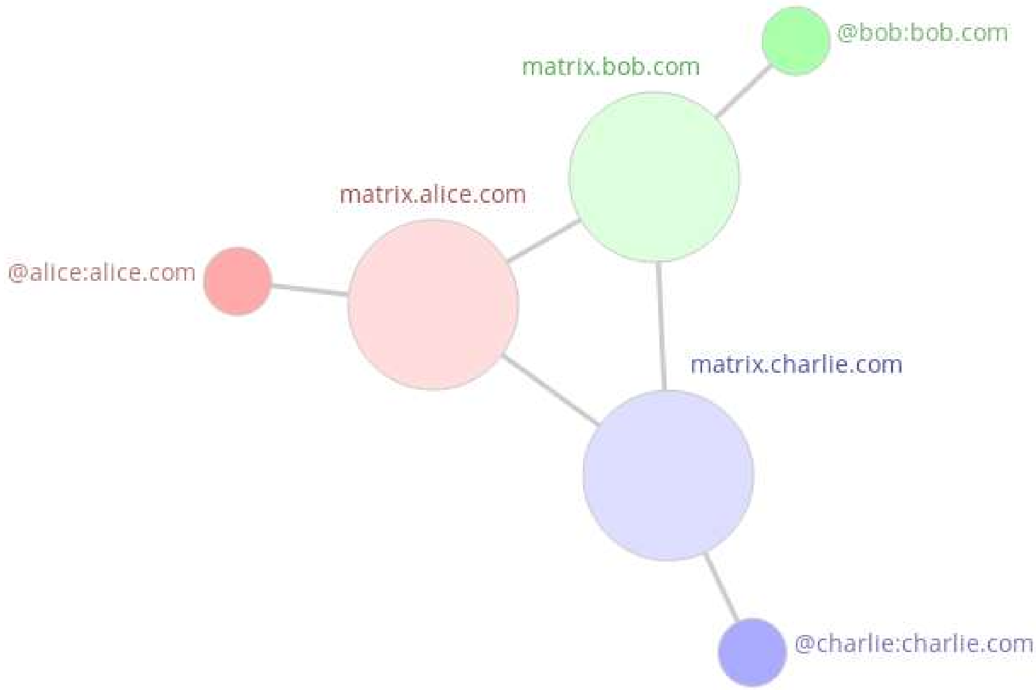


Figure 5: Structural scheme of a room in the Matrix protocol.

corresponding room. Any ostis-community can be a member of any number of rooms. Resources distributed by communities will be kept on a specialized servers. These are referred as storages and also will be used as homeservers. The given protocol might be used with enabled end-to-end encryption in order to implement private ostis-communities. This structure might be used to create the hierarchy of ostis-systems collectives in OSTIS Ecosystem by simultaneous use of the corporate ostis-systems as clients and homeservers.

It is also important to take into account the necessity of verification of the sources of fragments of knowledge bases (including agents) transmitted over the network within the Ecosystem. This task can be solved by usage of the existing digital signature protocols (for example, OpenPGP [19]).

C. Security of the personal ostis-assistant

A personal ostis-assistant is an ostis-system that provides comprehensive adaptive maintenance of a particular user on all issues related to his interaction with any other ostis-systems, as well as representing the interests of this user through the entire global network of ostis-systems [6]. Combination of a (human) user of the ostis-system and its corresponding personal assistant is a minimal ostis-community, where personal assistant takes a role of a ***corporate ostis-system**** of this community as shown on the Fig. 7.

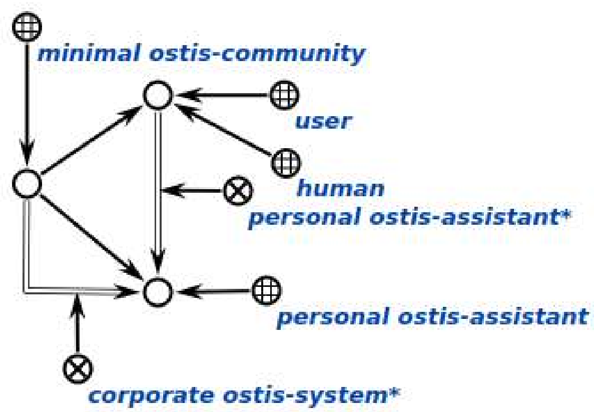


Figure 7: Example of minimal ostis-community members and their corresponding roles.

One of the main problems of ensuring the security of the personal ostis-assistant is the complex of the facts that the user is the main source of threats in the system (regardless of whether he is purposefully creating threats or not) and that the personal ostis-assistant is his representative, i.e. its main purpose is to be the user's bridge to the Ecosystem. It follows that when solving this problem, the assistant should pay special attention to the class of tasks that it is designed to solve, and limit the user from changing its main functions, while preserving its extensibility. To do this, it is necessary to create a separate fragment of the knowledge base of the personal ostis-assistant and limit it for direct access of the user for edit. Accordingly, the assistant must also have a certain set of agents which will be responsible for the verification of the user data and compute the corresponding degrees of trust. The problems of accounting for non-factors that arise in the knowledge base in the process of interacting with the user are extremely acute here. Given the non-triviality of the solution of the problem of automatic processing of all non-factors, the decision to develop methods of their detection and memorization is an acceptable one at the initial stage. In the future, when accessing this knowledge, the personal ostis-assistant can try to eliminate some of those non-factors through dialog with the user by providing the knowledge from the Ecosystem and its personal knowledge base as a source of reliable knowledge.

It is also worth paying attention that the personal ostis-assistant is responsible for the safety of the user himself. In this case, the main task lying on it is to preserve the confidentiality of the personal data of the user. On the one hand, this problem is solved by the user determining the data that he does not want to provide to third parties, and the assistant, accordingly, must save these preferences and follow them. On the other hand, at a certain stage of the development of the Ecosystem, it may be possible to automate the interaction of a personal ostis-assistant with third party systems that may require user data. In this case it is necessary at the level of communities requiring certain personal data to provide their personal data processing policy for open access. So these policies might be automatically processed by personal ostis-assistants. Depending on the preferences of the user through a special configuration of the assistant, you can set the rules that he can automatically accept or reject, and in extreme cases show the relevant parts of the policy to the user and explicitly request his decision.

D. Agents' code security

Since the main goal of existence of every system is solving certain tasks, we can say that the most large-scale source of threats are executable programs and their source code. In the multi-agent systems these are agents. The transmission of agents between ostis-systems for their storage and execution is one of the provided methods for the collective solution of problems. In turn, this requires providing ways to confirm the security of the received agent. The task of providing developers with tools that perform an automated verification of the code for errors and check the security of the developed agent is also no less important.

When receiving an executable file in traditional systems, accomplishment of security checks is not a trivial task due to very low level of machine instructions and manipulated data, i.e. the security degree of the executable code strictly depends on its interpretation. Currently this problem is solved via usage of the large-scale databases containing already known viruses and their signatures in the executable code. SCP being the native ostis-systems programming language uses relatively high-level code for machine instructions and manipulated data. So this allows to analyze the executable code developed for ostis-platforms with the same or even lower complexity as the source code written for traditional systems. Comparison of a "'Hello World" program written in SCP and machine code is shown on Fig. 8.

Note that due to the possibility of complete scanning of actions performed by the agent in the process of its work, one can completely analyze the tasks that it is designed to solve even without its execution. To evaluate the assurance level for agents, it is proposed to use an algorithm based on checking the number of attempts to perform prohibited actions for the user who initiated the execution of the agent. Thus, by using the mechanisms of an abstract interpretation [20], it is possible to implement not only a system that determines an agents' safety at the level of attempts to execute prohibited actions, but also analyzes possible bugs in the code during the development. Accordingly, each agent in the system will be assigned the appropriate assurance level. For example, if some agent gets in to the system from the outside, then the lowest assurance level must be set for it by default. Obviously, in the process of solving of specific problems some intermediate data might be produced and saved in the knowledge base so it is crucial to set up the appropriate access rights for the results of agents' work as well. The mentioned abstract interpretation algorithm in joint with more complex one as given in the [21].

With the growth of the security measures the attacks' complexity will grow as well. One can suppose that they will have more delayed nature and try to lurk behind actions that are harmless on their own. In order to detect those, the mechanisms based on analysis of events and states of the system might be used as given in the [22] in combination with semantic logging of events in the system [23].

V. Conclusion

The article examined the main threats and vulnerabilities relevant for individual ostis-systems and OSTIS Ecosystem in general. The differentiation of access to the knowledge bases of ostis-systems, the implementation of mechanisms of configuration of a personal ostis-assistant and the safety of the agents' source code were defined as the main directions for ensuring security.

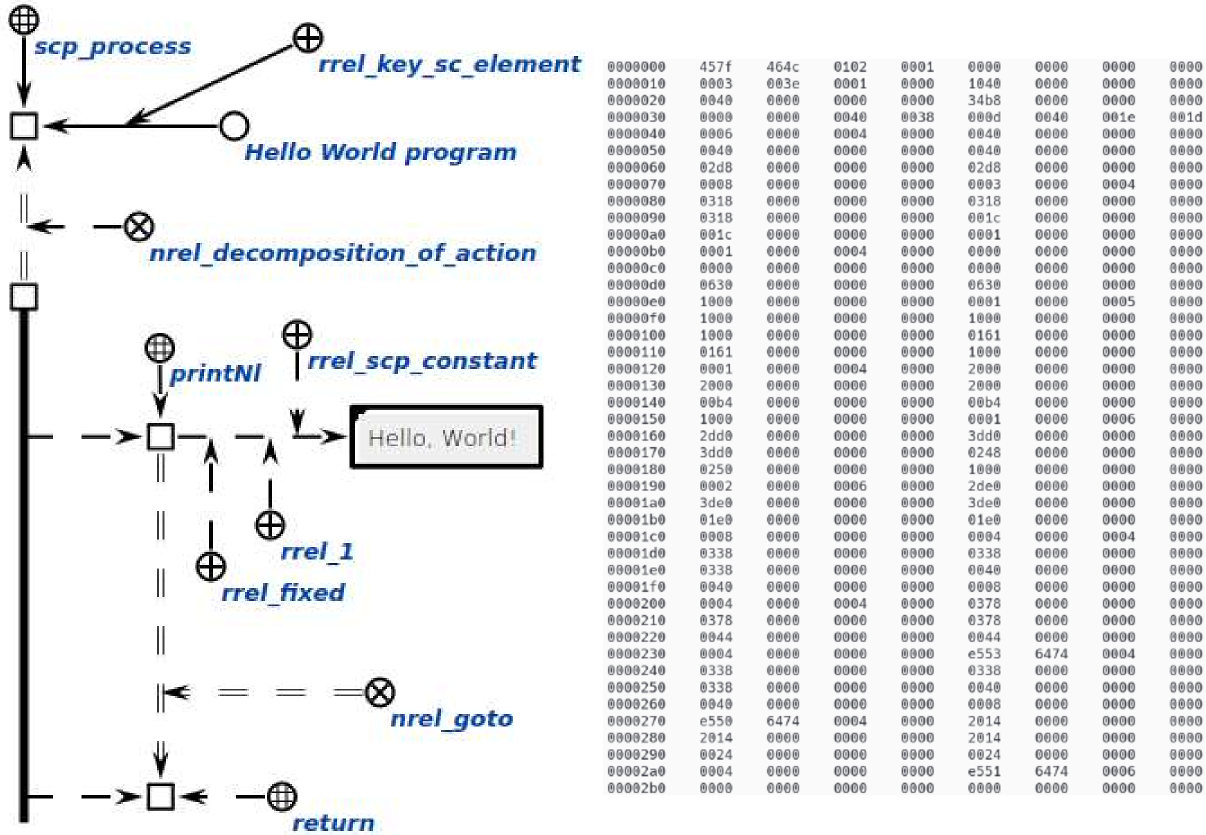


Figure 8: Example of a "Hello World" program written in SCP (left) and machine code (right).

An option was proposed to implement a mechanism for differentiation of access to the knowledge bases of ostis-systems based on the ABAC model. The work examined an example of the architecture of the OSTIS Ecosystem based on the Matrix protocol, as well as ideas for the implementation of safety measures of a personal ostis-assistant and for the agents' source code.

References

[1] S. Isoboev, D. Vezarko, and A. Chechelnitsky, "Intellektual'naya sistema monitoringa bezopasnosti seti besprovodnoi svyazi na osnove mashinnogo obucheniya [wireless communication network security intelligent monitoring system based on machine learning]," *Ekonomika i kachestvo sistem svyazi [Economics and quality of communication systems]*, vol. 1, no. 23, pp. 44–48, 2022.

[2] V. Chastikova and A. Mitiugov, "Metodika postroeniya sistemy analiza intsidentov informatsionnoi bezopasnosti na osnove neiroimmunnogo podkhoda [methodology for building a system for analyzing information security incidents based on the neuroimmune approach]," *Elektronnyi Setevoi Politematicheskii Zhurnal «Nauchnye Trudy Kubgtu» [Electronic Network Polythematic Journal «Scientific Works of the KubSTU»]*, no. 1, pp. 98–105, 2022.

[3] A. Skrypnikov *et al.*, "Reshenie zadach informatsionnoi bezopasnosti s ispol'zovaniem iskusstvennogo intellekta [solving information security problems using artificial intelligence]," *Sovremennye naukoemkie tekhnologii [Modern high technologies]*, no. 6, pp. 277–281, 2021.

[4] E. Sozinova, "Primenenie ekspertnykh sistem dlya analiza i otsenki informatsionnoi bezopasnosti [application of expert systems for analyzing and assessing the information security]," *Molodoi uchenyi [Young scientist]*, vol. 10, no. 33, pp. 64–66, 2011.

[5] *Information security, cybersecurity and privacy protection — Evaluation criteria for IT security — Part 1: Introduction and general model*, ISO/IEC Std. 15 408-1, 2022.

[6] V. Golenkov, Ed., *Tehnologija kompleksnoj podderzhki zhiznennogo cikla semanticheski sovmestimyh intellektual'nyh komp'juternyh sistem novogo pokolenija [Technology of complex life cycle support of semantically compatible intelligent computer systems of new generation]*. Bestprint, 2023.

[7] D. Abdurakhman, "Iskusstvennyi intellekt i mashinnoe obuchenie v kiberbezopasnosti [artificial intelligence and machine learning in cybersecurity]," *Sovremennye problemy lingvistiki i metodiki prepodavaniya russkogo yazyka v vuze i shkole [Modern problems of linguistics and methodology of teaching Russian language at university and school]*, no. 34, pp. 916–921, 2022.

[8] Cwe view: Research concepts. Available at: https://cwe.mitre.org/data/definitions/1000.html (accessed 2024, Feb).

[9] V. Ivashenko, *Modeli resheniya zadach v intellektual'nykh sistemakh. V 2 ch. Ch. 1 : Formal'nye modeli obrabotki informatsii i parallel'nye modeli resheniya zadach : ucheb.-metod. posobie [Models for solving problems in intelligent systems. In 2 parts, Part 1: Formal models of information processing and parallel models for solving problems: a tutorial]*. BGUIR, 2020.

[10] V. Druzhinin and D. Ushakov, Eds., *Kognitivnaya psikhologiya. Uchebnik dlya vuzov [Cognitive psychology. Textbook for universities]*. PER SE, 2002.

[11] R. Dushkin, *Metody polucheniya, predstavleniya i obrabotki znanii s NE-faktorami [Methods of gaining, representation and processing of knowledge with NON-factors]*, 2011.

[12] A. Gribkov, “Formirovanie dostovernogo znaniya: nakhozhdenie znanii i vyyavlenie defektov [the formation of reliable knowledge: Finding knowledge and identifying defects],” *Vestnik Leningradskogo gosudarstvennogo universiteta imeni A. S. Pushkina [A.S. Pushkin Leningrad State University Journal]*, pp. 74–90, 2023.

[13] A. Dementev, “Metriki semanticheskikh dannykh [metrics for semantic data],” *Molodoi uchenyi [Young scientist]*, vol. 24, no. 419, pp. 48–51, 2022.

[14] A. Narinyani, “Ne-faktory: Netochnost’ i nedoopredelennost’ - razlichie i vzaimosvyaz’ (doformal’noe issledovanie) [non-factors: Inaccuracy and under-determination - distinction and relationship (pre-formal study)],” *Mezhdunarodnyi Seminar DIALOG’99 Tarusa [International Seminar DIALOG’99 Tarusa]*, 1999.

[15] I. Davydenko, “Semantic models, method and tools of knowledge bases coordinated development based on reusable components,” *Otkrytye semanticheskie tehnologii proektirovanija intellektual’nyh sistem [Open semantic technologies for intelligent systems]*, pp. 99–118, 2018.

[16] A. Ostroukh, *Intellektual’nye sistemy: monografiya [Intelligent systems: monograph]*. Nauchno-innovatsionnyi tsentr, 2020.

[17] A. Baranovich, “Semanticheskie aspekty informatsionnoi bezopasnosti: kontsentratsiya znanii [semantic aspects of information security: concentration of knowledge],” *Istoriya i arkhivy [History and Archives]*, vol. 13, no. 75, pp. 38–58, 2011.

[18] Matrix specification. Available at: https://spec.matrix.org/v1.9/ (accessed 2024, Mar).

[19] Openpgp. Available at: https://www.openpgp.org/ (accessed 2024, Mar).

[20] P. Cousot and R. Cousot, “Abstract interpretation: Past, present and future,” *CSL-LICS ’14: Proceedings of the Joint Meeting of the Twenty-Third EACSL Annual Conference on Computer Science Logic (CSL) and the Twenty-Ninth Annual ACM/IEEE Symposium on Logic in Computer Science (LICS)*, pp. 1–10, 2014.

[21] V. Khoang and A. Tuzovskii, “Resheniya osnovnykh zadach v razrabotke programmy podderzhki bezopasnosti raboty s semanticheskimi bazami dannykh [solutions to the main problems in the development of a program to support the security of work with semantic databases],” *Doklady Tomskogo gosudarstvennogo universiteta sistem upravleniya i radioelektroniki [Proceedings of Tomsk State University of Control Systems and Radioelectronics]*, vol. 2, no. 28, pp. 121–125, 2013.

[22] A. Lajevardi and M. Amini, “Big knowledge-based semantic correlation for detecting slow and low-level advanced persistent threats,” *Journal of Big Data*, vol. 8, no. 148, 2021.

[23] V. Ivashenko, N. Zotov, and M. Orlov, “Semantic logging of repeating events in a forward branching time model,” *Pattern Recognition and Information Processing (PRIP’2021): Proceedings of the 15th International Conference*, p. 149–152, 2021.

ПРОБЛЕМЫ БЕЗОПАСНОСТИ ЭКОСИСТЕМЫ OSTIS

Хорошавин В. Д., Захаров В. В.

В данной работе рассматриваются угрозы и уязвимости, актуальные для ostis-систем. Разграничение доступа к базам знаний остис-систем, реализация механизмов настройки персонального остис-ассистента и безопасность исходного кода агентов определены как основные направлениями обеспечения безопасности остис-систем. Предложены варианты реализации соответствующих механизмов безопасности по этим направлениям.

Received 13.03.2024

Intelligent Tutoring System for Discrete Mathematics

Kanstantsin Shurmel and Artur Sharapov
Eugene Samokval and Vitaly Tsishchanka
Belarusian State University of Informatics and Radioelectronics
Minsk, Belarus
Email: shurmel.konstantin@mail.ru, artursha ser@gmail.com
samokvalz@gmail.com, vitalik.tsishchanka@gmail.com

***Abstract*—The article presents a model of intelligent tutoring system for discrete mathematics. The model of such system uses methods and tools designed to build intelligent tutoring systems for any discipline and easy integration of new disciplines into the existing tutoring system.**

***Keywords*—knowledge, knowledge base, intelligent systems, problem solver, interface, discrete mathematics.**

I. Introduction

Discrete mathematics is fundamental these days and finds wide application in various fields. These fields include logistics, geographic information systems, computer science, modeling of physical and mathematical phenomena. This variety of applications makes *discrete mathematics* attractive both to commercial organizations looking for process optimization and solving complex problems, and to non-profit organizations engaged in research and development of new methods and algorithms.

Moreover, there are many other fields in which *discrete mathematics* has potential for application, such as sociology, biology, chemistry, and economics. This emphasizes its importance and relevance in modern society. Hence, understanding *discrete mathematics* plays an important role in the progress of science and technology. Therefore, in order for people to learn this science in a convenient way, it is necessary to develop new teaching methods to make the educational material more effective and accessible.

Modern methods of tuition involve not only internal presence of the learner in a particular discipline, but also the possibility of distance learning. As a rule, many of those who already have higher professional education, wish to deepen their knowledge in the discipline of interest, to expand competence in a related professional field of activity and to obtain new skills and knowledge, giving the opportunity to occupy a more successful position in the professional environment.

The first mention of the concept of *intelligent tutoring systems* was defined in 1970 by J. Carbonell. More than 10 years later, real working *intelligent tutoring systems* appeared. The difference between *intelligent tutoring systems* and automated systems is that *automated system* is a *consolidated knowledge base*, based on the results of work with which the system gives the learner the results of correctly and incorrectly answered questions. In turn, *intelligent tutoring system* is aimed at the process of diagnosing learning, its correction. The essence of the work of such a system is not just in diagnosing the learner's mistakes, but also in issuing advice based on predetermined strategies of distance learning [1].

Intelligent tutoring system
:= [A set of software and hardware that uses *artificial intelligence* techniques to create interactive and adaptive educational tools. Such systems are usually able to adapt to the individual needs and knowledge level of each learner, offering personalized assignments, materials selection and feedback.]

Automated learning system
:= [A program or set of programs that facilitate or fully automate the learning process. They may include various functions such as organizing learning material, creating tests and assignments, and tracking student progress. Such systems are usually designed to optimize the learning process, reduce the time spent on routine teacher tasks, and improve learning efficiency.]

The main advantages of an intelligent tutoring system:

- personalized approach to learning, taking into account the individual needs and knowledge level of each student;
- the possibility of interactive classes and the use of visualization to explain theoretical concepts more clearly;
- automatic identification of students' weaknesses and suggestion of additional materials to reinforce the material;
- providing access to a wide range of educational

resources, including study materials, assignments, and tests;
- the ability to receive feedback from the system and teachers to assess progress and correct the learning process.

The main disadvantages of an intelligent tutoring system:

- limited ability of the system to adapt to the specific needs of students with different backgrounds and learning abilities;
- lack of accuracy in identifying student errors and failures, which may lead to underestimation of their real abilities;
- necessity of constant updating of educational content and algorithms of the system in accordance with changes in the discipline program and teaching methods;
- limited opportunities to interact with students outside the learning environment, which can make it difficult to solve individual questions and problems.

One of the best known systems in problem solving is **WolframAlpha** [2]. This system can solve problems from various fields, including graph theory problems.

This system supports functions such as graph editing, basic graph operations, providing examples with known graphs, and searching for elements in a subgraph.

However, the disadvantages of WolframAlpha are that it cannot provide step-by-step solutions to problems, it has a special query language that not every user can understand, and it does not provide training tasks for users.

ALEKS is an intelligent tutoring system that provides a customizable program for teaching students in different subject areas, including discrete mathematics [3]. The system uses an approach based on passive learning theory, which allows hypothesizing about a student's knowledge and determining the most effective learning path based on the student's learning data.

The ALEKS system uses mathematical algorithms that are the basis of its knowledge base and artificial intelligence technology. The ALEKS knowledge base is a system of concepts and tasks that are interconnected and divided into levels of complexity. Each assignment has its own unique difficulty level, which is determined based on the student's study data. This allows students to learn at their own pace and make the most efficient use of their time.

The disadvantages of this approach are its limited depth of understanding of topics and its inability to integrate with other systems. This can lead to limitations in using the system within broader programs of study and limitations in accessing deeper knowledge of discrete mathematics. In addition, a customizable curriculum may be less effective with students who already have a certain level of knowledge in discrete mathematics and do not require such customization.

Webwork is a system of online teaching materials [4]. This system allows teachers to create individual assignments and exercises that students can solve in real time, and also provides the ability to automatically create different versions of assignments and assessments, and use interactive elements such as graphs and visualizations to facilitate the understanding of the material.

The Webwork information base consists of mathematical formulas, algorithms for solving problems, and a database containing assignments and students' answers. The Webwork approach allows new assignments and answers to be quickly added to the database, which is its advantage. However, the disadvantage is the limited functionality, since Webwork cannot use complex mathematical objects such as graphs or matrices. Also, due to the lack of systematization and semantic structuring of knowledge, it is impossible to automate the process of creating tasks and generating answers, and in the absence of a certain problem in the database its solution will not be found.

Maple T. A. is a program that allows teachers to create and automatically check tests and assignments in discrete mathematics [5]. Maple T.A. uses a combination of several approaches. First, this system uses traditional mathematical methods to develop an information base, including the creation of test assignments and courses, and a semantic approach to create knowledge models and intelligent data processing systems. Second, it utilizes machine learning algorithms for adaptive learning and assessment of student knowledge, and incorporates automatic answer checking, manual evaluation, and proof checking. It is possible to add new mathematical objects such as graphs or matrices, which is an advantage of the system.

The system has limited functionality and no integration with other systems, which may limit its use in some areas. For example, in a business domain where it is important to have access to different systems and tools to solve problems, this limitation can be a significant disadvantage. In addition, the limited functionality may not meet the needs of users, which can lead to loss of interest and unproductive use of the system.

An intelligent tutoring system for discrete mathematics should not have the disadvantages inherent in the above systems, namely it should provide the ability to view step-by-step solutions, ease of integration with other systems, and ease of integrating new skills, knowledge, and themes into the existing system.

Also, the above systems are only able to solve a limited range of tasks that are predetermined by the developers, and the systems cannot generate tasks themselves for the user training.

II. Proposed approach

The proposed approach implies the developing of a system based on the *OSTIS Technology* and its basic principles to overcome these disadvantages [6].

OSTIS Technology is a complex *intelligent system* design technology based on semantic knowledge representation, which includes:

- library of generic, reusable and semantically compatible components and *intelligent systems* (components of *knowledge bases, intelligent problem solvers, intelligent user interfaces*);
- compatible semantic knowledge representation languages of various kinds, providing semantic compatibility not only for reusable components of *intelligent systems*, but also for entire *intelligent systems*;
- compatible semantic models of problem solving.

As a formal basis for knowledge representation within the framework of *OSTIS technology*, a unified *semantic network* with a set theory interpretation is used. Such a representation model is called *SC-code* (Semantic computer code). The elements of such a *semantic network* are called *sc-nodes* and *sc-connectors* (*sc-arcs, sc-edges*) [7]. A model of an entity described by means of *SC-code* is called a *sc-model*.

Intelligent systems developed with the use of the *OSTIS Technology* are called *ostis-systems*. Each *ostis-system* consists of a platform-independent unified logical-semantic model of this system (*sc-model* of a computer system) and a platform for interpretation of such models. In turn, each *sc-model* of a computer system can be decomposed into *sc-model of a knowledge base*, *sc-model of a knowledge processing machine*, *sc-model of an interface* and *an abstract sc-memory* where *SC-code* constructions are stored [7].

The creation of *sc-models of the knowledge base* is based on the ontological approach, which implies the creation of ontologies as systems of concepts describing a particular subject area [8].

The problem solver is the main part of the *sc-model of the knowledge processing machine*, which is built on the basis of a *multi-agent* approach, where the interaction of *agents*, called *sc-agents*, is carried out exclusively by means of *semantic memory*, which stores *SC-code* constructions [9]. This approach allows to ensure modularity and flexibility of the developed machine, and also provides the possibility of parallel execution of different knowledge processing processes.

III. Knowledge base

An important step in the design of *intelligent tutoring systems for discrete mathematics* is the creation of a *knowledge base* that will contain complete and structured information on *discrete mathematics*. However, in order for this *knowledge base* to be as useful and effective as possible, it is necessary to consider a number of criteria that will allow it to be evaluated.

Based on this, the following basic and most important criteria for analyzing a knowledge base were highlighted:

- *Knowledge base* for *discrete mathematics* can be organized as a tree structure, where each node represents a particular topic, or as a network structure, where each node represents a different theorem or algorithm. Evaluating the structure of the *knowledge base* may include analyzing the hierarchy of topics, the relationships between topics and individual elements of the *knowledge base*, and the ease of navigating the *knowledge base*.
- *knowledge base* for *discrete mathematics* should contain sufficient information about key concepts, theorems, algorithms, and applications. Assessment of the quality and completeness of the information may include verifying that all necessary definitions and theorems are present and that the information provided is accurate and reliable.
- *knowledge base* for *discrete mathematics* should be user-friendly and easily accessible for use by students, teachers and researchers. Evaluation of functionality and usability may include an analysis of the accessibility of the *knowledge base* and the ability to search for information.

An important factor affecting the efficiency and quality of the system is the structured nature of the *knowledge base*. For this purpose, it is necessary to have an organized hierarchy of partial *subject domains*, which are strongly connected with each other by many different relations related to some common (base) *subject domain*.

Figure 1 contains an example of the description of a **subject domain of graph theory**.

Determining the relations between concepts is an important step in the development of a *knowledge base*, and requires careful analysis and careful approach. According to *OSTIS Technology*, to achieve the best quality of the *knowledge base*, concepts and relations will be described using their *semantic neighborhoods*. The basic relations and concepts include identifier, definition, statement, inclusion and decomposition. The implementation of each relation and their representation in the *knowledge base* depends on the particular case.

The next step is to define the rules and principles of describing the elements of the *knowledge base*. If there are several concepts in the *knowledge base* that are related to each other, the correct description of these concepts and connections between them allows to avoid duplication of information and ensure the integrity of the *knowledge base*.

Figure 2 shows an example of the description of the concept **disconnected graph**.

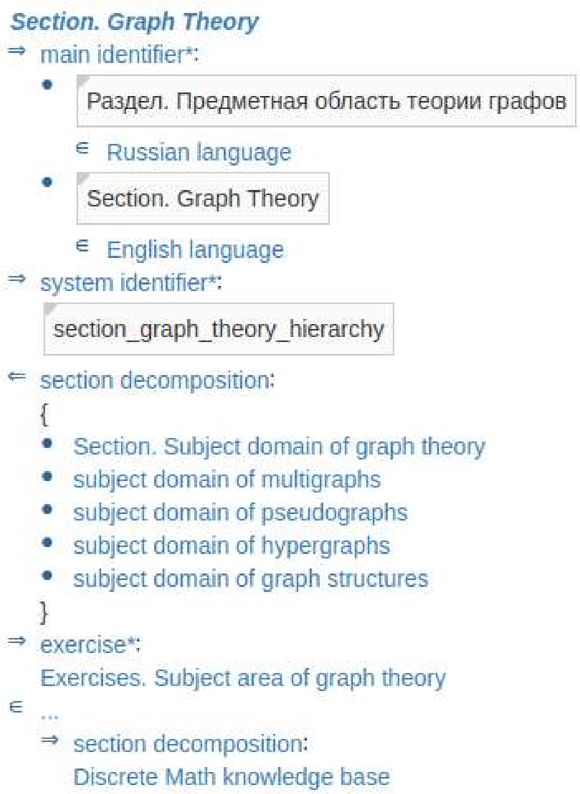


Figure 1. Section. Subject domain of graph theory

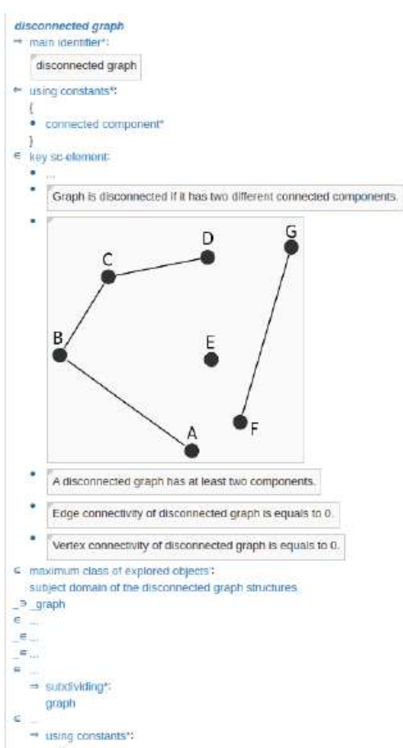


Figure 2. Concept disconnected graph

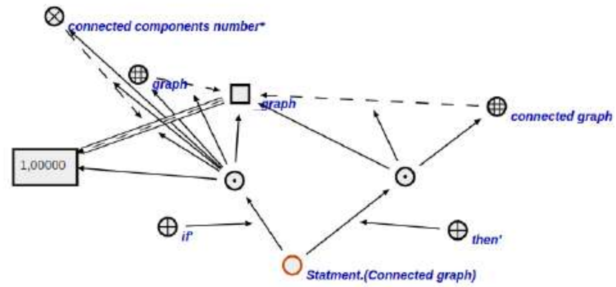


Figure 3. Statement about connected graph

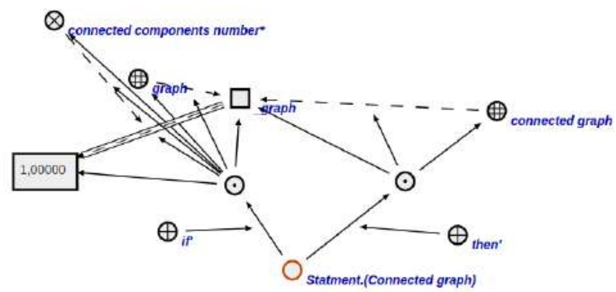


Figure 4. Specification of program for finding the union of two graphs

A unified description of the elements of the *knowledge base* creates an opportunity for its expansion in the future. The predefined rules and principles for describing elements will allow us to quickly and easily add new elements to the *knowledge base* without having to redesign its entire structure. Thus, we can be sure that system is flexible and scalable [10].

An example of a statement about a *connected graph* is shown in Figure 3, it is a logical formula written as an implication. The premise of the implication contains a pattern denoting that the graph has only one connectivity component. In the conclusion of the implication, the graph belongs to the set of connected graphs.

The *knowledge base* also contains specifications of programs that can be used by the *problem solver*. Thus, Figure 4 shows the specification of the *graph union finding program*. The result of this program is a graph that is the union of two input graphs.

An example of an exercise is shown on Figure 5.

IV. Problem solver

An important step in creating a *intelligent tutoring system* for *discrete mathematics* is to develop a *problem solver* that will solve problems in *discrete mathematics* with a full description of the step-by-step solution. But in order for the *problem solver* to fully fulfill its function, it is necessary that the *problem solver* meets the following criteria:

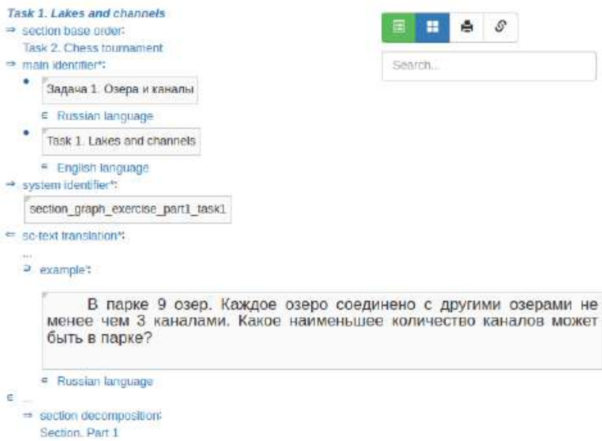


Figure 5. Exercise example

1) Correctness of solved problems is a key aspect when evaluating a *problem solver* in *discrete mathematics*. This means that the *problem solver* must be able to correctly solve any problem in its subject domain. This includes not only the ability to solve problems, but also the ability to adapt to different types of problems and conditions. If there are several ways to solve a problem, the solver should be able to choose between them and apply the most appropriate method depending on the specific task conditions. This may involve analyzing the complexity of different methods, evaluating their effectiveness, and determining the most optimal approach.
2) *problem solver* in an *intelligent tutoring system* should have the ability to describe in detail and step-by-step the process of solving the current problem. This includes a full or partial description of all algorithms used, which is critical for students to understand the logic and methodology of problem solving.
   Describing the problem solving process in detail helps students understand how to apply theoretical knowledge in practice. It also helps them develop critical thinking and analytical skills, as they can follow the problem solving process and understand how each step affects the final result.
   Describing the algorithms used is also an important part of the learning process. It helps students understand how different algorithms work and how they can be applied to solve specific problems. It can also help them develop programming and algorithmic thinking skills.
3) *problem solver* in an *intelligent tutoring system* should be user-friendly and feature-rich to ensure effective and productive learning. Usability may reduce the attractiveness of the *problem solver* as a learning tool, as it may increase the time and effort required to complete tasks, and thus may discourage users. Multifunctionality is also an important aspect of a *problem solver*. This means that a *problem solver* should have a wide range of features that can help users solve different types of tasks. Limited functionality may make the *problem solver* less useful to users, as it may not be able to solve all types of problems that users encounter.

Intelligent tutoring system for discrete mathematics problem solver consists of the the following modules:

- module for solving problems;
- module for generating and evaluating problem complexity;
- module for checking the correctness of the solution.

In the context of developing a *problem solver for an intelligent tutoring system for discrete mathematics*, *multiagent approach* can be used to implement different modules of the system, such as *problem solving module*, *solution correctness checking module*, *problems generation and complexity evaluation module* and others. Each module can be represented as an *agent* that performs its functions and interacts with other *agents* to achieve the goal of the system.

Designing a problem solving module. This module is part of the *intelligent tutoring systems for discrete mathematics* and provides solutions to problems based on a class of problems.

The functioning of this module consists of the interaction of *agents* from the following *sc-agent* hierarchy:

Abstract non-atomic sc-agent of problem solving
⇒ *decomposition of abstract sc-agent**:
  {• *Abstract sc-agent of task specification generation by template*
  • *Abstract sc-agent of solving a complex problem*
  • *Abstract non-atomic sc-agent of solving a simple problem*
    ⇒ *decomposition of abstract sc-agent**:
      {• *Abstract sc-agent of finding the relation of a given sc-element with a given concept*
      • *Abstract sc-agent of unisg an unary operation*
      • *Abstract sc-agent of using a binary operation*
      }
  }

Designing a module for generating and evaluating problem complexity. This module is an important part of the *intelligent tutoring systems for discrete mathematics* and provides problem generation and complexity evaluation. It is a tool that allows the generation of a variety of problems, taking into account different parameters and requirements, and at the same time estimating their complexity in order to adapt problems to the learners' level of knowledge.

Problem generation is an important function of this module, providing the ability to create tasks using specified parameters such as problem type, number of variables, constraints and other factors. This module generates unique tasks each time, promoting variety and fun for learners.

Problem difficulty evaluation is another important feature of the module, based on analyzing the generated problems and determining their difficulty based on specified criteria. The evaluation criteria may include the number of steps to solve, the use of complex algorithms or mathematical concepts. Conducting such an evaluation allows system to objectively assess the complexity of tasks and compare them.

Designing a module for checking the correctness of the solution. This module is part of the *intelligent tutoring systems for discrete mathematics* and is designed to check the correctness of the solution of problems. It analyzes the solution provided by the user and checks if it corresponds to possible solutions of the problem.

The main functions of the module are:

- module analyzes the structure of the solution, checks the presence of the necessary blocks of the solution, the correctness of their location and links;
- module checks the logic of the solution, analyzes the correctness of algorithms and logical operations used in problem solving;
- module checks the correct syntax of the program code in the solution;
- module checks the answer by comparing the obtained result with the expected one and determines whether the solution is correct or not.

Advantages of the solution correctness checking module:

- module allows system to automatically analyze the solution, which significantly speeds up the verification process and reduces the probability of errors;
- module is based on the specified conditions and requirements, which allows system to make an objective assessment of the correctness of the solution;
- module conducts a detailed check of all aspects of the solution, including structure, logic and syntax, which allows system to identify and point out errors.

The use of the solution checking module:

- the user provides their solution in the form of program code or algorithm;
- the solution validation module analyzes the provided solution with using specified algorithms and rules;
- the module displays the result of the check, indicating the detected errors or confirming the correctness of the solution.

V. User interface

The interface for *intelligent tutoring systems for discrete mathematics* is of particular interest because graphical representation of the main objects of study of graph theory and set theory, the two main components of *discrete mathematics*, graphs and sets, is the most convenient and effective for human understanding.

There are many software solutions related to the visualization of graph and set structures, but most of them are focused on solving specific highly specialized problems. In this connection, when it is necessary to solve a new problem, or conceptually the same problem, but from another subject area, the development of a new software solution for the task becomes the only way out, including the construction of its own visualization.

Proceeding from the fact that the most effective learning takes place in practice, when solving specific and possibly real-life problems, and through the acquisition of relevant experience, the tutoring system should be able to provide all the necessary tools and elements of the graphical interface for the appropriate practice of learners.

In this regard, a virtual space for working with graphs was developed - an element of the graphical interface, which contains a graph visualizer and editor, as well as elements of control and manipulation of graphs. This interface element allows creating, editing, and loading graphs stored in the *sc-memory* of the *sc-machine*, a software implementation of the semantic network storage and processing. In addition, the controls, which are graphical interpretations of the corresponding elements of the *knowledge base* stored in the *sc-memory*, allow to perform a certain set of actions on the graphs contained in the workspace. It is important to note that this set is defined exclusively by the description of the corresponding actions in the *knowledge base*, thus ensuring automatic and dynamic integration of new actions.

The graph editor used in the interface was designed to support the SCg alphabet and behave similarly to the SCg-editor during user interaction, but many improvements were introduced as well. The following were implemented: moving with touchbar, zooming with "pinch" on touchbar, etc.

In addition, the developed graph editor was designed in accordance with a modular architecture, where each element is an independent component responsible exclusively for its functions and extending the capabilities of the basic graph structure visualization component [11]. Such a solution is particularly time-consuming during design and initial development, but it allows to extend the capabilities of the editor in the future using the developed internal programming interfaces for connecting new components or plug-ins, without any modification of the source code of the editor or the visualizer. As a result, the implemented graph editor offers developers interested in using it not only the simplest and most efficient way to extend the editor functionality, but also an ability to customize the editor according to specific requirements,

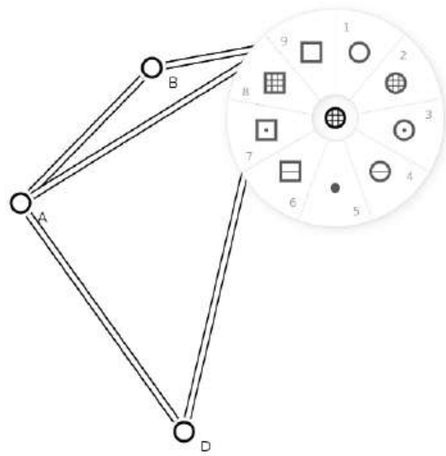


Figure 6. Graph editor example

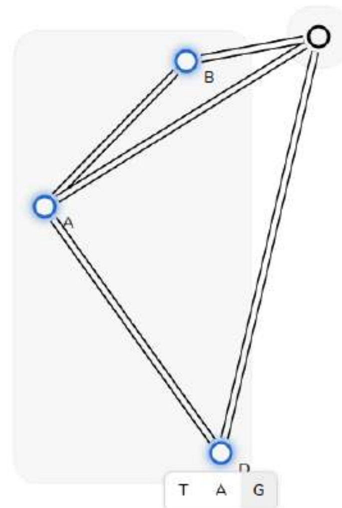


Figure 7. Floating menu example

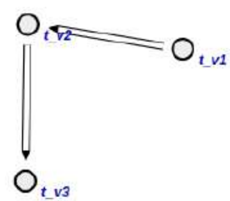


Figure 8. First graph

excluding certain plug-ins built into the editor by default or changing their configuration.

An example of a graph editor is shown in the Figure 6.

One of the improvements of the graph editor is also a floating menu, supplied as a plugin (independent component) of the editor as an alternative to the classic menu. This decision can be justified by the following laws.

Fits Law allows to quantify the fact that the farther an object is from the current cursor position or the smaller the size of this object, the more time the user will need to move the cursor to it.

Hick's Law quantifies the observation that the more options of a given type you provide, the longer it takes to choose [12].

The floating menu appears automatically near the user's cursor when selecting certain objects, thereby minimizing the distance required for the cursor to overcome to perform a particular action on the selected objects. In addition, the menu is automatically hidden after receiving a signal indicating cursor movement away from it, provided that the user has pointed to this menu at least once, thus informing the user of its existence on the one hand, but not interfering with the user in his work on the other hand, conditionally guaranteeing that the user has paid attention to the existing menu under the selected objects.

An example of the floating menu is shown in the Figure 7.

In addition, depending on the type of selected objects in the floating menu, the system will offer only those actions that can be performed only on objects of the selected type. Thus, this solution allows system to significantly reduce the number of options for selecting an action and, consequently, the time required for the user to make a choice, which significantly improves the usability of the interface.

VI. Demonstration of results

As an example, the problem of determining whether the union of two graphs, shown in Figure 8 and Figure 9, is a tree is given.

Figure 10 shows a task template for determining whether the union of two graphs is a tree. This template specifies the input arguments of the problem and the goal of the solution.

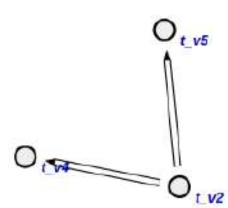


Figure 9. Second graph

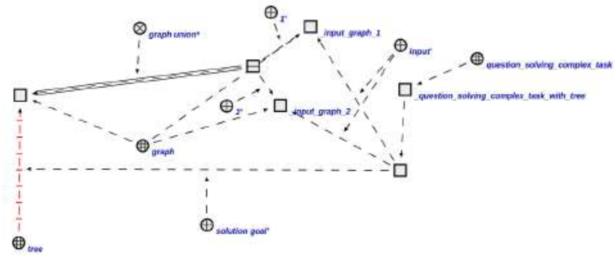


Figure 10. Task template

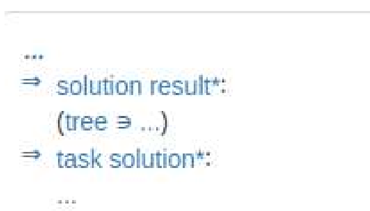


Figure 11. Result of problem

The result is shown in the figure 11. First the *sc-agent of task specification generation by template* created a specification of a task and then it called *sc-agent of solving a complex problem* which solved the problem because the *knowledge base* had the statement that *the tree is a connected acyclic graph*, the *problem solver* knows how to determine whether the graph is connected or not, whether the graph is acyclic or has cycle and the *problem solver* knows how to find the union of two graphs.

VII. Conclusion

Discrete mathematics is applied in various fields including logistics, geographical information systems, computer science, modeling of physical and mathematical phenomena, as well as sociology, biology, chemistry and economics, among others. Therefore, the development of an *intelligent system* to solve problems directly or indirectly related to discrete mathematics is of great relevance and importance in modern society.

Based on this work, the main components of *intelligent tutoring systems for discrete mathematics* such as *knowledge base*, *problem solver* and *user interface* have been identified and described. In addition to this, the requirements that all the components of the *intelligent tutoring systems* should follow and their functions were also identified.

Based on all the above, a prototype of *intelligent tutoring systems for discrete mathematics* has been developed. However, this is only the beginning, and options for further development include the implementation of user tutoring and a personalized learning approach.

Acknowledgment

The authors would like to thank the research groups of the Department of Intelligent Information Technologies of the Belarusian State University of Informatics and Radioelectronics

References

[1] A. Liliya, “Social technologies and processes,” *International scientific journal "BULLETIN OF SCIENCE" No. 1*, pp. 149–155, 2018.

[2] (2024, March) Wolfram|Alpha. [Online]. Available: https://www.wolframalpha.com/

[3] (2024, March) ALEKS - Assessment and LEarning in Knowledge Spaces. [Online]. Available: https://www.aleks.com/

[4] (2024, March) The WeBWorK Project. [Online]. Available: https://openwebwork.org/

[5] (2024, March) Maple T.A. [Online]. Available: https://www.maplesoft.com/

[6] V. Golenkov, N. Gulyakina, I. Davydenko, and D. Shunkevich, “Semanticheskie tekhnologii proektirovaniya intellektual’nyh sistem i semanticheskie associativnye komp’yutery [Semantic technologies of intelligent systems design and semantic associative computers],” *Otkrytye semanticheskie tekhnologii proektirovaniya intellektual’nykh system [Open semantic technologies for intelligent systems]*, pp. 42–50, 2019, (In Russ.).

[7] V. V. Golenkov and N. A. Gulyakina, “Graphodynamic models of parallel knowledge processing: principles of construction, implementation and design,” *Otkrytye semanticheskie tekhnologii proektirovaniya intellektual’nykh system [Open semantic technologies for intelligent systems]*, pp. 23–52, 2012.

[8] K. Bantsevich, “Structure of knowledge bases of next-generation intelligent computer systems: a hierarchical system of subject domains and their corresponding ontologies,” *Otkrytye semanticheskie tekhnologii proektirovaniya intellektual’nykh system [Open semantic technologies for intelligent systems]*, pp. 87–98, 2022.

[9] D. Shunkevich, “Agentno-orientirovannye reshateli zadach intellektual’nyh sistem [Agent-oriented models, method and tools of compatible problem solvers development for intelligent systems],” *Otkrytye semanticheskie tekhnologii proektirovaniya intellektual’nykh system [Open semantic technologies for intelligent systems]*, pp. 119–132, 2018.

[10] A. S. Sharapov, “Designing knowledge bases of intelligent learning systems based on ostis technology,” *A young scientist*, pp. 17–19, 2023. [Online]. Available: https://moluch.ru/archive/478/105252/

[11] W. G. Griswold, M. Shonle, K. Sullivan, Y. Song, N. Tewari, Y. Cai, and H. Rajan, “Modular software design with crosscutting interfaces,” *IEEE software*, vol. 23, no. 1, pp. 51–60, 2006.

[12] V. V. Golenkov, N. A. Gulyakina, D. G. Kolb, “Intelligent user interface,” pp. 35–36, 2013. [Online]. Available: https://libeldoc.bsuir.by/bitstream/123456789/998/2/Golenkov_Int.pdf

ИНТЕЛЛЕКТУАЛЬНАЯ ОБУЧАЮЩАЯ СИСТЕМА ПО ДИСКРЕТНОЙ МАТЕМАТИКЕ

Шурмель К. А., Шарапов А. С.,
Самохвал Е. С., Тищенко В. Н.

В статье представлена модель интеллектуальной обучающей системы по дискретной математике. Модель такой системы использует методы и средства, рассчитанные на построение интеллектуальных обучающих систем по любой дисциплине и простую интеграцию новых дисциплин в существующую обучающую систему.

Received 01.04.2024

Methods and Means of Constructing Plans for Solving Problems in Intelligent Systems on the Example of an Intelligent System on Geometry

Natallia Malinovskaya and Anna Makarenko
Belarusian State University of
Informatics and Radioelectronics
Minsk, Belarus
Email: natasha.malinovskaya.9843@gmail.com, anna.makarenko1517@gmail.com

***Abstract*—In this paper we propose an approach to the development of methods and tools for constructing plans for solving problems in intelligent systems on the example of an intelligent system for geometry. The described approach is aimed at improving the accuracy of answers due to the possibility of decomposition of problems into simpler ones, and also aims to overcome the shortcomings of modern intelligent systems. An intelligent system realizing the proposed approach is described.**

***Keywords*—problem solving, knowledge, knowledge base, intelligent systems.**

I. Introduction

Nowadays, the use of intelligent systems in various fields is becoming relevant. However, the quality of ISs is largely determined by its answers and the ability to solve complex problems, which is why it is necessary that its answers are as accurate and reliable as possible. In order to increase the accuracy of answers it is necessary to be able to decompose problems into simpler ones, in turn, the automation of this process will allow the IS to solve not only simple problems, but also complex, non-trivial ones. The relevance of systems that have the ability not only to perform the above functions, but also have the ability to partially satisfy the human need for quick answers, allowing you to automate some of the routine actions.

Problem solver is one of the key components of ISs, allowing them to solve a wide range of problems. Unlike other modern software systems, the peculiarity of problem solvers in ISs is the necessity to solve problems in conditions when the necessary information for their solution is not explicitly localized in the knowledge base of the IS and must be found in the process of problem solving on the basis of certain criteria.

The composition of the problem solver in each particular system depends on its target, classes of solved problems, subject area and other factors. In general, a problem solver provides the ability to solve problems related both to the core functionality of the system and to ensure its efficient operation and development automation. A problem solver that performs all these functions is called a unified problem solver for a given IS.

Expanding the areas of application of ISs requires their ability to solve complex problems, which involve the joint use of different models of knowledge representation and problem solving models. In addition, solving complex problems involves the use of shared information resources, such as a knowledge base, by different components of the solver that specialize in solving different subproblems. Since a complex problem solver integrates different problem solving models, it is called a hybrid problem solver.

Examples of complex problems are:

- problems related to understanding natural language texts (both printed and handwritten), understanding speech messages and images. In each of these cases it is required to perform syntactic analysis of the processed file or signal, remove insignificant fragments, classify significant fragments, relate them to concepts known to the system, etc.;
- automation of adaptive learning for schoolchildren and students, which implies that the system is capable of autonomously solving various problems from a certain subject area, as well as managing the learning process, generating tasks for students and controlling their independent fulfillment by the student;
- the problems of planning the behavior of intelligent robots, which involve both understanding a variety of external information and making various decisions using both credible methods and methods that rely on probabilistic estimates and plausible assumptions;
- problems related to complex and flexible automation of various enterprises.
- etc.

The use of different problem solving models within an IS implies decomposition of a complex problem into subproblems that can be solved using one of the known IS problem solving models. Due to the combination of different problem solving models, the set of problems solved by the hybrid solver will be much wider than the combination of sets of problems solved separately by all problem solvers included in its composition [1].

The existing variety of approaches to problem solving in computer systems can be divided into two classes:

- problem solving using stored programs. In this case, it is assumed that the system has a program for solving a problem of a given class in advance, and the solution is reduced to searching for such a program and interpreting it on the given input data. The systems oriented on such approach to problem solving include systems using:
  - programs written in both imperative and declarative programming languages, including logical and functional programming [2];
  - genetic algorithm implementations [3], [4];
  - neural network models of knowledge processing [5], [6], [7].

  It should be noted that even in the case of using a stored program, the solution of the problem is not always trivial, because, first, it is required to find such a stored program on the basis of some specification, and second, to provide its interpretation;
- solving problems in conditions where the solution program is not known.

In this case, it is assumed that the system does not necessarily contain a ready-made solution program for the class of problems to which some formulated problem to be solved belongs. In this connection, it is necessary to apply additional methods of searching for ways to solve the problem, which are not designed for any narrow class of problems (e.g., splitting the problem into subproblems, methods of searching for solutions in depth and width, method of random solution search and trial-and-error method, etc.), as well as various models of logical inference (classical deductive, [8], inductive [9], [10], abductive [8]; models based on fuzzy logics [11], [12], [13], the logic of default [14], temporal logic [15], and many others).

For example, currently one of the most popular approaches to text generation for natural language problems is the use of *large language models* [16], which are models consisting of neural networks with many parameters trained on a large amount of unlabeled text, but these models have a number of disadvantages, partial solution of which can be prevented by integration of such systems with knowledge bases [17], but this approach is not able to solve all the disadvantages of large language models in solving complex problems.

Thus, although there have been significant advances in the development of problem solvers for ISs, there are still outstanding challenges related to provisioning:

- compatibility of different private problem solvers, i.e. the possibility of their coordinated use in solving the same complex problem;
- possibilities to modify the hybrid solver without significant additional costs in the process of operation of the IS. This includes expanding the number of used problem solving models without restrictions on their type. This requirement is due to the fact that when solving a complex problem, it may be unknown what specific problem-solving models and types of knowledge will be needed.

The modern development of *artificial intelligence* is moving towards the creation of intelligent computer systems of a new generation [18]. These systems are capable not only of solving problems from various fields of knowledge, but also of explaining their solutions. However, the disadvantages described above do not allow such systems to be built solely on the basis of existing solutions. Instead, new-generation ISs are built on a unified knowledge base that integrates problems, subject domains, and methods of their solution. Thus, we conclude that the use of modern solutions such as neural networks, models using specialized software interfaces between different system components, large language models can and should become a powerful tool for solving problems of ISs [19], but they cannot completely replace these systems [20].

That is, despite the fact that currently there is a large number of problem-solving models, many of which are implemented and successfully used in practice in various systems, the problem of low consistency of the principles underlying the implementation of such models and the lack of a single unified framework for the implementation and integration of different models remains relevant, which leads to the fact that:

- it is difficult to simultaneously use different models of problem solving within one system when solving the same complex problem;
- it is practically impossible to use technical solutions realized in one system in other systems;
- in fact, there are no complex methods and tools for building problem solvers, which would provide the possibility of designing, realizing and debugging solvers of different kinds.

Therefore, the ability of an IS to independently solve non-trivial problems will expand the possible functionality of the IS, detail and improve the accuracy of its answers.

The purpose of this study is to refine the ontologies of actions and problems, to develop a collective of agents that allows to divide the problem into subproblems, as well as to develop a methodology for the application of the obtained subsystem in specific application systems on the example of an IS for geometry.

II. Proposed approach

In the development of the module of building plans for solving the problems of the IS, it is proposed to use the OSTIS Technology, focused on the development of a class of systems, which are called knowledge-managed computer systems, as well as its basic principles [21], since any *ostis-system* consists of *knowledge base*, *problem solver* and *user interface*, which corresponds to the classical definition of an IS [22].

This technology uses a unified semantic network with a set-theoretic interpretation as a formal basis for knowledge representation. This representation model is called SC-code (Semantic computer code). Elements of such semantic network are called sc-nodes and sc-connectors (sc-arc, sc-edges). Agents in the developed system described by means of SC-code will be called a semantic agent or simply sc-agent.

As it was mentioned earlier, in order to be able to solve some non-trivial problems it is necessary to be able to divide them into trivial ones, i.e. to build a plan for problem solving. When constructing plans for solving the problem of an IDS, it is suggested to divide the main problem into separate subproblems taking into account various separate independent components necessary for solving the main problem.

Let us introduce the concepts *problem*, *subproblem*, *action*, *subaction*.

A. Fragment of the obtained ontology

planning of IDS problems
:= [partitioning a problem into smaller problems - allocation of separate subproblems for solving the main problem]

The concept of **problem** is directly related to the concept of **action**, so let's consider the classification of the concept of **action**.

1) action classification: Let us consider the specification of the concept of *action* in SCn code.

action
:= [*impact*, in which the *subject*$'$ carries out the *action* purposefully, i.e. according to some *target**]
:= [targeted action performed by one or more actors (cybernetic systems) with the possible application of some tools]
⊂ *impact*
    := [a process in which at least one influencing entity (subject of influence$'$) and at least one entity to be influenced (object of influence$'$) can be clearly distinguished]
    ⊂ *process*
:= [a purposeful ("conscious") process performed (controlled, realized) by a certain subject]
:= [work]
:= [problem solving process]
:= [target-oriented process]
:= [holistic piece of some activity]
:= [A purposeful process controlled by some entity]
:= [the process of performing some action by some subject (executor) on some objects]

target*
:= [target situation*]
⊂ *specification*
:= [description of what is to be obtained (what situation is to be achieved) as a result of performing a given (specified) action*]

Each *action* performed by one or another *subject* is interpreted as a process of solving a certain problem, i.e., a process of achieving a given *target** under given conditions, and, therefore, is performed purposefully. However, an explicit indication of the *action* and its relation to a particular problem may not always be present in memory. Some *problems* may be solved by certain subjects permanently, e.g., optimizing a knowledge base, searching for incorrectness, etc., and for such problems it is not always necessary to explicitly introduce the *structure* that is the formulation of the *problem*.

In its turn, the concept of a *subproblem* or *action* is a separate independent *problem* (action), which is performed within the framework of some other extensive *problem* (action).

B. Architecture of a problem solver for partitioning problems into subproblems

To realize an IS, which includes methods and means of constructing plans for solving problems of the IS, it was necessary to build a decomposition of the problem solver of the system, where the key agent of the IS is an abstract non-atomic sc-agent of constructing a plan for solving problems of the IS, taking into account the above classification.

The following is the decomposition of the IS problem solver in the form of SCn-code:

Problem solver for partitioning a problem into subproblems of an intelligent system
⇒ *decomposition of abstract sc-agent**:
  {• *Abstract sc-agent for classifying a message by topic*
      ⇒ *realization**:
          C++ language
  • *Abstract sc-agent for classifying a message by type*
      ⇒ *realization**:
          C++ language
  • *Abstract sc-agent for generating a problem condition from a template*
      ⇒ *realization**:

C++ language

- *Abstract sc-agent for finding the meaning of a problem*
  ⇒ *realization**:
    C++
- *Abstract sc-agent targeting*
  ⇒ *realization**:
    C++ language
- *Abstract non-atomic sc-agent for constructing a problem-solving plan for an intelligent dialog system*
  ⇒ *realization**:
    C++ language
- *Abstract sc-agent for knowledge base replenishment*
  ⇒ *realization**:
    C++ language
- *Abstract sc-agent for generating a response to a message*
  ⇒ *realization**:
    C++ language

}

In this article we consider the approach to solving complex problems, so we will consider in detail the abstract non-atomic sc-agent of building a plan for solving problems of the IS.

As a part of the abstract non-atomic sc-agent of constructing a plan for solving problems of an intelligent dialog system, the following group of agents that break the problem into separate subproblems was identified:

Abstract non-atomic sc-agent for constructing a plan for solving problems of an intelligent system
⇒ *decomposition of abstract sc-agent**:
{
- *Abstract sc-agent for solving a composite problem*
  ⇒ *realization**:
    C++ language
- *Abstract non-atomic sc-agent for solving a simple problem*
  ⇒ *realization**:
    C++ language
- *Abstract sc-agent for interpreting a non-atomic action*
  ⇒ *realization**:
    C++ language

}

C. Principles of application of the developed solver in application systems

The approach proposes the introduction of separate modules with their ontologies and problem solvers into the developed system, which will increase the range of capabilities of the developed system.

Let us consider *abstract non-atomic sc-agent of constructing a plan for solving problems of an IS* on the example of a system with implemented module of an IS on geometry.

Examples will be considered on the basis of dialog-based ISs.

The abstract sc-agent for solving a composite problem takes as input one parameter, which specifies the condition of the composite problem written in the form of sc-code and a node denoting the target of solving the problem. The task of this agent is to divide the original problem into simple subproblems, and also to call for each simple subproblem one of the agents for solving simple problems in a certain sequence, which is set by the abstract sc-agent of the interpretation of non-atomic action, in some cases there is a need to execute the agent of the application of the logical formula [23]. The result of solving the subproblem is recorded in the decision tree, which is ultimately the solution of the user's problem.

That is, the abstract non-atomic sc-agent of constructing a plan for solving the problems of the IS involves breaking the main problem-request into separate subproblems to form a better response to the user's message, i.e., solving smaller problems allows increasing the quality of knowledge immersed in the database. As an input construct, we obtain a situation that is formed due to an abstract sc-agent of target-setting developed within the framework of the IS under development, the input of which is a user's user's message. As an example, let us take the user's request "What is the area of the figure?".

Depending on the knowledge in the KB, the formed target may not be deployed at all, or it may be more deployed, which allows to break the problem into subproblems more qualitatively, i.e., to solve the problem in the most accurate way in the future.

The figure 1 shows an example of a more extended generated situation, namely the target of "find out the area of a figure".

When receiving a question from a user about the area of a figure without additional instructions, it is impossible to answer it directly. As a result of the abstract target-setting agent's work in the IS, a situation arises that requires solving a non-trivial problem. To break this problem into simpler subproblems, an abstract non-atomic sc-agent of constructing a problem-solving plan is used.

We apply a depth-first search method over a graph in which we define the corresponding subproblems that are trivial. Unknown values for these subproblems can be discovered by searching the KB or by using an agent included in the IS module on geometry. If it is still not possible to solve the problem, we ask for clarifications from the user to obtain additional information that will contribute to expanding the description about the situation.

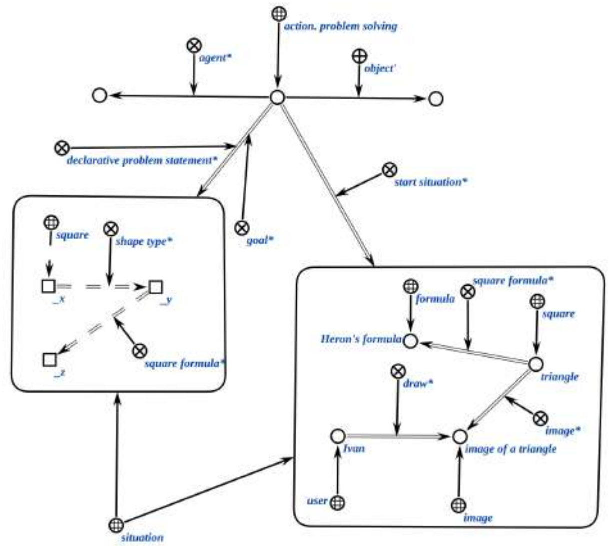


Figure 1. Example of a detailed situation generated as a result of a dialog with a user

Therefore, the agent of building a plan for solving problems of the IS in turn consists of an agent for solving a composite problem and an agent for solving a simple problem. That is, in the context of the developed dialog system, a simple problem will be an indivisible problem that does not require additional information, such a problem, the result of which can be solved as a result of a search in the KB, or as a result of the work of an abstract sc-agent from additional implemented modules.

In general, a *problem solving plan* consists of a sequence of simple and compound problems. The composite problems are solved by *Abstract sc-agent for interpreting non-atomic action*. It receives 1) a processing program in the form of a template and 2) sets of arguments that it retrieves from the semantic neighborhood of the composite problem. The template data is matched with the arguments and the corresponding agents are invoked, to solve the atomic actions belonging to the composite problem. The invoked agents perform the actions of the composite problem.

In essence, this agent is an interpreter that generates a processing program, i.e., a program that includes a sequence of agent executions. A processing program — is a description of a non-atomic action to be performed to solve a composite problem. Sequential execution of agents is produced by sequential processing of transitions between agents.

The transition to the next action depends on the result of the previous one. After the action is completed, i.e. after it is added to the problem_executed class, its success is checked to determine the required transition. There are 3 variants of transition:

1 Transition on successful completion of the action. It is defined by the then* relation. Transition by this relation is performed at successful completion of the action from which the transition is performed (its addition to the class of successfully completed action).

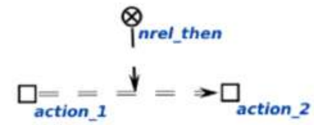


Figure 2. An example of recording a transition between two actions, depending on the result of the previous action

2 Transition at unsuccessful completion of the action. It is set by the then* relation. The transition by this relation is performed at the successful completion of the action from which the transition is performed (its addition to the class of unsuccessfully completed action).

3 Unconditional transition.

The figure 2 shows an example of recording a transition between two actions, depending on the result of the previous action.

In addition to transitions depending on the result of the previous action, conditional transitions are introduced by means of the state* relation. The first element of pairs of this relation are pairs (arcs) of transitions according to the success of action completion, the second element is a logical formula. In this case, an additional condition is imposed on the transition pair (besides the success / failure of action completion) - the truth of the logical formula. In this case, the truth of the formula is calculated for the same substitutions that were used to generate the process by the program.

The figure 3 shows an example of writing a conditional transition between two actions.

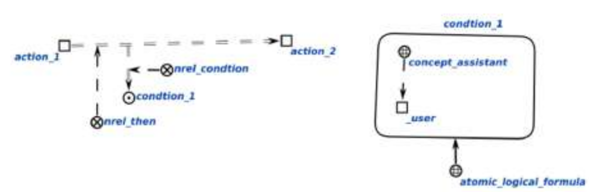


Figure 3. Example of writing a conditional transition between two actions

However, in order to constantly replenish the *knowledge base* during the dialog process, it is necessary to extract *knowledge* from the user's messages and immerse them into the *knowledge base*. For this purpose, the problem solver of the described system implements an action whose problem is to transform the natural-language text of messages into *knowledge base* constructs.

Such a solution allows the system to better utilize knowledge for a more accurate and coherent answer, taking into account the previously mentioned entities and understanding how they relate to the current topic of conversation. This makes the dialog with the system more productive and natural, similar to talking to a person who remembers all the details to solve a problem.

III. Example of application of the results obtained

An example of realization of the proposed approach is an intelligent geometry learning system consisting of three components: a KB, a problem solver and a UI.

Below is a decomposition of the *Geometry Intelligent System Problem Solver*. It consists of a search module — agents that search for constructs in the KB and a computational problem solver, i.e., agents that implement algorithms for solving geometry problems.

Solver of intelligent system problems in geometry
⇒ *decomposition of an abstract sc-agent**:
{• *Abstract non-atomic sc-agent search agent*
• *Computational Problem Solver*
⇒ *decomposition of an abstract sc-agent**:
{• *Abstract sc-agent for interpreting arithmetic expressions*
• *Abstract non-atomic sc-agent for interpreting logical rules*
• *Abstract sc-agent of constructing a strategy for finding a solution to a problem in width*
}
}

Next, a decomposition of *Abstract non-atomic sc-agent search* is presented. It consists of a necessary set of sc-agents that can be used in solving specific problems. For example, if a theorem proving problem needs to be solved, it is appropriate to use *Abstract sc-agent for searching axioms of a given ontology* or *Abstract sc-agent for searching theorems of a given ontology*.

Abstract non-atomic sc-agent search agent
⇒ *decomposition of an abstract sc-agent**:
{• *Abstract sc-agent for finding an annotation for a given section*
• *Abstract sc-agent for searching axioms of a given ontology*
• *Abstract sc-agent for searching theorems of a given ontology*
• *Abstract sc-agent for finding direct links between two objects*
• *Abstract sc-agent for searching concepts through which a given concept is defined*
• *Abstract sc-agent for searching the scope of a relation definition*
• *Abstract sc-agent to find a definition or explanation for a given object*
• *Abstract sc-agent for finding examples for a given concept*
• *Abstract sc-agent for finding a formal statement record for a given statement sign*
• *Abstract sc-agent for finding illustrations for a given object*
• *Abstract sc-agent of finding key sc-elements for a given subject area*
• *An abstract sc-agent searches for concepts that are defined on the basis of a given*
• *Abstract sc-agent search for all constructs isomorphic to a given pattern*
• *Abstract sc-agent for finding the sc-text of a proof for a given assertion*
• *Abstract sc-agent for searching relations defined on a concept*
• *Abstract sc-agent for searching sc-text of condition and problem solution*
• *Abstract sc-agent for searching statements about an object*
}

The following is a decomposition of the *Abstract non-atomic sc-agent problem solver*.

Abstract non-atomic sc-agent problem solving agent
⇒ *decomposition of an abstract sc-agent**:
{• *Abstract sc-agent for searching the value of an unknown quantity*
• *Abstract sc-agent for verifying the truth of an assertion*
• *Abstract sc-agent application of problem-solving strategies*
• *Abstract sc-agent of performing logical inference*
• *Abstract non-atomic sc-agent for calculating mathematical expressions*
⇒ *decomposition of an abstract sc-agent**:
{• *Abstract sc-agent for coordinating the calculation of mathematical expressions*
• *Abstract sc-agent for degree expansion, root extraction and finding the natural logarithm*
• *Abstract sc-agent for addition and subtraction of quantities and numbers*
• *Abstract sc-agent of product and division of quantities and numbers*
• *Abstract sc-agent for comparing quantities and numbers*
• *Abstract sc-agent for calculating trigonometric expressions*
}
}

Abstract sc-agent for coordinating the calculation of mathematical expressions takes a formula as a parameter. An example is shown below in Fig. 4.

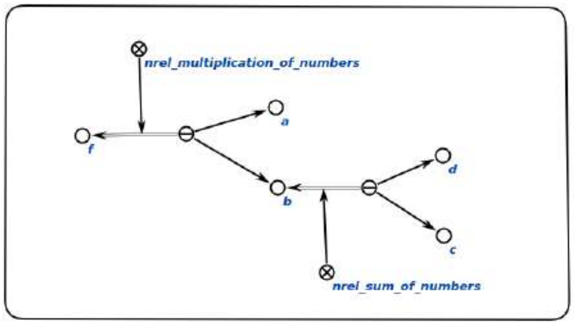


Figure 4. Example formula as input parameter of sc-agent

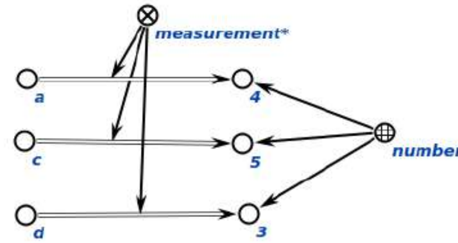


Figure 5. Example formula as input parameter for sc-agent

A formula is represented as an sc-structure, which contains sc-bindings of mathematical operation relations and sc-nodes, which are signs of numbers or variables whose value is known or to be calculated. In this example, the formula consists of:

- variables:
  - a,
  - b,
  - c,
  - d,
  - f,
- relations:
  - nrel_sum_of_numbers,
  - nrel_multiplication_of_numbers,
- arcs and edges.

The abstract sc-agent of coordination of calculation of mathematical expressions of formula processing after initiation searches for all sc-edges of relations of arithmetic operations and forms a structure for calling the operation calculation agent with the parameter of the sc-edges connecting the node of the relation and the binary arc of the basic kind. In turn, each of the sc-agents for operation computation checks whether it can compute an operation of the given type. If it can, it computes the operation and creates an sc-node with the answer, otherwise it does not continue its work.

Thus, the abstract sc-agent coordinating the computation of mathematical expressions processing formula does not know in advance which agent to call specifically. All agents react to the initiated action by checking the input parameters as the initial condition of the problem.

The values of the variables in the formula can be specified in advance, or they can be found in the course of solving the problem. Below are the steps of calculating the values of the variables in the formula, if the values of a, d, c are known in advance (otherwise the formula would not have a specific value).

Since the values of a and b must be known to compute f, the sc-agent checks if their values are known. Since the value is known only for a, the agent will generate a

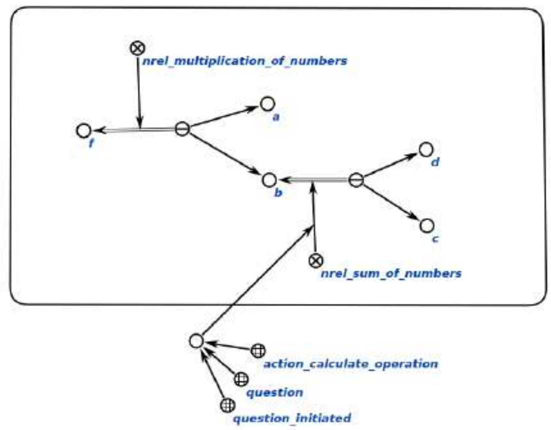


Figure 6. Example of operation calculation agent initiation

structure to initiate the agent to compute an arithmetic operation, after which the value of b will be known.

An example of such a structure for initiating the sc-agent is given below in Fig.6.

Once completed, the agent will create the following construct in the KB. 7

Thus, if all initiated agents are successfully executed, the value of the value of f.

This agent can be used to calculate the values of area, perimeter, etc. using predetermined formulas.

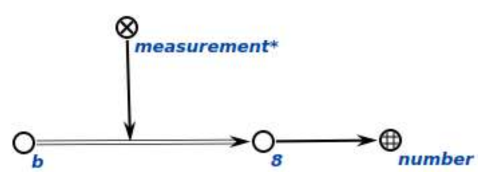


Figure 7. Example of the result of the operation calculation agent execution

IV. Conclusion

The paper proposes an approach to the development of methods and means of constructing plans for problem solving in ISs, which allows us to improve the accuracy of answers, as well as to overcome the shortcomings of modern ISs.

The proposed model allows us to consider the developed problem solver at different levels of detail, which provides the possibility of step-by-step design of solvers, as well as their modifiability.

Classification and specification of actions, problems are specified.

The architecture is considered and the IS itself, realizing the proposed approach, is described.

The obtained results will allow to increase the efficiency of designing ISs and means of automating the development of such systems, as well as to provide an opportunity not only for the developer, but also for the IS to automatically supplement the system with new knowledge and skills.

Acknowledgment

The authors would like to thnk the research groups of the Department of Intelligent Information Technologies of the Belarusian State University of Informatics and Radioelectronics.

References

[1] I. Z. Batyrshin, *Fuzzy hybrid systems. Theory and practice*, H. G. Y. M., Ed. Physmatlit, 2007.

[2] T. Pratt, *Programming languages: development and realization : transl. from English / T. Pratt, M. Zelkovits*, A. Matrosov, Ed. SPb. Peter Print, 2002.

[3] L. A. Gladkov, *Genetic algorithms : textbook*, V. M. K. n. r. L. A. Gladkov, V. V. Kureichik and supplementary ed. M., Eds. Physmatlit, 2006.

[4] V. V. Emelyanov, *Theory and practice of evolutionary modeling*, V. M. K. V. V. Emelyanov, V. V. Kureichik, Ed. Physmatlit, 2003.

[5] M. B. Berkinblit, *Neural networks : an experimental textbook*, M. B. Berkinblit, Ed. MIROS : All-Russian extramural multidisciplinary school of the Russian Academy of Sciences., 1993.

[6] V. A. Golovko, *Neural networks: training, organization, and applications : a textbook*, V. A. Golovko, Ed. Journal. "Radiotekhnika", 2001.

[7] A. N. Gorban, *Neural networks on a personal computer*, D. A. R. A. N. Gorban, Ed. Novosibirsk : Nauka, 1996.

[8] V. N. Vagin, *Reliable and plausible inference in intelligent systems*, 2nd ed., D. A. P. B. N. Vagina, Ed. Fizmatlit, 2008.

[9] B. A. Kulik, *The logic of natural reasoning*, B. A. Kulik, Ed. St. Petersburg. : Nev. dialect, 2001.

[10] D. Poya, *Programming languages: design and implementation : transl. from English*, M. Z. e. b. A. M. T. Pratt, Ed. Peter Print, 2002.

[11] I. Z. Batyrshin, *Basic operations of fuzzy logic and their generalizations*, I. Z. Batyrshin., Ed. Kazan : Fatherland, 2001.

[12] N. P. Demenkov, *Fuzzy control in technical systems : textbook*, N. P. Demenkov, Ed. Moscow State Technical University Publishing House, 2005.

[13] D. A. Pospelov, *Modeling reasoning: experience in analyzing thought acts*. Radio and communications, 1989.

[14] R. A. Reiter, *A logic for default reasoning*, R. Reiter, Ed. Artificial Intelligence., 1980, vol. 13, no. 13.

[15] A. P. Eremeev, *Construction of ternary logic-based decision functions in decision-making systems under uncertainty*, A. P. Eremeev, Ed. Famous Russian Academy of Sciences. Theory and systems of management., 1997.

[16] A. Tamkin, M. Brundage, J. Clark, and D. Ganguli, "Understanding the capabilities, limitations, and societal impact of large language models," *CoRR*, vol. abs/2102.02503, 2021. [Online]. Available: https://arxiv.org/abs/2102.02503

[17] K. Bantsevich, M. Kovalev, and N. Malinovskaya, "Integration of large language models with knowledge bases of intelligent systems," *Otkrytye semanticheskie tekhnologii proektirovaniya intellektual'nykh system [Open semantic technologies for intelligent systems]*, pp. pp 213–219, 2023.

[18] V. V. Golenkov and N. A. Gulyakina, "Next-generation intelligent computer systems and technology of complex support of their life cycle," *Otkrytye semanticheskie tekhnologii proektirovaniya intellektual'nykh system [Open semantic technologies for intelligent systems]*, pp. 27–40, 2022.

[19] M. V. Kovalev, "Convergence and integration of artificial neural networks with knowledge bases in next-generation intelligent computer systems," *Otkrytye semanticheskie tekhnologii proektirovaniya intellektual'nykh system [Open semantic technologies for intelligent systems]*, pp. 173–186, 2022.

[20] (2023, March) Opinion | Noam Chomsky: The False Promise of ChatGPT. [Online]. Available: https://www.nytimes.com/2023/03/08/opinion/noam-chomsky-chatgpt-ai.html

[21] V. Golenkov, N. Gulyakina, and D. Shunkevich, *Open technology for ontological design, production and operation of semantically compatible hybrid intelligent computer systems*, G. V.V., Ed. Minsk: Bestprint, 2023.

[22] A. N. Averkin, *A Comprehensive Dictionary of Artificial Intelligence*, H. G. Yarushina, Ed. Radio and communications, 1992.

[23] "Github:scl-machine [electronic resource]." [Online]. Available: https://github.com/ostis-ai/scl-machine

МЕТОДЫ И СРЕДСТВА ПОСТРОЕНИЯ ПЛАНОВ РЕШЕНИЯ ЗАДАЧ В ИНТЕЛЛЕКТУАЛЬНЫХ СИСТЕМАХ НА ПРИМЕРЕ ИНТЕЛЛЕКТУАЛЬНОЙ СИСТЕМЫ ПО ГЕОМЕТРИИ

Малиновская Н. В., Макаренко А. И.

В данной работе предлагается подход к разработке методов и средств построения планов решения задач в интеллектуальных системах на примере интеллектуальной системы по геометрии. Описанный подход направлен на повышение точности ответов засчет возможности декомпозиции задач на более простые, а также направлен на преодоление недостатков современных интеллектуальных систем. Описана интеллектуальная система, реализующая предлагаемый подход.

Received 01.04.2024

Bringing the Subject Domain Ontology to Optimal Canonical Form

Anatoli Karpuk
Belarusian State Academy of Communications
Minsk, Belarus
Email: a_karpuk@mail.ru

***Abstract*—A formal definition of the subject domain ontology is given. The concept of the canonical form of a subject domain ontology is considered, which is built on the basis of an analysis of functional dependencies between concepts and properties of the subject domain ontology concepts. An algorithm for bringing the subject domain ontology to a canonical form is described. The concept of the optimal canonical form of a subject domain ontology is introduced, containing the minimum number of classes and the minimum number of attributes in the classes. A method for bringing the subject domain ontology to the optimal canonical form is roposed.**

***Keywords*—ontology, domain, canonical form, functional dependence, optimal canonical form**

I. Introduction

In computer science, ontology is a comprehensive and detailed formalization of a certain area of knowledge in the form of a conceptual diagram. A conceptual schema is a set of concepts and information about concepts, which includes properties, relationships, restrictions, axioms and statements about concepts necessary to describe the processes of solving problems in a selected subject domain. For each knowledge area, an applied ontology is built, which consists of a top-level ontology, a subject domain ontology, and a task ontology. Subject domain ontology are simultaneously developed and used by many users. For this reason, in subject domain ontology, the same property of a concept can be represented in different ways. Such ambiguity can lead to difficulties when solving problems using subject domain ontology. To eliminate this drawback, works [1], [2] propose to bring the subject domain ontology to the so-called canonical form, which is based on the analysis of functional dependencies between the concepts and properties of the ontology. However, the developed algorithm for bringing the ontology to a canonical form also does not provide a unique solution. Depending on the order in which the functional dependencies between attributes are analyzed, it is possible to obtain a different number of classes with different numbers of attributes in them. This article introduces the concept of an optimal canonical form of a subject domain ontology, containing a minimum number of classes and a minimum number of attributes in classes, and proposes a method for bringing a subject domain ontology to an optimal canonical form.

II. Formal Definition the Subject Domain Ontology

Let us define the subject domain ontology in the form of a quadruple $O = \langle K, R, F, I \rangle$ [3], [4] where K is a finite set of concepts of the subject domain ontology; R – a finite set of relations between concepts; F – a finite set of interpretation functions defined on concepts and relationships; I – a finite set of axioms, each of which is always a true statement on concepts and relations. The set of concepts has the form $K = \langle D, A, Q \rangle$, where D is a finite set of domains; A — a finite set of attributes; Q — a finite set of classes in subject domain ontology.

Domains are used as sets of possible attribute values. Each domain's data has one of the data types allowed in the XML language [5]. Based on the number of data elements in the value, domains are divided into atomic, union, and list. An atomic domain consists of indivisible data elements of a specific type and format. The value of a federated domain is a data aggregate (structure) consisting of other aggregates and data elements. The value of any merged domain can be represented as a union of the values of its constituent atomic domains. A list domain value is a list (repeating group) of atomic or concatenated domain values. The number of list elements can be any. The value of any list domain can be represented as a repeating group of values from one or more atomic domains.

Based on value restrictions, atomic domains are divided into primitive, built-in, and constructed. Primitive data types of the XML language are used as primitive atomic domains. Built-in domains are derived from primitive domains by applying fixed constraints to them. For example, the primitive domain decimal produces the built-in domains integer, long, int, short, byte, nonNegativeInteger, positiveInteger, unsignedLong, unsignedInt, unsignedShort, unsignedByte, nonPositiveInteger, negativeInteger. Derived domains are derived from primitive and built-in domains by applying various facets to them. For example, the constraints length, minLength, maxLength, pattern, enumeration, whiteSpace, assertions can be applied to the primitive atomic domain string. The constraints totalDigits, fractionDigits, pattern, whiteSpace, enumeration, maxInclusive, maxExclusive, minInclusive, minExclusive, assertions can be applied to the primitive atomic domain decimal.

Depending on the identification method, domains can be unnamed or named. Unnamed domains do not carry semantic load and are used only as data types. Unnamed domains cannot be merged or list domains. Named domains are distinguished from unnamed domains by having a user-defined domain name and can be atomic primitives, inline and derived domains, as well as federated and list domains.

Concepts from set A represent properties (attributes) of subject domain classes. Each attribute is specified by its unique name and the domain to which the attribute values belong. Depending on what domain the attribute is defined on, it can be atomic, aggregated, or list.

Concepts from set Q represent classes (objects, entities) of the subject area. One class includes real or abstract people, objects, phenomena, events, processes that have the same or similar set of properties (attributes), knowledge about which is stored in the ontology and used when solving problems from a given subject domain. When constructing a subject domain ontology, classes are first described, and then knowledge about the individuals of each class is recorded in the ontology. In Russian-language literature, individuals of classes are often called instances of classes or objects. Each class is given its own unique name and its own set of attributes.

Each attribute within a class can have cardinality and functional properties. Cardinality values indicate the minimum and maximum number of attribute values that one individual of a class can have. By default, a class individual can have any number of attribute values. If an attribute value may not be present in an individual of a class, then the ontology must indicate a minimum cardinality equal to 0. The functionality sign shows that any individual of a class can have no more than one attribute value. If the functionality attribute is set, then the maximum cardinality of this attribute should not be specified, or should be equal to 1. The set of class attributes, the values of which uniquely determine an individual of the class, is declared as the key of the class. A class can have more than one key.

The set of relations between concepts R includes relations between domains for constructing derived domains, relations between attributes and domains for determining the scope of attributes, relations between classes and attributes for determining the composition of the attributes of each class, and relations between classes. Relationships between classes reflect "whole-part", "genus-type" connections, as well as hierarchical and other connections between classes that exist in the subject domain. Each relationship between classes can also have cardinality and functionality properties. In addition, relationships between classes can be inverse, inverse functional, transitive, symmetric, asymmetric, reflexive, and irreflexive.

The set of functions F consists of n-ary relations between classes or attributes in which the value of an element with a number n is uniquely determined by the values of previous $(n-1)$ elements. Using functions, you can describe class keys, hierarchical relationships between classes and attributes, and any other functional dependencies between classes and attributes that exist in the subject domain.

The set of axioms I serves to represent in the ontology statements about classes, attributes, domains and relations that are always true. Each axiom is formulated in the form "if <condition on the values of domains for given attributes of given classes or relations> then <statement about the values of domains for given attributes of given classes or relations>". Axioms are included in the ontology to check restrictions on the values of attributes, to check the correctness of the description of the ontology, to derive new true statements about classes, attributes, domains and relationships.

III. Functional Dependencies between Attributes of the Subject Domain Ontology

Let $X \subset A$ be a subset of attributes of the subject domain ontology, $Z \in A$ — some attribute. We will say that in the subject domain ontology there is a functional dependence (FD) $X \to Z$, if any combination of attribute values from X always corresponds to a single value of the attribute Z.

The FD structure on a set of attributes A satisfies Armstrong's axioms [6]:

if $X \subseteq A$ then $A \to X$ (reflexivity axiom);

if $X \to Y$ and $YC \to D$, then $XC \to D$ (axiom of pseudotransitivity).

If there is a FD $X \to Y$, then they say that X functionally determines Y or Y functionally depends on X. From the given axioms, one can derive a number of properties of the FD structure, which in the literature (for example, [7]) are often also called axioms, although it is more accurate to call them rules of inference. The most important are the following inference rules:

if $X \to YC$, then $X \to Y$ and $X \to C$ (decomposition rule), indeed, by the axiom of reflexivity we have $XYC \to Y$ and $XYC \to C$, then by the axiom of pseudotransitivity we obtain $X \to Y$ and $X \to C$;

if $X \to Y$, then $XC \to YC$ (replenishment rule), indeed, by the axiom of reflexivity we have $XYC \to YC$, then by the axiom of pseudotransitivity we obtain $XC \to YC$;

if $X \to Y$ and $X \to C$, then $X \to YC$ (the union rule), indeed, by the completion rule we have $X \to XY$ and $XY \to YC$, then by the axiom of pseudotransitivity we obtain $X \to YC$.

Obviously, the inclusion relation $\subseteq$ determines the FD structure on the set A, which is called the trivial structure of the FD, and the FDs included in it are called trivial FDs. To specify a FD structure that differs from the trivial one, it is necessary to postulate a finite set of FD $F = \{F_j = X_j \to Y_j \mid X_j \subset A, Y_j \subseteq A, j = \overline{1,m}\}$, which in the article [8] was called a system of generators of the FD structure on the set of attributes A. The FD structure, specified by the system of generators F, will be denoted by $S(F)$.

It is obvious that from $Z \in X$ it follows that $X \to Z$. Such an FD, in which the dependent attribute is part of the left side of the FD, is called trivial. In what follows, we will consider only non-trivial FDs between attributes. In the subject domain ontology, the following non-trivial FDs between attributes can be distinguished:

- each functional attribute of a class that is not a subclass of another class, that is not a subordinate attribute of another class attribute, functionally depends on each class key;
- each functional attribute of a subclass that is not a subordinate attribute of another attribute of a subclass is functionally dependent on each subset of attributes obtained by combining each key of the parent class with each key of the subclass;
- each functional attribute of a class that is not a subclass of another class that is a subordinate attribute of another class attribute functionally depends on each subset of attributes obtained by combining each class key with a parent attribute;
- each functional attribute of a subclass, which is a sub-attribute of another attribute of a subclass, functionally depends on each subset of attributes obtained by combining each key of the parent class with each key of the subclass and the parent attribute;
- each functional relationship between classes from the set R specifies the FD of attributes of each key of the parent class from each key of the subordinate class;
- each function from the set F, defined on the attributes of the subject domain ontology, sets the FD of the last attribute of the relation from the previous attributes of this relation;
- each function from the set F, defined on the classes of the subject domain ontology, specifies the FD of the attributes of each key of the last class of the relation from each subset of attributes containing any one key of the previous classes of this relation.

When defining the FD between the attributes of the subject domain ontology, it is possible to write more than one attribute on the right side of the FD, since for the FD between attributes the property of the cluster decomposition of the right side is valid, namely, if $Y_1 \in A$, $Y_2 \in A$ and $Y = Y_1 \cup Y_2$, then the record $X \to Y$ corresponds to the simultaneous presence of FD $X \to Y_1$ and $X \to Y_2$. Moreover, instead of the last two FD, you can write $X \to Y_1Y_2$. Each FD between attributes in the subject domain ontology can be considered as the simplest rule for deriving new knowledge from the knowledge available in the ontology. Indeed, if the values of the attributes of the left side of the FD are known, then either the ontology already contains uniquely corresponding values of the attributes of the right side of the FD, or the values of the attributes of the right side of the FD can be obtained by solving the problem from the ontology of tasks. The input data of this problem are the values of the attributes of the left side of the FD, and the output data are the values of the attributes of the right side of the FD.

Let us single out in the subject domain ontology all FDs between attributes and represent them as a set of FDs $P = \{P_j = X_j \to Y_j \mid X_j \subset A, Y_j \subseteq A, j = \overline{1,m}\}$, which is called the system of forming FD structures on the set of attributes of the ontology. The structure of the FD, given by the system of generators P, will be denoted $S(P)$.

The closure of the set of attributes $X \subset A$ concerning to the structure of FD $S(P)$ is a set $X^+(P) \subseteq A$ such that for any $Y \subseteq A$ from $X \to Y$ follows $Y \subseteq X^+(P)$. In other words, the closure of the set of attributes X includes all the attributes, the values of which can be obtained from the known values of the attributes of set X, using the FD derivation from the set P. The algorithm for constructing the closure $X^+(P)$ consists of the following steps [8].

1) Put $X^+(P) = X$ and $p_j = 0$, $j = \overline{1,m}$.
2) Put $q = 0$ and for each $j = \overline{1,m}$ perform step 3.
3) If $p_j = 0$ and $X_j \subseteq X^+(P)$ then put $X^+(P) = X^+(P) \cup Y_j$, $q = 1$ and $p_j = 1$.
4) If $q = 1$, then go to step 2, otherwise finish the job.

The structures of FD $S(P^1)$ and $S(P^2)$ on the set of attributes A with systems of generators $P^1 = \{{X_i}^1 \to {Y_i}^1 \mid {X_i}^1 \subset A, {Y_i}^1 \subseteq A, i = \overline{1,m_1}\}$ and $P^2 = \{{X_j}^2 \to {Y_j}^2 \mid {X_j}^2 \subset A, {Y_j}^2 \subseteq A, j = \overline{1,m_2}\}$, respectively, are called equivalent if for any $X \subset A$ the equality $X^+(P^1) = X^+(P^2)$. In article [9] it is proven that necessary and sufficient conditions for the equivalence of structures FD $S(P^1)$ and $S(P^2)$ on the set of attributes A are the fulfillment of equalities ${X_i}^{1+}(P^1) = {X_i}^{1+}(P^2)$ and ${X_j}^{2+}(P^1) = {X_j}^{2+}(P^2)$ for all $i = \overline{1,m_1}$, $j = \overline{1,m_2}$. The system of generators $E = \{H_j \to T_j \mid H_j \subset A, T_j \subseteq A, j = \overline{1,m}\}$ is called an elementary basis of the structure of FD $S(E)$ if the removal of any attribute from the left or right side of any FD from E leads to the structure of FD that is not equivalent to $S(E)$.

IV. Canonical Form of the Subject Domain Ontology

We will say that the subject domain ontology is in canonical form if the following conditions are met [1]:

- all attributes from the set A participating in the definition of classes, functions, and axioms ontology are atomic;
- all attributes of each class have a functionality flag and have no subordinate attributes;
- the system of FD structure generators between the attributes of the ontology is the elementary basis of this FD structure.

The algorithm for reducing the subject domain ontology to the canonical form consists of the following steps.

1) For each composite attribute, add to set A the atomic attributes that make up the com-posite attribute. If this composite attribute is part of some class with a flag of functionality, then replace it in this class with atomic attributes with a flag of functionality. If a composite attribute is a part of a class without a functionality flag (it is a list attribute), then represent it as a subclass consisting of atomic attributes with a functionality flag included in the composite attribute. At the same time, determine the keys of a new class depending on the presence of an FD between atomic attributes within a composite attribute.
2) Each composite attribute included in the functions and axioms of the ontology should be replaced with atomic attributes included in its composition.
3) Each atomic attribute that is part of some class without a flag of functionality, to represent in the form of a subclass consisting of this attribute with a flag of functionality, and the class key consists of this atomic attribute.
4) Each atomic attribute that is part of a class and has subordinate attributes in it should be represented as a subclass consisting of this attribute and subordinate attributes with a flag of functionality. If this attribute was without the functionality flag, then the key of the new class consists of this atomic attribute otherwise all of its attributes are included in the key of the new class.
5) Select non-trivial FD between attributes according to the rules described above and form a system of generators of FD structure on the set of attributes $P = \{P_j = X_j \to Y_j \mid X_j \subset A, Y_j \subseteq A, j = \overline{1,m}\}$.
6) Remove redundant attributes from the left sides of the FD from set P. Attribute $B \in X_j$ is considered redundant in X_j, if $B \in (X_j \backslash B)^+(P)$.
7) Remove redundant attributes from the right sides of the FD from set P. Attribute $B \in Y_j$ is considered redundant in Y_j, if $B \in {X_j}^+(P')$, where P' denotes the system of generators of the FD structure, obtained from P by replacing FD $X_j \to Y_j$ with $X_j \to (Y_j \backslash B)$. As a result of performing steps 6 and 7, an elementary basis of the FD structure on a set of attributes $E = \{H_j \to T_j \mid H_j \subset A, T_j \subseteq A, j = \overline{1,m}\}$ will be obtained.
8) Bring the subject domain ontology in accordance with the obtained elementary basis of the FD structure between attributes by performing the following steps:
   - remove from the classes the attributes that turned out to be redundant in the right parts of the corresponding FD of the elementary basis;
   - remove from the composition of the keys of the classes the attributes that turned out to be redundant in the left parts of the corresponding FD of the elementary basis;
   - unite into one class those ontology classes that have the same closures of their keys concerning the elementary basis of the FD structure;
   - remove functions from the set F, in which all the attributes of the right-hand sides in the corresponding FD of the elementary basis turned out to be redundant;
   - from the left-hand sides of the functions from the set F, remove the attributes that turned out to be redundant in the left-hand sides of the corresponding FD of the elementary basis.

V. Optimal Canonical Form of the Subject Domain Ontology

The canonical form of a subject domain ontology is called optimal if it contains a minimum number of classes with a minimum number of occurrences of attributes in them. The task of bringing the subject domain ontology to optimal canonical form comes down to finding the optimal elementary basis of the structure FD between the attributes of the ontology, which contains the minimum number FD with the minimum number of occurrences of attributes in them.

In article [9], the concept of P-dependencies in the elementary basis of the FD structure was introduced and studied. Let $E = \{H_j \to T_j \mid H_j \subset A, T_j \subseteq A, j = \overline{1,m}\}$ be the elementary basis of the FD structure $S(E)$ on the set A on the set A. We will say that in the FD $(H_s \to T_s) \in E$ there is a P-dependence of a non-empty ${T_s}^{'} \subseteq T_s$ on H_s if there exists an $P \subset {H_s}^+(E)$, $P \neq H_s$ such that ${T_s}^{'} \subseteq (P^+(E) \backslash P)$ and no subset of P possesses these properties. Let E' us denote the system of generators of the FD structure obtained from E by replacing the FD $H_s \to T_s$ with $H_s \to T_s \backslash {T_s}^{'}$. We will distinguish three types of P-dependence: P_1-dependence occurs if $P \subseteq {H_s}^+(E')$, P_2-dependence if simultaneously $P \nsubseteq {H_s}^+(E')$ and ${T_s}^{'} \nsubseteq (P^+(E' \backslash P)$, P_3-dependence if simultaneously $P \nsubseteq {H_s}^+(E')$ and ${T_s}^{'} \subseteq (P^+(E' \backslash P)$.

We formulate the main properties of P-dependencies in the form of the following statements, the proof of which is carried out by checking the fulfillment of sufficient conditions for the equivalence of the FD structures.

If in the FD $(H_s \to T_s) \in E$ there is a P_1-dependence of $T_s^{'} \subseteq T_s$ on H_s, and the system of generators of the structure of the FD Q is obtained from E by replacing the FD $H_s \to T_s$ with $H_s \to T_s \backslash T_s^{'}$ and adding the FD $P \to T_s^{'}$ to the E, then the structures of the FD specified by the systems of generators E and Q are equivalent.

If in the FD $(H_s \to T_s) \in E$ there is a P_2-dependence of $T_s^{'} \subseteq T_s$ on H_s, and the system of generators of the FD structure Q is obtained from E by replacing the FD $H_s \to T_s$ with $H_s \to ((T_s \backslash T_s^{'}) \cup (P \backslash H_s))$ and adding the FD $P \to T_s^{'}$ to the E, then the FD structures specified by the systems of generators E and Q are equivalent.

If in the FD $(H_s \to T_s) \in E$ there is a P_3-dependence of $T_s^{'} \subseteq T_s$ on H_s, and the system of generators of the FD structure Q is obtained from E by replacing the FD $H_s \to T_s$ with $H_s \to ((T_s \backslash T_s^{'}) \cup (P \backslash H_s))$, then the FD structures specified by the systems of generators E and Q are equivalent.

The given properties of P-dependencies make it possible to move from one elementary basis of the FD structure to other elementary bases and find the optimal elementary basis of the FD structure.

In article [10], the concept of a cycle in the elementary basis of the FD structure is introduced and it is proved that the presence of cycles in the elementary basis of the FD structure is a necessary condition for the existence of P-dependencies in the elementary basis. The elementary basis of the FD structure on set A can be associated with a bipartite oriented graph (A, E, H, T), in which A – the set of vertices of the first part of the graph, E – the set of vertices of the second part of the graph, H – the set of arcs of the graph directed from the vertices of the first part to the vertices of the second part of the graph (showing the occurrence of elements from A to the left parts of the FD), T is a set of arcs of the graph directed from the vertices of the second part to the vertices of the first part of the graph (showing the occurrence of elements from A in the right parts of the FD). It is easy to verify that each cycle in an elementary basis corresponds to a family of cycles in the corresponding bipartite graph.

In general, the problem of finding all cycles in a bipartite directed graph is NP-hard, but its difficulty is determined by the fact that the maximum possible number of cycles in a graph depends exponentially on the dimension of the graph. The search time for one cycle in a directed graph using a standard depth-first search algorithm depends linearly on the dimension of the graph. In real optimization problems of the canonical form of a subject domain ontology, with the number of attributes on the order of 10^3, the number of cycles in the elementary basis of the FD structure does not exceed 10^2, therefore all cycles in the elementary basis of the FD structure can be found in an acceptable time.

VI. Software for Bringing the Subject Domain Ontology to Canonical Form

The input data of the software is the subject domain ontology in the OWL-2 language in the input file. The output data is the equivalent subject domain ontology, which is in the canonical form, presented as an owl-file. The software extracts from the original owl-file and presents attributes, classes, class hierarchy, links between attributes and classes in the form of database tables. Then, atomic attributes and the system of forming the FD structure between the attributes are extracted from the database tables. The software for bringing the subject domain ontology to the canonical form includes the Attribute, AttributeSet, FuncDepen, FDStructure classes developed in C++.

The Attribute class is used to represent a single atomic attribute. The class data is an attribute identification number, an attribute name, and an attribute purpose.

The AttributeSet class is used to represent any subset of atomic attributes. The class data is an array of attribute identification numbers. The class methods return the number of attributes in a subset, add an attribute to a subset, remove an attribute from a subset, check for the presence of an attribute in a subset, check if a given subset of attributes is in a subset, get the union and intersection of a given subset of attributes with a subset.

The FuncDepen class is used to represent a single functional dependency between atomic attributes. The data of the class are an object of the AttributeSet class, corresponding to the left part of the FD, and an object of the AttributeSet class, corresponding to the right part of the FD. Class methods add an attribute to the left or right part of the FD, remove an attribute from the left or right part of the FD.

The FDStructure class is used to represent a system of generators and an elementary basis for a structure of functional dependencies between atomic attributes. The class data is an array of objects of the FuncDepen class. The class methods return the number of FDs in the structure, add FDs to the structure, remove FDs from the structure, obtain the closure of a given subset of attributes with respect to the FD structure, and find the elementary basis of the FD structure.

Software for bringing the subject domain ontology to the optimal canonical form is under development.

The software for bringing the subject domain ontology to canonical form was used in the development of the domain ontology of radio communication networks. As a result, the main ontology classes and their subclasses were obtained.

The TransmiterTypes class and its subclasses are designed to store data about transmitter types. The classes include attributes: name of the transmitter type, boundaries of operating ranges frequencies, power range boundaries, emission codes, emission bandwidth for each emission code, dependence of the attenuation of out-of-band and noise emissions on detuning from the operating frequency, attenuation of radiation at harmonic and reference oscillator frequencies, attenuation of radiation at combination and intermodulation frequencies.

The ReceiverTypes class and its subclasses are designed to store data about receiver types. The classes include attributes: the name of the receiver type, the boundaries of the operating frequency ranges, sensitivity, codes of received emissions, bandwidth and the required signal-to-noise ratio for each received radiation, the dependence of the sensitivity attenuation on detuning from the operating frequency, intermediate frequencies, the radiation power of the receiver at local oscillator frequencies, weakening of sensitivity at intermediate frequencies, local oscillator frequencies, mirror local oscillator frequencies, combination and intermodulation frequencies.

The AntennaTypes class is designed to store data about antenna types. The class includes the following attributes: antenna type name, antenna type code, polarization code, minimum and maximum electrical center height, isotropic gain for horizontal and vertical polarization, antenna half power beamwidth in horizontal and vertical plane, side lobe attenuation relative to to an isotropic antenna, attenuation in the feeder.

The RadioDevTypes class is designed to store data about types of radio devices (RD). The class includes the following attributes: name of the RD type, code of the RD type (transmitter, receiver, radio station), code of the RD operating mode (simplex, duplex).

The ObjectCommInds class is designed to store data about individuals of communication objects. The class includes the following attributes: name of the communication object, geographic coordinates of the center of the stationary communication object or the center of the movement zone of the mobile communication object, the radius of the movement zone of the mobile communication object, the height of the point of standing of the stationary communication object above sea level.

The AntennaInds class is a subclass of the AntennaTypes class and is intended to store data about antenna individuals. The class includes the following attributes: the name of the antenna, the height of the electrical center of the antenna above the communication object, the direction angles of the antenna in the horizontal and vertical plane, the coordinates of the antenna relative to the center of the communication object.

The Radiolines class and its subclasses are designed to store data about radio lines (RL) and radio networks. The classes include the following attributes: name of RL, RL importance code, RL type code (with fixed radio frequencies, with pseudo-random switching of the operating frequency, radio relay line interval), RL operating mode code (simplex, duplex with time division, duplex with frequency division), codes of emissions used in RL, frequencies or average frequencies of frequency bands assigned to RL for transmitting and receiving in the main RD.

VII. Conclusion

To eliminate ambiguity and redundancy in the domain ontology, the concept of the canonical form of the domain ontology was introduced and algorithms were proposed for bringing the domain ontology to a canonical form and an optimal canonical form. Software has been developed that implements bringing the domain ontology to a canonical form.

References

[1] A. A. Karpuk and A. V. Havorka, “Bringing the Subject Domain Ontology of Radio Communication Networks to Canonical Form”, Problems of Infocommunications, № 2(14), pp. 25–30, 2021, (in Russian).

[2] A. V. Havorka and A. A. Karpuk, “Reduction the Subject Domain Ontology to Canonical Form”, International Journal of Information and Communication Technologies. Special Issue, pp. 43–47, May 2022.

[3] A. V. Palagin, S. L. Kryvy and N. G. Petrenko “Ontological Methods and Means of Processing Subject Knowledge: Monograph”, Lugansk, 324 p., 2012, (in Russian).

[4] A. A. Karpuk and A. V. Havorka, “Construction of an Applied Ontology of Radio Communication Networks”, Vestnik suvjazi, № 6, pp. 36–40, 2021, (in Russian).

[5] “W3C XML Schema Definition Language (XSD) 1.1 Part 2: Datatypes. W3C Recommendation”, 5 April 2012 [Electronic resource] URL: http://www.w3.org/TR/xmlschema11-2.

[6] W. W. Armstrong, “Dependency structure of data base relationships”, Proc. IFIP Congress. Geneva, Switzerland, pp. 580–583, 1974.

[7] S. D. Kuznecov, “Database. Models and Languages”, M., Binom–Press, 720 p., 2008, (in Russian).

[8] A. A. Karpuk and V. V. Krasnoproshin, “Methodology of Data Domain Description for Databases Design in Complex Systems”, International Academy Journal Web of Scholar, Vol. 1, № 4(13), pp. 11–20, 2017.

[9] A. A. Karpuk, “Analysis of Structure of Functional Dependencies between Attributes of a Relational Database”, Economics and Management of Control Systems, № 3(25), pp. 64–70, 2017, (in Russian).

[10] A. A. Karpuk and V. V. Krasnoproshin, “Cycles in Structures of Functional Dependencies”, International Journal of Open Information Technologies, Vol. 5, № 7, pp. 38–44, 2017.

ПРИВЕДЕНИЕ ОНТОЛОГИИ ПРЕДМЕТНОЙ ОБЛАСТИ К ОПТИМАЛЬНОЙ КАНОНИЧЕСКОЙ ФОРМЕ

Карпук А.

Дано формальное определение онтологии предметной области. Рассмотрено понятие канонической формы онтологии предметной области, которая строится на основе анализа функциональных зависимостей между понятиями и свойствами понятий онтологии. Описан алгоритм приведения онтологии предметной области к каноническоой форме. Введено понятие оптимальной канонической формы онтологии предметной области, содержащей минимальное количество классов и минимальное количество атрибутов в классах. Предложен метод приведения онтологии предметной области к оптимальноой каноническоой форме.

Received 13.03.2024

Designing an IoT Network for the Diagnosis of Alzheimer's Disease Using OSTIS

Vishniakou Uladzimir
Belarusian State University of Informatics and Radioelectronics
Minsk, Belarus
vish@bsuir.by

Chuyue Yu
Belarusian State University of Informatics and Radioelectronics
Minsk, Belarus
ycy18779415340@gmail.com

***Abstract*—The report is devoted to the development of the Internet of Things for the diagnosis of Alzheimer's disease (AD) using OSTIS technology. The structure of the ontology for describing the elements of AD disease is given. The article considers the construction of an IT diagnostic network of BA, which uses the semantic capabilities of the OSTIS platform for processing and analyzing medical data. The elements of describing the knowledge base, solvers and user interfaces using a component-based design approach are presented.**

***Keywords*—IoT network, Alzheimer's disease, user interface, IT diagnosis, ostis**

I. Introduction

In the contemporary era, marked by the advent of information technology, the progression of technological and scientific theories has led to the digitization and informatization of traditional medicine [1]. Smart hospitals, relying on an environment founded upon information and communication technologies, especially those optimized and automated by the Internet of Things (IoT) [2], have enhanced the efficiency and reliability of IT systems. IT diagnostics refers to the process of using information technology (IT) tools for diagnosing, analyzing, and solving technical issues. In the medical field, it involves the use of information technology, such as artificial intelligence, machine learning, and data analysis, to aid in medical diagnosis. This technology has the potential to improve the quality and safety of healthcare services [3]. Given that each disease requires a complex medical process, the IoT infrastructure must adhere to the medical rules and steps involved in the diagnostic process to meet the requirements of healthcare providers [4].

An IT network for intelligent detection of Alzheimer's disease can be defined as a comprehensive information technology system that integrates the collection, processing, pattern recognition, and interactive user interfaces of medical health data. This data may involve a variety of multimodal data, such as human behavior data, brain imaging data, or sound audio data, etc., providing support in the process of identifying and predicting signs of Alzheimer's disease, with system tasks that can be broken down into:

1) System Analysis and Simulation Task: The intelligent detection system for Alzheimer's disease analyzes data from diverse sources to build IT network, aiming to identify characteristics unique to Alzheimer's patients.
2) Prediction Task: By using information extracted from features, it effectively distinguishes between Alzheimer's patients and healthy control groups, achieving early detection or prediction of disease progression.
3) Management Task: Management of patient data in the Alzheimer's detection system involves ensuring the privacy and security of the data while providing necessary data to doctors and researchers to support the decision-making process.

The purpose of this paper is to introduce an approach that integrates IT diagnostic networks with OSTIS technology. The objective of this work is to enhance data interpretation capabilities and apply the principles of intelligent disease detection. Through a highly structured and semantic approach, this method aims to improve information processing, knowledge representation, and intelligent decision-making capabilities, enabling the model to be seamlessly optimized and utilized across various scenarios.

II. The ontology

OSTIS (Open Semantic Technology for Intelligent Systems) is closely related to ontology as a framework aimed at developing and implementing intelligent systems. The use of semantic technologies and ontology advocated by it finds application in IoT networks.

The ontology of the knowledge base of the intelligent Alzheimer's detection system can be subdivided into:

1) Theoretical concepts of Alzheimer's recognition
2) Applications of Alzheimer's recognition theory

To achieve interoperability between intelligent systems, it is necessary to develop an ontology for the theoretical concepts of Alzheimer's recognition. This ontology will cover the framework and basic theoretical foundation of Alzheimer's recognition, defining a hierarchical system to build the foundation for IT networks for Alzheimer's recognition, enabling these knowledge to be efficiently shared, understood, and utilized in IT networks.

Alzheimer's Recognition Theory Concept Ontology

:= [A collection of theoretical concepts of Alzheimer's recognition.]

:= [Instances of theoretical objects in the subject area of "Alzheimer's recognition" include key concepts and their relationships.]

⊂ *Ontology*

∋ *Specialized Terminology Categories for Alzheimer's Disease*

∋ *Data Form Categories for Alzheimer's Disease*

∋ *Feature Extraction Categories for Alzheimer's Disease*

∋ *Diagnostic Method Categories for Alzheimer's Disease*

Specialized Terminology Categories for Alzheimer's Disease

:= [Describes specialized terminology related to Alzheimer's, from the domain perspective, including corresponding medical terms and technical synonyms, defining the basic concepts, symptoms, and progression stages of the disease.]

:= [The pathological phenomenon existing in the human brain in the form of a neurodegenerative disease, whose origins can be genetic, environmental factors, or an interaction of both. At a given time point, the pathological state is determined through clinical assessment and biomarker testing, characterized by cognitive dysfunction, memory decline, executive function impairment, and language difficulties. Specialized terminology for Alzheimer's not only describes the clinical manifestations and neuropathological features but also includes concepts related to disease diagnosis and recognition process.]

∋ *Disease Process*

The disease process, referring to the entire process of disease development and its impact on an individual's health status, starting from its initial factors, including the mechanisms of disease onset, development stages, symptoms produced, and the final outcome (recovery, persistence, or deterioration). To describe and understand the disease process and its manifestations in different contexts, here is a hierarchical structure description of subcategories classified by different dimensions or properties for the disease [5].

Classified disease-relevant process by properties

⇒ *subdividing**:

{• *disease-relevant process by Frequency*

:= [The frequency or commonness of the disease process.]

⇒ *includes**:

{• *communly frequent disease-relevant process*

• *frequent disease-relevant process*

• *occasional disease-relevant process*

• *very frequent disease-relevant process*

}

• *disease-relevant process by pathogenic intensity*

:= [The intensity or severity of the disease process's impact on individual health.]

⇒ *includes**:

{• *highly impacting disease-relevant process*

• *lowly impacting disease-relevant process*

• *midly impacting disease-relevant process*

}

• *disease-relevant process by specificity*

:= [The degree of association of the disease process with specific diseases, conditions or factors.]

⇒ *includes**:

{• *Alzheimer Disease disease-relevant process*

• *Alzheimer Disease type disease-relevant process*

• *individual disease-relevant process*

• *life disease-relevant process*

• *life-style disease-relevant process*

• *neurodegenerative disease disease-relevant process*

• *pathogenic disease-relevant process*

}

}

In this way, the disease process is organized into different subcategories, each defined by a shared set of properties. Based on this structure, the following subcategory "identified disease-relevant process" supplements specific content and examples, including processes that have been confirmed in disease research to have a specific impact or role.

Identified disease-relevant process
⇒ *includes**:
{
- *environmental disease-relevant process*
- *gene-related disease-relevant process*
- *molecular disease-relevant process*
- *other disease-relevant process*
- *pathological disease-relevant process*
- *physiological disease-relevant process*

}

III. IT diagnostics network structure

When adopting IT diagnostics for the intelligent identification of neurological diseases, it is necessary to construct an IoT network structure tailored to the specific neurological condition. Based on the analysis, Alzheimer's disease is characterized by long-term progression and gradual worsening, affecting various aspects of an individual's cognitive functions, motor skills, and daily living abilities. It typically requires long-term management and treatment, and there is currently no possibility of complete cure through short-term treatment, making early stage identification and intervention particularly important.

Utilizing acoustic features for the automated detection of Alzheimer's disease is a promising research area. Broadly speaking, the application of IoT networks in IT diagnostics can be divided into three key steps: collection, processing, and analysis of sensor data. To achieve automation in detection tasks, we are developing an IoT network:

1) Sensor data collection: When users enter the client interaction page, the main page displays the required steps and instructions. Participants respond or describe based on the given questions or images, and the system collects and saves their voice data, which is then uploaded to a test library for assessment. The data can be stored locally, uploaded to the cloud, or decentralized.
2) Sensor data processing: Data processing can occur locally or on cloud servers, mainly involving data cleaning, extracting acoustic features, or preprocessing voice data.
3) Sensor data analysis: The extracted data features are sent to a previously trained model prediction agent, where machine learning algorithms deployed in the model provide analysis results. These data are then sent to servers for storage and also serve as a training set to improve the model's accuracy.

The client application displays the likelihood of a participant having Alzheimer's disease on the screen. The doctor's mobile application, on every device that has the app installed, shows real-time patient monitoring information from sensors in the cloud database. The detection results for participants are sent based on the user's registration information. The overall structure of the proposed framework is shown in Fig. 1, illustrating the network's basic structure, which consists of datasets, model prediction agents, participant data, the IoT platform, and mobile applications.

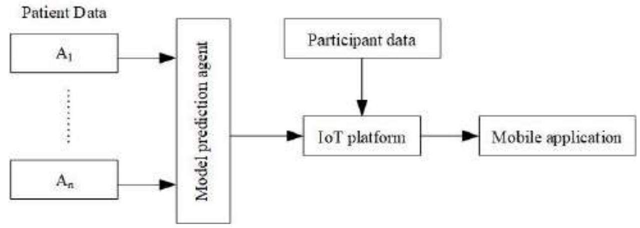


Figure 1. Overall structure of the proposed network.

IV. Distinguishing Voice Characteristics of Alzheimer's Patients

Taking the application of IoT technology and voice analysis for the diagnosis of neurological diseases as an example, the human ear is incapable of perceiving variations in sound and voice rhythm. However, advancements in technology enable automatic voice analysis to identify and extract acoustic and temporal parameters [6], making voice analysis of Alzheimer's patients a focal point of scholarly research. Clinically, these patients exhibit symptoms of unclear speech characterized by slowness, fluctuation, monotony, trembling, hesitation, weakness, inability to control over-breathing, loss of voice, low melody levels, and slower rhythms [7]. In semantic tasks, patients with Alzheimer's disease also perform poorly, often confusing names or unable to accurately name objects [8]. Non-verbal indicators predictive of dementia include the continuity of speech, the duration and proportion of silent segments, pitch (periodicity) related parameters. Useful linguistic markers require automatic voice recognition for assessment, including the richness of speech and the proportion of different parts of speech [9].

Based on the aforementioned voice characteristics of Alzheimer's patients, research in [10] has demonstrated the feasibility of using machine learning methods to distinguish between the voice transcripts of Alzheimer's patients and healthy individuals by analyzing the semantic information in patients' natural language expressions. By converting voice data into text and analyzing these texts with a Random Forest classifier, researchers were able to identify specific language features associated with Alzheimer's disease.

V. Proposed approach

In the development of an IT network for Alzheimer's disease diagnosis using OSTIS, the core components involve the development of a problem solver, knowledge base, and user interface. This is due to the OSTIS system architecture, which integrates a semantic model interpretation platform with a semantic model of the system itself described in SC-code, encompassing sub-models for the knowledge base, interfaces, and problem solvers. A detailed description follows:

A. Development of the Solver and Knowledge Base.

The introduction of SC-code by OSTIS technology serves as a universal language for semantic expression of information in intelligent computing systems. It employs set theory for defining semantic aspects and graph theory for constructing syntactic structures, ensuring compatibility across different types of knowledge through semantic memory (sc-memory). This approach not only simplifies the description of various types of knowledge and models but also allows for the construction of an ontological model (sc-model) of any entity based on SC-code.

Developing a problem solver and knowledge base, especially for Alzheimer's disease using OSTIS technology, involves several key steps:

a) Knowledge Representation: This step involves defining knowledge about Alzheimer's disease, such as symptoms, diagnostic criteria, etc., and how this knowledge is represented within the OSTIS framework. The method of representation includes creating a semantic network to model the relationships and entities related to Alzheimer's disease.
b) Data Collection and Organization: Gathering and documenting data, where the structure of the data must be consistent with the chosen method of knowledge representation.
c) Knowledge Base Development: Importing data into the OSTIS system to establish a knowledge base, which may include concepts, predicates, and rules for controlling the logic of diagnosing Alzheimer's disease within the system.
d) Solver Implementation: Developing the solver involves creating algorithms that can navigate the knowledge base to provide diagnoses based on input data. This includes implementing reasoning mechanisms that can assess patient data against the knowledge base to identify the presence of Alzheimer's disease.
e) Data Storage: In compliance with medical data storage regulations, the system needs not only to possess a static knowledge base but also to have the capability to store dynamic data such as patient records and diagnostic results.

The semantic knowledge base is the cornerstone of semantic communication, forming the foundational part of the decision-making system. The data it contains are authoritative, referential, and diverse. For different domains, the objects encompassed by the knowledge base vary, aiming to integrate scattered knowledge within the domain, including historical data, real-time data, statistical analysis results, predictive models, as well as expert knowledge and experience. Through computer technology, originally scattered knowledge is recombined and stored in different locations, enabling decision-makers to quickly access relevant information and resources.

In application, the development of the knowledge base is primarily based on the information extracted, combined with the OSTIS framework, to construct a knowledge graph. This graph is further processed and refined to form a structured knowledge base. Storing semantically linked data about neurological diseases, it provides the necessary knowledge support for disease prediction.

System Knowledge Base
$\supset$ *Declarative Knowledge*
 $\ni$ *Alzheimer's Recognition Theory Concept Ontology*
$\supset$ *Procedural Knowledge*
 $\supset$ *System Knowledge*
  $\Rightarrow$ *splitting**:
   {• *Model Building Knowledge*
   • *Model Optimization Knowledge*
   • *Results Display Knowledge*
   }
 $\supset$ *Static Knowledge*
  $\ni$ *System Process*
 $\supset$ *Dynamic Knowledge*
  $\ni$ *Conditional Response in Specific Situations*

The problem solver performs basic actions of directly modifying and managing the knowledge base, offering a more dynamic component compared to the knowledge base itself. It handles information stored in the knowledge base to address specific problems, applying algorithms, rules, and reasoning techniques to interpret, analyze, and derive conclusions or solutions from available knowledge, making decisions or generating new knowledge. Semantically, these operations are considered actions performed within the memory of the acting entity, typically the OSTIS system itself, with the knowledge base viewed as its memory. Actions are executed based on the set problem, describing the internal state of the intelligent system or the required state of the external environment for different problem categories.

Figure 2 presents the SCG-code description of the speech classification task within the set-theoretical ontology.

Within the OSTIS Technology framework, to achieve the task of natural language processing, a specialized ostis-system can be developed. The sc-model of the problem solver is developed as a hierarchical system of agents (sc-agents), providing flexibility and modularity for the developed sc-agents. Abstract sc-agents are func-

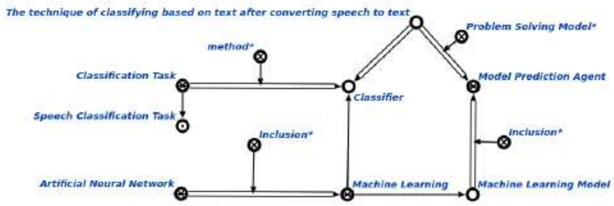


Figure 2. A fragment of the set-theoretic ontology for speech classification tasks.

tionally equivalent sc-agents of a certain category, with different instances implemented in various programming languages to address specific problems. In our work, to implement an intelligent IT diagnostic system for Alzheimer's disease integrated with OSTIS, converting natural language speech into diagnostic indicators required describing the entire process of natural language processing and building extraction rules. Since in the OSTIS technology framework, every action internally represents some transformation executed by a specific sc-agent (or a group of sc-agents), we consider the sc-model of the problem solver. The entire data processing workflow is a core component of the system, ensuring efficient and accurate collection, transmission, processing, and storage of data. An example of processing users' data within the system is shown in Figure 3.

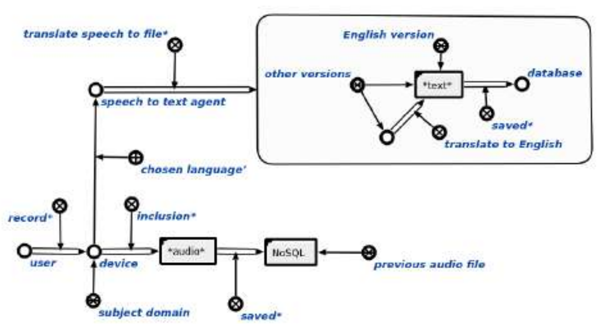


Figure 3. An example of processing users' data within the system.

B. Designing the User interface.

Designing the user interface for the diagnostics system requires careful consideration of the end-users, typically healthcare professionals and possibly patients or their families. The interface should be intuitive and facilitate easy navigation through the diagnostic process:

1) Information Input: Designing forms and input fields that allow for easy entry of patient data and symptoms.
2) Diagnostic Process Visualization: Providing visual cues or progress indicators that guide the user through the diagnostic process, helping them understand what stage the diagnosis is at and what steps are next.
3) Results Presentation: Displaying the diagnostic results in a clear, understandable format. This could include a summary of findings, confidence levels, and recommendations for further actions or treatments.

Due to the use of component approach, the development of the entire natural language interface comes down to development and improvement of separate specified components (e.g. knowledge base on natural language processing, component for natural language texts generation). The model of the process of responding to user needs and the components of inference engine was shown in figure 4.

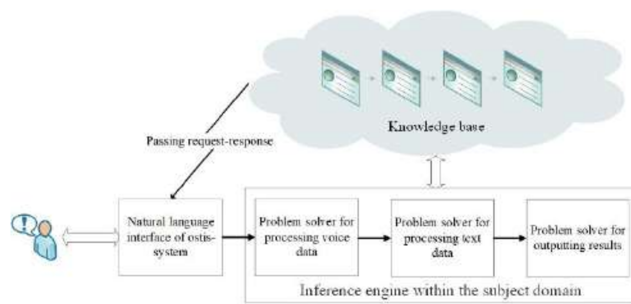


Figure 4. The model of the process of responding to user needs.

The user interface includes all the components required for user interaction and obtaining results. The general process of interacting with the user interface can be described as follows:

- The user reads all the necessary instructions.
- The user begins to record voice data according to the instructions.
- The user ends the recording action and uploads the data.
- The user presses the "Predict Data" button to receive the probability of illness corresponding to their voice data.
- The result is displayed in the result feedback area.

In the design of the user interface, taking the collection of voice data as an example, the client-side page allows participants to select from three functions: recording voice information, converting voice to text, and predicting outcomes. The voice-to-text conversion serves as an illustrative example of one method of data preprocessing. Developers may adapt or modify this functionality based on specific design requirements. After participants have recorded their voice information onto their phones, the device will save this voice data and carry out data preprocessing and feature extraction tasks. Upon activation of the "predict" function by the participant, the phone, acting as a client, will transmit the generated feature file to the server. Following the receipt of prediction results sent from the server or another agent, the phone will interpret these results and display them on the page.

User interface design is based on a component-based approach, any user interface component can be described in the ostis-system knowledge base. An illustration of the user interface displaying the probability of a participant being diagnosed with Alzheimer's disease is presented in Figure 5.

Figure 5. An example of the user interface

For the corresponding SCg-code description fragment within the OSTIS system knowledge base, see Figure 6.

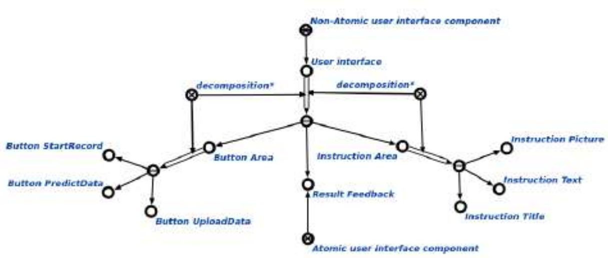


Figure 6. SCg-code of the user interface.

VI. Conclusion and future works

In the report, the authors present an approach to the design of an Internet of Things-based diagnostic network aimed at using the OSTIS platform to expand the possibilities of data interpretation within the intelligent process of diagnosing Alzheimer's disease. The structure of the ontology for the description of the subject area is given. The process of developing solvers, knowledge bases and user interfaces adapted for healthcare professionals and patients is described taking into account the stages from data collection to presentation of diagnostic results. In the future, the work will focus on improving the OSTIS-based OTNET to increase the accuracy of recognition and study its applicability to the treatment process, thereby expanding the possibilities of intelligent diagnostics in the field of medicine.

References

[1] T. Shuo, Y. Wenbo, L. G. Jehane Michael, W. Peng, H. Wei, and Y. Zhewei, "Smart healthcare: making medical care more intelligent," *Global Health Journal*, vol. 3, no. 3, pp. 62–65, 2019.

[2] J. Zhang, Y. Li, L. Cao, and Y. Zhang, "Research on the construction of smart hospitals at home and abroad," *Chinese Hospital Management*, vol. 38, pp. 64–66, 2018.

[3] K. Clemens Scott and B. Amanda, "Health information technology continues to show positive effect on medical outcomes: systematic review," *Journal of medical Internet research*, vol. 20, no. 2, 2018, article e41.

[4] B. Pradhan, S. Bhattacharyya, and K. Pal, "Iot-based applications in healthcare devices," *Journal of healthcare engineering*, vol. 2021, pp. 1–18, 2021, article ID 6632599.

[5] (2020) Alzheimer disease relevance ontology by process. [Online]. Available: https://bioportal.bioontology.org/ontologies/AD-DROP/?p=classes&conceptid=root

[6] G. Szatloczki, I. Hoffmann, V. Vincze, J. Kalman, and M. Pakaski, "Speaking in alzheimer's disease, is that an early sign? importance of changes in language abilities in alzheimer's disease," *Frontiers in aging neuroscience*, vol. 7, 2015, article 195.

[7] F. Martínez-Sánchez, J. J. G. Meilán, J. Carro, and O. Ivanova, "A prototype for the voice analysis diagnosis of alzheimer's disease," *Journal of Alzheimer's disease*, vol. 64, no. 2, pp. 473–481, 2018.

[8] R. D. Nebes, C. B. Brady, and F. J. Huff, "Automatic and attentional mechanisms of semantic priming in alzheimer's disease," *Journal of Clinical and Experimental Neuropsychology*, vol. 11, no. 2, pp. 219–230, 1989.

[9] A. Satt, R. Hoory, A. König, P. Aalten, and P. H. Robert, "Speech-based automatic and robust detection of very early dementia," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014, p. 2538—2542.

[10] U. Vishniakou and C. Yu, "Using machine learning for recognition of Alzheimer's disease based on transcription information," *Reports of BSUIR*, vol. 21, no. 6, pp. 106–112, 2023.

РАЗРАБОТКА СЕТИ ИНТЕРНЕТА ВЕЩЕЙ ДЛЯ ДИАГНОСТИКИ БОЛЕЗНИ АЛЬЦГЕЙМЕРА С ИСПОЛЬЗОВАНИЕМ OSTIS

Вишняков В.А., Чуюэ Юй

Доклад посвящен разработке Интернета вещей для диагностики болезни Альцгеймера (БА) с использованием технологии OSTIS. Приведена структура онтологии для описания элементов заболевания БА. В статье рассматривается построение ИТ-диагностической сети БА, использующей семантические возможности платформы OSTIS для обработки и анализа медицинских данных. Представлены элементы описания базы знаний, решателей и пользовательского интерфейса с использованием компонентного подхода к проектированию.

Ключевые слова: сеть интернета вещей, болезнь Альцгеймера, база знаний, решатель, пользовательский интерфейс, ИТ-диагностика, OSTIS.

Received 31.03.2024

IoT Network for Diagnosis of Parkinson's disease Using Neural Networks and OSTIS

Vishniakou Uladzimir
Belarusian State University of Informatics and Radioelectronics
Minsk, Belarus
vish@bsuir.by

Yiwei Xia
Belarusian State University of Informatics and Radioelectronics
Minsk, Belarus
xiayiwei4@gmail.com

***Abstract*—The aim of the work is to propose a model for the diagnosis of Parkinson's disease (PD) within the framework of the Internet of Things (IoT) using OSTIS. The report is devoted to the development of the Internet of Things for the diagnosis of Parkinson's disease (PD) using OSTIS technology. The structure of the ontology for describing the elements of PD disease is given. The construction of an IT diagnostic network of the PD is considered, which uses the semantic capabilities of the OSTIS platform for processing and analyzing medical data of the PD. The elements of the description of the knowledge base, solvers and user interfaces for PD using a component-based approach are presented.**

***Keywords*—Internet of Things network, IT diagnostics of Parkinson's disease (PD), PD ontology, neural networks, knowledge base, OSTIS**

I. DIAGNOSIS METHOD APPLIED TO PARKINSON'S DISEASE

The early diagnosis of Parkinson's disease has been a challenge for the medical community, and usually about 60 % of nigrostriatal neurons have degenerated and 80 % of striatal dopamine is depleted by the time the disease is diagnosed [1]. Currently, the diagnosis of the disease is based on medical history, clinical signs and symptoms, and response to antiparkinsonian drugs, but because the disease starts slowly and clinical symptoms appear only when the nigrostriatal dopamine neurons are depleted to a certain extent, patients are often at an advanced stage of the disease when they are diagnosed, missing the best time for treatment. Therefore, the development of new treatments in this field depends on two main aspects:

1) The early diagnosis of the disease.
2) The correct and constant evaluation of the effectiveness of the treatment.

In article [2] we proposed the method for complex recognition of Parkinson's disease using machine learning, based on markers of voice analysis and changes in patient movements on known data sets. The time-frequency function, (the wavelet function) and the Meyer kepstral coefficient function are used. The KNN algorithm and the algorithm of a two-layer neural network were used for training and testing on publicly available datasets on speech changes and motion retardation in Parkinson's disease. A Bayesian optimizer was also used to improve the hyperparameters of the KNN algorithm. The constructed models achieved an accuracy of 94.7 % and 96.2 % on a data set on speech changes in patients with Parkinson's disease and a data set on slowing down the movement of patients, respectively. The recognition results are close to the world level. The proposed technique is intended for use in the IoT network of IT diagnostics of PD.

II. KNOWLEDGE BASE AND ONTOLOGY

The construction of the Intelligent Parkinson's Disease Diagnosis System is based on the following three main components:

1) Knowledge Base: Collects and stores knowledge and data about Parkinson's Disease;
2) Problem Solver: Utilizes information from the knowledge base to solve specific problems;
3) Interaction Interface: Provides a way for doctors and patients to interact with the system.

A. Construction of the Knowledge Base for the Intelligent Parkinson's Disease OSTIS System

The ontology construction of the knowledge base of the Smart Parkinson's Disease Diagnostic OSTIS system is subdivided into the following areas.

Parkinson's Disease Ontology
$\subset$ *ontology*
$\subset$ *ostis-system*
$\subset$ *Familial neurodegenerative disease*
$\Rightarrow$ *Decomposition*:*:*
{• *Basic Classification of Parkinson's Disease*
• *Clinical Features of Parkinson's Disease*
• *Parkinson-Plus Syndromes [3]*
• *Etiology of Parkinson's Disease*
• *Neuropathology of Parkinson's Disease*
• *Information Models of Parkinson's Disease*
$\Rightarrow$ *Decomposition*:*:*
{• *Research Models [4]*
• *Predictive Analytic Models*
}
}

Predictive Analytic Models
⇒ *Decomposition*:*:*
{• *Medical Imaging Data [5]*
:= [Processing MRI, PET, SPECT, etc., imaging data for the diagnosis and research of Parkinson's Disease.]
• *Biomarkers [6]*
:= [Analyzing α-synuclein, inflammatory markers, oxidative stress markers, etc., to monitor disease progression.]
• *Genetic Information [7]*
:= [Analyzing Single Nucleotide Polymorphisms (SNPs) and gene mutations associated with an increased risk of Parkinson's Disease.]
• *Clinical Feature Data Processor*
⇒ *Decomposition*:*:*
{• *Motion Data Processor*
• *Voice Data Processor*
}
}

Voice Data Processor
:= [Changes in speech and language in patients with Parkinson's Disease, such as softer voice, slower speech, and reduced intonation variability, can be quantified through voice analysis technology, serving as a tool for assessing the disease and its progression.]
⇒ *Decomposition*:*:*
{• *Specific Voice Data Collection Methods*
⇒ *Decomposition*:*:*
{• *Model Selector*
⇒ *Decomposition*:*:*
{• *LSTM Model*
• *GRU Model*
• *KNN Model*
• *Random Forest Model*
}
• *Voice Data Feature Analyzer*
⇒ *Decomposition*:*:*
{• *Voice Feature Analyzer:*
• *Raw Voice Data Processor*
}
}
}

Motion Data Processor
:= [Involves patient's motor abilities and control, including tremors, muscle rigidity, bradykinesia, gait, and balance issues, analyzed through motion analysis technologies like wearable devices or motion capture systems.]
⇒ *Decomposition*:*:*
{• *Specific Motion Data Collection Methods*
⇒ *Decomposition*:*:*
{• *Model Selector*
⇒ *Decomposition*:*:*
{• *LSTM Model*
• *GRU Model*
• *KNN Model*
• *Random Forest Model*
}
• *Motion Data Feature Analyzer*
⇒ *Decomposition*:*:*
{• *db6 Wavelet Feature Analyzer:*
• *Raw Motion Data Processor*
}
}
}

B. Problem Solver for the Intelligent Parkinson's Disease Diagnosis System

The task resolver of the Intelligent Parkinson's Disease Diagnosis System is a collective of interacting sc-agents that facilitate the resolution of diagnostic and management issues related to Parkinson's disease. Herein is a fundamental decomposition of the task resolver for the Intelligent Parkinson's Disease Diagnosis System, based on the principal sc-agent classes designated for Parkinson's disease diagnostic purposes.

Intelligent Parkinson's Disease Diagnosis System Problem Solver
⇒ *Decomposition*:*
{• *Clinical Data Analysis abstract sc-agent*
⇒ *Decomposition*:*
{• *Motion Data Analysis abstract sc-agent*
• *Voice Data Analysis abstract sc-agent*
• *Genetic Information Analysis abstract sc-agent*
}
• *Symptom Severity Assessment abstract sc-agent*
• *Medical Database and Diagnostic Tool Interface abstract sc-agent*
⇒ *Decomposition*:*
{• *Medical Imaging Database Access abstract sc-agent*
• *Biomarker Analysis abstract sc-agent*
• *Time Series Analysis abstract sc-agent*
}
• *Symptom and Treatment Plan Correlation abstract sc-agent*
• *Diagnosis and Treatment Knowledge Base Verification abstract sc-agent*
⇒ *Decomposition*:*
{• *Treatment Plan Efficacy Verification abstract sc-agent*
• *Disease Diagnosis Accuracy Verification abstract sc-agent*
• *Patient Data Privacy and Security Protection abstract sc-agent*
}
• *Intelligent Question-Answering abstract sc-agent*
⇒ *Decomposition*:*
{• *User Interaction abstract sc-agent*
• *Medical Knowledge Graph*
• *Answer Generation abstract sc-agent*
}
}

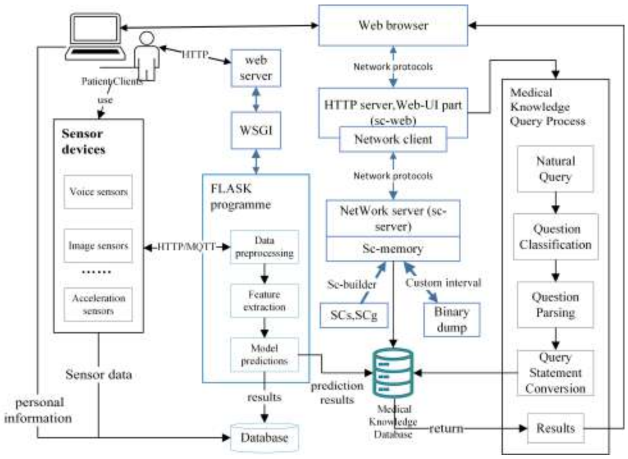


Figure 1. Framework Diagram for the Intelligent Parkinson's Disease OSTIS System's Automated Diagnostic and Question-Answering Tool.

III. SYSTEM ARCHITECTURE OVERVIEW

This section offers a comprehensive overview of the architectural underpinnings governing the deployment of our Parkinson's disease diagnosis model [2] within the IoT framework. It expounds upon the intricate interplay between system components designed to facilitate efficient data processing, storage, and presentation. Fig 1 illustrates an IoT system architecture enhanced with neural network capabilities using OSTIS technology.

IoT Device (Client): Data collection and preprocessing occur on the client-side. This could be a smart device such as a sensor or a smartphone, responsible for data acquisition and preliminary data preprocessing and feature extraction.

Local Flask Server [8]: The Flask server is located locally and serves as the receiver for data from the IoT device. It acts as an intermediary for data transmission, forwarding the received feature data to the OSTIS server.

OSTIS [9] Server: The OSTIS server is a knowledge graph platform that receives and processes data from the local Flask server. Running on this server is a Neural Network Predictor Agent responsible for loading internal neural network model files.

Neural Network Predictor Agent: This agent is responsible for loading and executing neural network models, processing incoming feature data, and making predictions. Predictions can be associated with knowledge within the OSTIS system and ultimately saved to a local database.

Local Database: The local database is used to store the prediction results returned from the OSTIS server, along with other relevant information.

The workflow of the entire system is as follows:

1) IoT device collects and preprocesses data.
2) Preprocessed data is sent to the local Flask server.
3) The local Flask server forwards the data to the OSTIS server.
4) The Neural Network Predictor Agent on the OSTIS server loads the neural network model and processes the data.
5) Determining the subclass, type, and subtype based on the Parkinson's disease diagnostic classifier, i. e., the types of diagnostic entities in medical ontology;
6) Establishing the inherent attributes and characteristics of the diagnostic category;
7) Determining the values of features for that diagnostic category;
8) Resolving polysemy in the diagnostic process;
9) Establishing the corresponding connections between diagnostic entities and concepts with medical semantic features in the knowledge base;
10) Establishing relationships between diagnostic entities belonging to a specific category of Parkinson's disease.
11) The processed results are associated with the knowledge graph and finally saved to the local database.

This system allows real-time data processing and complex object recognition in an IoT environment, combining data with a knowledge graph to support advanced analysis and decision-making. Proper configuration and management of each component are required to ensure the efficient operation of the system.

IV. IOT DEVICE AND DATA COLLECTION

Within this subsection, we embark on an in-depth exploration of the IoT device deployed for data collection. We elucidate its pivotal role in the acquisition of two critical data modalities: movement data and audio data. We delve into the intricacies of data preprocessing and transmission to the local server, underscoring the cardinal importance of real-time data acquisition capabilities. As shown in Figure 2, the IoT system architecture based on neural network with OSTIS technology is illustrated.

First, let us introduce the IoT device used. The device plays a key role in data acquisition by capturing two important data types: motion data and audio data. Motion data records the movement characteristics of Parkinson's patients, while audio data captures their sound characteristics.

To capture the acceleration data of the cell phone:

1) use a third-party library in Python (PySensors) to access the phone's acceleration sensor.
2) set up the sensor parameters, setting the sampling rate to 64hz and the sensor precision to 16 bits.
3) create a data storage structure to hold the acquired acceleration data.
4) In a loop, periodically read the acceleration sensor data and store it in the data structure.
5) Store the data in a local file for further processing and analysis.

Collect the voice data from the cell phone:

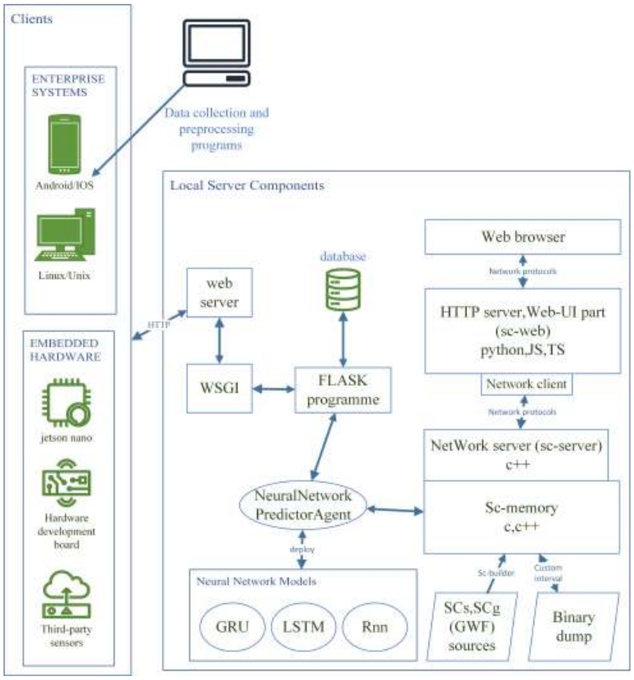


Figure 2. IoT system architecture based on neural network with OSTIS technology.

1) Use a third party library (PyAudio) in Python to access the microphone of the phone.
2) Initialize the microphone and set the audio parameters . The sample rate is 44100hz, the number of channels is 1 and the bit depth is 16.
3) Create an audio stream for receiving and recording voice data.
4) Start recording voice data and store it on a local file.

The next step after collecting and recording accelerometer and voice data from Parkinson's disease patients is to preprocess and extract features from these data. This process is aimed at preparing the data for further analysis and model training.

Data Preprocessing and Feature Extraction Process:

1) Data Cleaning and Calibration: Firstly, the collected data undergoes initial cleaning and calibration to remove potential noise and outliers. For accelerometer data, sensor calibration is performed to ensure data consistency within a common reference frame.
2) Time and Frequency Domain Analysis: Time and frequency domain analyses are conducted on voice data to obtain fundamental signal characteristics. Time domain analysis includes waveform shape, energy, duration, and more. Frequency domain analysis encompasses spectral distribution, frequency components, and related features.
3) Feature Extraction: Features are extracted from accelerometer and voice data, representing informative and meaningful attributes that aid in subsequent model training and analysis. For accelerometer data, statistical features of motion patterns such as mean, standard deviation, energy, etc., can be extracted. For voice data, sound features like fundamental frequency, pitch, spectral features, etc., can be extracted.
4) Feature Normalization: Extracted features are normalized to ensure they share similar scales, preventing certain features from disproportionately influencing model training.
5) Feature Selection: With a large number of extracted features, feature selection can be performed to choose the most relevant and useful ones, reducing dimensionality and enhancing model efficiency and performance.
6) Data Storage: The feature data is saved as a Comma-Separated Values(CSV) file for local storage.

The specific process of transferring data from a mobile device to the Flask server on a computer is as follows:

1) The mobile device collects accelerometer data or voice data and performs preprocessing and feature extraction.
2) The feature data is packaged in JSON format. Below is a JSON example:

```
{
  "data_type": "acceleration",
  // or "voice"
  "features":
    [
      { "feature_name":
      "feature_1",
        "value": 0.123},
      { "feature_name":
      "feature_2",
        "value": 0.456}
      // More features...
    ]
  }
```

3) Send the JSON data to the Flask server using the HTTP protocol with a POST request. Target URL: http://192.168.100.14:5000/getdata. The request header includes data type information and can be set as 'Content-Type: application/json'.
4) The Flask server receives and handles the request at the 'getdata' route.
5) Upon reception of the data on the server side, a range of distinct processing strategies can be applied contingent on the data type. These strategies encompass data storage within local files, conducting more in-depth analytical procedures, or the utilization of predictive models for decision-making.

In this process, data is transferred from the mobile device to the Flask server, which receives the data via the specified route and performs the necessary processing.

V. LOCAL SERVER SETUP WITH OSTIS SYSTEM

The OSTIS Web Platform [10] is a web-oriented software platform of the OSTIS Project. It serves as a robust framework for deploying existing OSTIS systems and creating new ones.

The OSTIS Web Platform includes the following components:

1) Knowledge Base [11]: It contains top-level ontologies to assist in developing various information models.
2) Knowledge Processing Machine [12]: This component features semantic network storage and agent-based knowledge processing.
3) Web-Oriented Semantic Interface [13]: It allows users to interact with the intelligent system.

Below are the steps for a quick start using Docker Compose, suitable for windows:

1) Clone the repository and navigate to the directory.
2) Download images from Docker Hub:

```
docker compose pull
```

3) Build the knowledge base:

```
docker compose run machine
build
```

4) Launch the web platform:

```
docker compose up
```

Using Docker Compose:

1) Build the Knowledge Base (required before the first startup or if you've made updates to KB sources):

```
docker compose run machine
build
```

2) Start platform services and access the web interface locally (address: localhost:8000):

```
docker compose up
```

3) Launch the knowledge processing machine:

```
./scripts/run_sc_server.sh
```

4) In another terminal, launch the semantic web interface (address: localhost:8000):

```
./scripts/run_sc_web.sh
```

When the Flask server receives data, it forwards it to the agent within the OSTIS server, namely the neural network predictor agent. The primary role of this agent is to load the appropriate neural network model and process the received data. During the processing, the model makes predictions and generates results.

Subsequently, the model stores these results in a database for further analysis and persistent storage. Once the results are successfully stored, the Flask server constructs a response, encapsulating the prediction results in JSON format, and sends it back to the client for users to access real-time analysis results.

This process ensures data integrity and persistence while allowing users to interact with the OSTIS server through the Flask server to access the information they need.

VI. DISPLAYING DATA IN THE OSTIS SYSTEM

Once the prediction results are stored in the OSTIS system's database, they can be accessed and displayed within the system's knowledge base. This allows users to interact with and visualize the results for further analysis or decision-making. The process of displaying data in the OSTIS system involves the following steps:

1) Knowledge Base Integration: The prediction results, which are stored in the database, need to be integrated into the OSTIS knowledge base. This integration typically involves creating or updating knowledge structures within the system to represent the newly acquired data.
2) Semantic Representation: The data should be represented in a semantically meaningful way using the OSTIS Semantic Computer Code (SC-code) [14] to ensure compatibility with the system's knowledge processing capabilities. This may involve defining new semantic relationships or entities to represent the data.
3) User Interface: OSTIS provides a web-oriented semantic interface that allows users to interact with the intelligent system. This interface can be customized to display the prediction results in a user-friendly and informative manner. Users can query the system to access specific data and visualize it through the interface. Figure 3 displays SCn-nodes featured within the OSTIS web interface's knowledge base.
4) Visualization and Analysis: Depending on the nature of the prediction results, the OSTIS system may offer various visualization tools and analysis capabilities. Users can explore the data, generate reports, or perform further analysis within the system. Figure 4 presents SCg-nodes displayed within the knowledge base of the OSTIS web interface.

VII. Conclusion

The report presents the architecture of the IoT network for the diagnosis of Parkinson's by voice and movement of patients. The IoT device collects and preprocesses patient data using a smartphone to collect data and extract the necessary functions from them. This recognition data is transmitted through the local Flask server, which acts as the main intermediary. The heart of the system, the OSTIS server, serves as a knowledge graph platform hosting a neural network prediction agent that downloads, executes and links predictions with existing knowledge. The advantage of this system lies in its ability to facilitate real-time data processing and recognition of complex objects in the IoT network. Combining data and

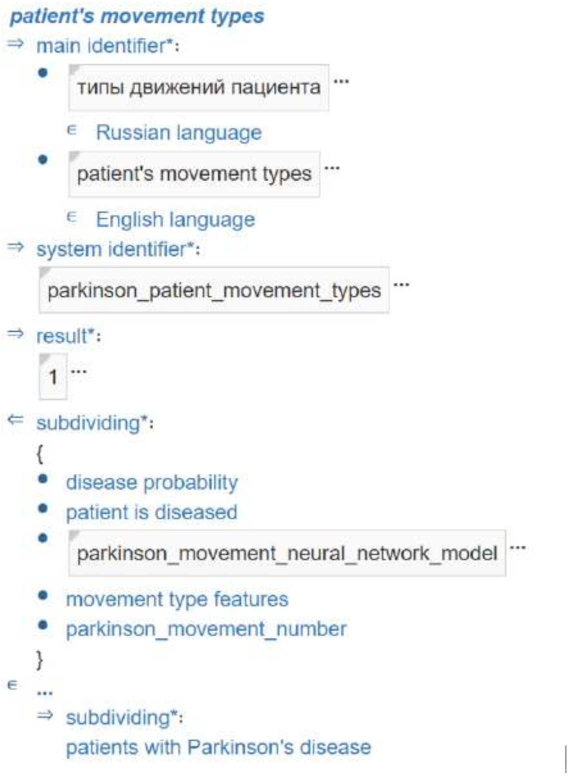


Figure 3. SCn-nodes showcased in the OSTIS web interface's knowledge base.

knowledge base allows for in-depth analysis and informed decisions in the field of diagnosis of Parkinson's disease. This work highlights the potential of the IoT network and deep learning technologies for rapid remote IT diagnostics. An innovative combination of data processing, neural networks and knowledge base integration promises a more accurate and timely diagnosis of Parkinson's disease, which will ultimately improve the further treatment of patients.

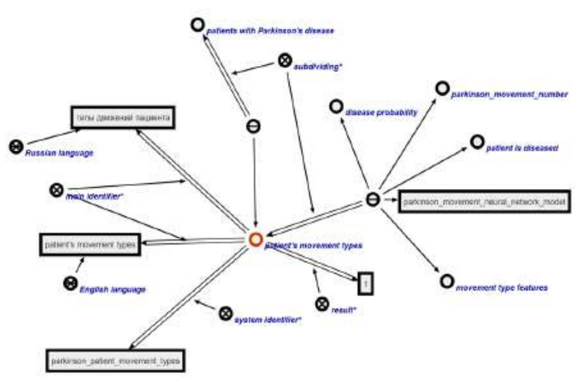


Figure 4. SCg-nodes showcased in the OSTIS web interface's knowledge base.

References

[1] W. Dauer and S. Przedborski, “Parkinson's disease: mechanisms and models,” *Neuron*, vol. 39, no. 6, pp. 889–909, 2003.

[2] U. Vishnyakou and X. Yiwei, “Recognition of signs of parkinson's disease based on voice markers and motor activity,” *Informatics*, no. 3, pp. 106–114, 2023.

[3] A. C. Vertes, M. R. Beato, J. Sonne, and M. Z. K. Suheb, “Parkinson-plus syndrome,” in *StatPearls [Internet]*. StatPearls Publishing, 2023.

[4] S. J. Chia, E.-K. Tan, and Y.-X. Chao, “Historical perspective: models of parkinson's disease,” *International journal of molecular sciences*, vol. 21, no. 7, p. 2464, 2020.

[5] J. Zhang, “Mining imaging and clinical data with machine learning approaches for the diagnosis and early detection of parkinson's disease,” *NPJ Parkinson's disease*, vol. 8, no. 1, p. 13, 2022.

[6] U. Ganguly, S. Singh, S. Pal, S. Prasad, B. K. Agrawal, R. V. Saini, and S. Chakrabarti, “Alpha-synuclein as a biomarker of parkinson's disease: good, but not good enough,” *Frontiers in aging neuroscience*, vol. 13, p. 702639, 2021.

[7] H. Pan, Z. Liu, J. Ma, Y. Li, Y. Zhao, X. Zhou, Y. Xiang, Y. Wang, X. Zhou, R. He *et al.*, “Genome-wide association study using whole-genome sequencing identifies risk loci for parkinson's disease in chinese population,” *npj Parkinson's Disease*, vol. 9, no. 1, p. 22, 2023.

[8] V. R. Vyshnavi and A. Malik, “Efficient way of web development using python and flask,” *Int. J. Recent Res. Asp*, vol. 6, no. 2, pp. 16–19, 2019.

[9] N. Zotov, “Implementation of information retrieval subsystem in the software platform of ostis-systems,” *Open Semantic Technologies for Intelligent Systems (OSTIS)*, vol. 7, pp. 77–94, 2023.

[10] M. Sadouski, “User interface of the ostis ecosystem,” *Open Semantic Technologies for Intelligent Systems (OSTIS)*, vol. 7, pp. 153–158, 2023.

[11] I. Davydenko, “Ontology-based knowledge base design,” *Open Semantic Technologies for Intelligent Systems (OSTIS)*, pp. 57–72, 2017.

[12] D. Shunkevich, “Ontology-based design of knowledge processing machines,” *Open Semantic Technologies for Intelligent Systems (OSTIS)*, pp. 73–94, 2017.

[13] M. Sadouski, “Ontological approach to the building of semantic models of user interfaces,” *Open Semantic Technologies for Intelligent Systems (OSTIS)*, vol. 5, pp. 105–116, 2021.

[14] K. Bantsevich, “Metasystem of the ostis technology and the standard of the ostis technology,” *Open Semantic Technologies for Intelligent Systems (OSTIS)*, vol. 6, pp. 357–368, 2022.

СЕТЬ IOT ДЛЯ ДИАГНОСТИКИ БОЛЕЗНИ ПАРКИНСОНА С ИСПОЛЬЗОВАНИЕМ НЕЙРОННЫХ СЕТЕЙ И OSTIS

Вишняков В.А., Ивей С.

Цель работы состоит в том, чтобы предложить модель диагностики болезни Паркинсона (БП) в рамках Интернета вещей (IoT), используя ОСТИС. Доклад посвящен разработке сети Интернет вещей для ИТ-диагностики болезни Паркинсона (БП) с использованием технологии OSTIS. Приведена структура онтологии для описания элементов заболевания БП. Рассматривается построение ИТ-диагностической сети БП, которая использует семантические возможности платформы OSTIS для обработки и анализа медицинских данных БП. Приведены элементы описания базы знаний, решателей и пользовательских интерфейсов для БП с использованием подхода, основанного на компонентном прое.

Ключевые слова: сеть Интернет вещей, ИТ-диагностика болезни Паркинсона (БП), онтология БП, нейронные сети, база знаний, ОСТИС.

Received 31.03.2024

Neural Network Technology for Real-Time IT Service Management

Viktor Krasnoproshin
Faculty of Applied Mathematics and Computer Science
Belarussian State University
Minsk, Belarus
Email: krasnoproshin@bsu.by

Aleksandr Starovoitov
Faculty of Applied Mathematics and Computer Science
Belarussian State University
Minsk, Belarus
Email: StarovoytovAA@bsu.by

***Abstract*—The paper explores a relevant applied problem related to building decision-making systems for resource management of critical IT services. The uncertainty of external load is an important factor that affects operational management. Neural network forecasting is used to improve control systems. A model system of a critical IT service is described, an original technology, structure and architecture of the control system are proposed. Experiments have been conducted to confirm the workability of the proposed technology.**

***Keywords*—decision-making, information system, proactive management, uncertainty of external load, neural networks, critical IT service**

I. Introduction

Support for computing systems in critical infrastructures (such as banking, telecommunications, industrial systems) in an operational state (with guaranteed computational resource levels) is a relevant problem in today's digital society.

Uncertainty in external loads and outages of computing equipment lead to operational failures and performance degradations of critical IT systems. As a result, the loss of operational efficiency in processing information and conducting banking and other operations can have serious consequences, including financial losses and major incidents.

Making operational decisions for the management of critical IT services allows for the reduction or prevention of negative consequences. However, the human factor often contributes to a decrease in operational efficiency. Therefore, various automated solutions are actively being developed to enhance efficiency and proactivity in the management of critical systems.

In laboratory conditions, it is difficult to develop relevant systems (without interacting with the critical infrastructure itself). Therefore, one of the current problems is the creation of model systems that enable researchers to use them for the development of proactive management systems for critical IT services.

Several authors develop various management systems for critical IT services using neural network models, which contribute to efficient decision-making [1]–[4]. However, in these systems, only one neural network model is trained for specific types of external load. It is assumed that these models will successfully forecast the values of necessary parameters for other types of load associated with uncertainty. In these works, training datasets are prepared in advance, containing long time series with a large number of elements. Training on such datasets takes a considerable amount of time and requires high-performance resources (GPU, TPU) to effectively train models within an acceptable timeframe.

This paper considers the construction of a model system conceptually corresponding to a critical IT service and an operational management system with a neural module.

A multi-model approach is used, based on the idea that the managed IT system can be in various states (in terms of computational resource volume), and for each state during its existence, a neural network model can be created and trained to predict the average %CPU utilization across computational modules.

A combined management system is used, in which the control decision for state changes is formed based on both reactive and proactive approaches.

II. Critical IT Service: Conceptual Model

Let's consider the conceptual model of a critical information system, which describes its basic elements and significant parameters. It can be described in the form of the following tuple:

$$\text{System} = (\text{Clients} \cup \text{Balancers} \cup \text{APPs} \cup \text{DBs}, \text{Links}), \quad (1)$$

where

$\text{Clients} = \{\text{Client}_i\}$ — the set of system clients. These can be both user devices and other external systems;

$\text{Balancers} = \{\text{Balancers}_j\}$ — the set of balancers that distribute requests from clients and external systems across services. Services within the system can also communicate with each other through balancers;

$\mathrm{APPs} = \{\mathrm{APP}_k\}$ – the set of application services within the system;

$\mathrm{DBs} = \{\mathrm{DB}_l\}$ — the databases of the system (can be of any type);

$\mathrm{Links} = \{\mathrm{Link}_m\}$ — the set of bidirectional temporal links that arise during communication between elements of the system;

The elements Client_i, $\mathrm{Balancers}_j$, APP_k, DB_l, Link_m can also be represented as tuples:

$\mathrm{Client}_i = (\mathrm{Protocol}_i, \mathrm{Profile}_i)$, where $\mathrm{Protocol}_i$ defines the specification, and $\mathrm{Profile}_i$ represents the interaction profile (requests, their parameters, frequency, etc.) for the $\mathrm{i^{th}}$ client;

$\mathrm{Balancers}_j = (\mathrm{Protocol}_j, \mathrm{RPS}_j, \mathrm{Throughput}_j)$, where RPS_j is the number of requests per second, and $\mathrm{Throughput}_j$ is the throughput (Mbps) for the $\mathrm{j^{th}}$ balancer. It is assumed that the balancer supports all interaction protocols, and the delay on the balancer is insignificant compared to the response time;

$\mathrm{APP}_k = (\mathrm{Compute\_Modules}_k, \mathrm{Soft\_Service}_k)$ the set of instances of application software, where $\mathrm{Compute\_Modules}_k$ is the set of computational modules, and $\mathrm{Soft\_Service}_k$ is the type of application service used on the computational modules $\mathrm{Compute\_Modules}_k$;

$\mathrm{DB}_l = (\mathrm{Compute\_Modules}_l, \mathrm{Soft\_Service\_DB}_l)$ — describes a set of database instances serving one type of application service instances, where $\mathrm{Compute\_Modules}_l$ is the set of computational modules, and $\mathrm{Soft\_Service\_DB}_l$ is the type of database software used in the computational modules $\mathrm{Compute\_Modules}_l$. These can be any (not necessarily relational) databases;

$\mathrm{Link}_m = \{(\mathrm{P}_{mn}, \mathrm{P}_{mo})\}$ — the set of bidirectional temporal links between elements of the system, established through communication via specific open ports;

$\mathrm{Compute\_Modules} = \{\mathrm{Compute\_Module}_c\}$ — the set of computational modules of the system, where each computational module can also be described by the following tuple:

$$\mathrm{Compute\_Module} = (\mathrm{CM\_Type}, \mathrm{CM\_CPU\_Limit}, \mathrm{CM\_CPU\_Perf}, \mathrm{CM\_RAM\_Limit}, \mathrm{CM\_RAM\_Perf}, \mathrm{CM\_Storages}, \mathrm{CM\_NET\_Throughput}),$$

where

- $\mathrm{CM\_Type} \in \{\text{"physical server"}, \text{"logical or hw partition"}, \text{"virtual machine"}, \text{"container"}\}$ — the type of module;
- CM_CPU_Limit — the number of processors (CPU or vCPU) available to the module;
- CM_CPU_Perf — the maximum performance of one processor in the module;
- CM_RAM_Limit — the available volume of RAM in the module (GiB);
- CM_RAM_Perf — the maximum performance of RAM in the module (determined by bandwidth and memory access time);
- CM_NET_Throughput — the maximum throughput capacity of the module (Gbps).

$\mathrm{CM\_Storages} = \{\mathrm{CM\_Storage}_p\}$ — storage modules available to the computational module. Each storage module is defined by a tuple:

$$\mathrm{CM\_Storage} = (\mathrm{CM\_Storage\_Type}, \mathrm{CM\_Storage\_Capacity}, \mathrm{CM\_Storage\_Perf}),$$ where

- $\mathrm{CM\_Storage\_Type} \in \{\text{"local"}, \text{"external"}\}$ — determines the type of connection of the storage module to the computational module. In this case, the local option describes local disk resources (HDD, SSD). The "external" component refers to storage resources external to the computational module. These can be connected using various input-output devices that support different block protocols (iSCSI - SCSI over Ethernet, Fiber Channel Protocol – SCSI over Fiber Channel, NVMe-oF – NVMe over Fiber Channel), file protocols (CIFS, NFS) and object protocol (S3).
- CM_Storage_Capacity — the storage capacity of the module (GiB).
- CM_Storage_Perf — the performance of the module can be defined as the number of Input/Output Operations Per Second (IOPS) (with minimum response time) or throughput (GBps), for different workload profiles.

The workload profile is determined by the percentage of read operations (%Read), the size of data blocks (KiB), and the distribution, which indicates the presence of block sizes in requests, their proportion, and the delays associated with them in the total stream of requests.

$\mathrm{Soft\_Service} = \{\mathrm{Soft\_Service}_k\}$ — types of application services in the system used by computational modules in the system.

$\mathrm{Soft\_Service\_DB} = \{\mathrm{Soft\_Service\_DB}_l\}$ — types of database software used in computational modules.

In information systems, for integrating various services, elements such as message brokers (e. g., RabbitMQ, IBM WebSphere MQ, ActiveMQ Artemis, Apache Kafka, etc.) are often used. In our case, these services are not allocated to a separate class since essentially they can be attributed either to the set of entities in APPs or to DBs if this functionality is implemented at the database level (e. g., Oracle Advanced Queuing).

The model structure explicitly does not include: telecommunication equipment (Switches, Routers), resource management systems, and various perimeter control systems of the information system. All of these may restrict access to the system from outside and between components based on ports and protocols (Firewalls). Additionally, they may lead to deeper inspection of the exchange at the application level (WAF — Web Application Firewalls) and the necessity to perform analysis of exchanges at OSI Model layers 3 and 4 (IPS — Intrusion Prevention System), tracking anomalies and conducting checks for known vulnerabilities and attack vectors based on signature databases and established policies.

The structure also explicitly does not include monitoring and logging systems, antivirus protection, encryption, as well as other technological services and systems (LDAP, DNS, backup and recovery services, CI/CD systems, time synchronization service, etc.). It is assumed that these services are properly configured, and their operation does not affect the functionality of the system being considered.

III. Model System: Task Statement

With consideration of the described model, the following task is formulated:

To develop the system that conceptually corresponds to the previously described model of a critical information system (1), meeting the following criteria:

- The system can operate on a workstation, laptop, or virtual machine with resources not exceeding 4 CPU cores and 8 GiB of RAM;
- The set of APPs is represented by a compact web application capable of handling external requests;
- An instance of the web application service operates within the computational module Compute_Module, which has a specific limit on CPU resources (other limits are also possible but not mandatory);
- The system has the capability to scale within the specified resources mentioned above, meaning the number of instances of the web application service can be adjusted during operation under load by increasing or decreasing the number of computational modules. Scaling management is available both programmatically and manually;
- For each computational module of the system CPU utilization metrics are collected, with a metric collection period $\sim 2s$. The data is stored in CSV files with the specified frequency. When additional computational modules are added, statistics collection is automatically enabled for them;
- The system should have a reactive scaling module that operates according to the following algorithm: after receiving metrics of %CPU utilization for the running computational modules hosting an application web service, the average utilization percentage $\overline{\%\text{CPU}}_N$ is calculated based on the number (N) of running modules. If $\overline{\%\text{CPU}}_N > 50\%$, the system automatically adds another computational module with the application web service. If $\overline{\%\text{CPU}}_N < 20\%$, the system automatically shuts down one computational module with the application web service. If $\overline{\%\text{CPU}}_N$ is in the range from 20% to 50%, the system does not perform any configuration changes;
- After each change in the system's state (during scaling), a stabilization mode must be activated, during which the reactive control system does not perform any configuration changes to the application web service for a specified period of time;
- The set of Balancers is represented by a request balancing service between instances of the web application;
- When a new instance of the web application is added, it is automatically included in the load balancing. Similarly, when an instance of the web application is turned off, it is automatically excluded from the load balancing;
- The set of Clients is represented by a load testing system where you can define a workload profile for the application system based on requests and various user profiles. The workload profile can be defined using a function, pre-prepared data, or through the operation of web and CLI clients. The load testing system should display real-time statistics during testing and have the capability to save reports containing statistics on specific requests such as RPS, Response Time, errors during request execution, and the number of users;
- The load testing system operates within the computational module and can support a distributed configuration of instances running simultaneously across multiple computational modules;
- During a load test with maximum load, the CPU resource consumption by the computational module, where the load testing system operates, should not exceed the capacity of one CPU core;
- The presence of a database instance is possible but not mandatory;
- The system supports the Infrastructure as Code (IaC) model.

IV. Model System: Implementation

The algorithm used to solve the given task includes the following main stages:

- Defining the key functional blocks of the system. Preparation of a high-level schematic diagram;
- Identifying possible implementation options for each block considering the system requirements;
- Analyzing possible implementations considering the following criteria: availability of ready-made components that can be used to build functional blocks, simplicity of implementation, and implementation time;
- Choosing an implementation option for prototyping the system;
- Creating a prototype of the system;
- Qualitative assessment of the prototype's compliance with the task criteria during testing;
- If the criteria are not met, return to step 2;
- Perform the necessary number of iterations (steps 2-6) until the prototype meets the task criteria.

Next, the functional blocks of the system were defined and a high-level schematic of the prototype was constructed (Fig. 1):

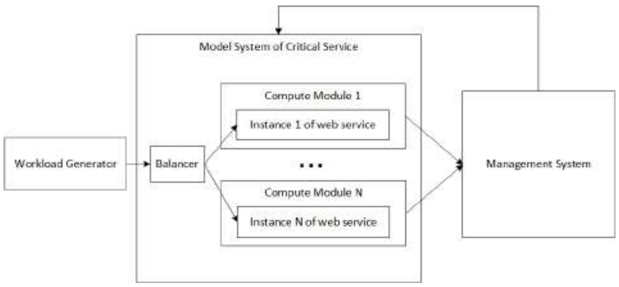


Figure 1. Structural block diagram of Model Critical Service.

A. Computational Modules

This functionality is core as it forms the basis for various solutions. In works [1]–[4], solutions for proactive management tasks were based on virtual machines deployed in various public clouds (such as Amazon, Google, etc.) or Private Clouds or Data Centers. For managing virtual machines (creation, launch, stop, deletion), cloud service capabilities or data center management systems were used. Virtual machines provide good application isolation but require more computational resources since each virtual machine needs resources for the operating system. Additionally, it is necessary to use a module that replicates the functionality of a cloud service for managing virtual machines (creation, launch, stop, deletion), as well as perform configuration of operating systems in virtual machines, installation, configuration, and management of the application web service, and support the Infrastructure as Code (IaC) model.

As a result of the research, the decision was made to use containers as the computational modules. They are less resource-intensive, simpler to implement, and meet the criteria III.

Several solutions were considered: Docker Compose [5], Kubernetes Cluster [6]. The Docker Compose option turned out to be simpler, although the use of the Kubernetes Cluster with a single worker node (e. g., Minikube [7]) is also possible.

As a result, the decision was made to use Docker Compose for the computational module block in the system prototype. This solution allows for the creation of computational modules (Docker containers), their management, and the collection of container utilization metrics. There are ready-made libraries (e. g., Docker SDK for Python [8]) for working with the Docker API Engine.

Load balancing across containers with web application is achieved using Docker Compose's built-in features based on the service name. The configuration of services and resources is described in a YAML file.

B. Web Application Service

To conserve resources, ensure stability, and simplify setup and operation, it was decided to implement the web application as a microservice based on the popular high-performance minimalist web framework echo.labstack [9], written in the high-level language Go [10].

C. Load Generator

Two options of load generation software were considered: Apache JMeter [11], written in Java, and Locust [12], written in Python. Apache JMeter is more feature-rich, complex to configure, and resource-intensive, whereas Locust has less functionality, is easier to set up, allows load profiles to be described as Python classes, is less resource-intensive, supports distributed instance configuration, and allows defining user load profiles as functions or pre-prepared data. Locust was chosen as the load generator for the prototype.

During debugging, it was discovered that in the case of a large number of lightweight requests, the CPU utilization of the container running Locust noticeably exceeds the CPU utilization of containers hosting the web application services. It was necessary to add an additional endpoint (/load) at the web application level, which invokes Go code. This code, using parallel goroutines, achieves the desired increase in CPU utilization.

The addition of this more CPU-intensive request allowed the rebalancing of the resource utilization between the load generator and the system under test. Experimentally, it was determined that the prototype system, with the resource limit specified in criterion III and the generated load profile, can handle up to 700-800 active users. Further increasing the load leads to reaching the limit of resource utilization on the virtual machine hosting Docker Compose.

Next, we will discuss the principles of operation and implementation of a combined control system with a neural module.

V. Combined Control System: Operation Principle

To manage the resources of a critical IT system, an agent is used, which in real-time receives utilization data from computational modules and makes decisions regarding the scaling of the managed system. The agent employs a combination of reactive and proactive management. For each state of the managed system, a unique initial dataset is automatically generated, a neural network model is trained based on this dataset, and predictions of resource utilization parameters are made.

The agent compares the current data of average load across computational modules with the forecast results for a specific system state and makes decisions regarding state changes (scaling).

If there is a sharp peak in load and the neural network model is not yet ready or there is no forecast for the

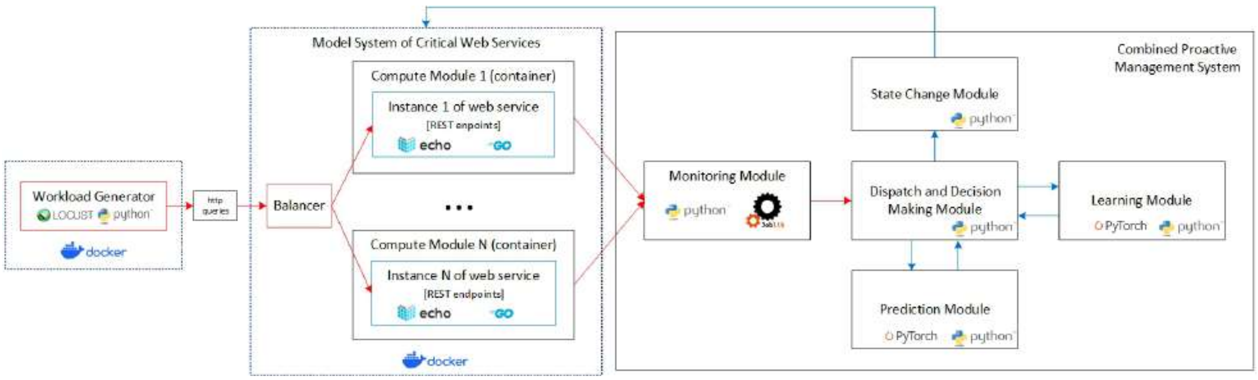


Figure 2. Structural block diagram of the combined control system.

average utilization of modules, or if such behavior is not included in the forecast, the reactive component is triggered.

If the forecast of average parameters exceeds threshold values, a proactive decision is made in advance regarding the state change of the system.

VI. Combined Control System: Implementation

The system architecturally consists of 5 main modules, which are depicted in (Fig. 2).

A. Monitoring Module

The monitoring module is responsible for collecting performance metrics from the computational modules of the system and adding/removing new metrics (when the system state changes). It is worth noting separately that when working with the Docker API Engine, there is a peculiarity related to the fact that the API does not return statistics for all containers in a single request, as the docker stats console command does. Additionally, the request itself takes $\sim 1s$ to execute because the Docker daemon needs a certain interval to calculate the corresponding average metric values. Therefore, the joblib library [13] was used for correct metric collection, which allows for implementing parallel requests.

B. State Change Module

The state change module is responsible for sending control commands (which modify the state of the managed system) and ensuring the correctness of state changes.

C. Dispatch and Decision-Making Module

The dispatch and decision-making module is central. It and the other modules are implemented in the high-level language Python. More detailed algorithm of its operation is presented in Figure 3 (see [16] for more details).

During initialization, resource utilization thresholds (system SLAs) are set, reaching which leads to a change in the state of the managed system. Separate thresholds are set for adding (A) and removing (D) computational resources. A system stabilization parameter (cool_period) is set, determining the number of cycles during which no state changes are performed in the system. A parameter is set to determine the number of data accumulation iterations for one state to create a model (M). A lead time parameter (Z) is set – the number of forecast points into the future.

In the main process, after the initialization stage, a loop is implemented in which each iteration involves refining the composition of computational modules of the managed system, obtaining current values of utilization of computational modules, and making decisions regarding the change of the managed system's state.

Metrics are collected with a period $\sim 2s$. Historical data on the utilization of each computational module for different system states are saved as CSV files. Data on module utilization for the current state are accumulated in memory in a dictionary. Using the latest collected data, the main process calculates the average value across the set of computational modules (R).

If R exceeds the threshold for addition (A), a decision is made to add resources to the managed system. If R is less than the removal threshold (D), a decision is made to remove resources. After the decision is made, a command is sent to the state change module (reactive component).

In this process, mechanisms of non-blocking interaction with other processes are implemented, which run in parallel with the main process and responsible for creating and training neural networks and forecasting utilization parameters for a specific state with a given lead time. This mechanism is implemented using the multiprocessing [14] library. Inter-process communication uses the multiprocessing.Queue mechanism, which forms FIFO (first input first output) queues.

Three queues (model_list_q, model_state_q, model_result_q) are used for interaction between the main process and the process responsible for creating and training the neural network model, and three queues

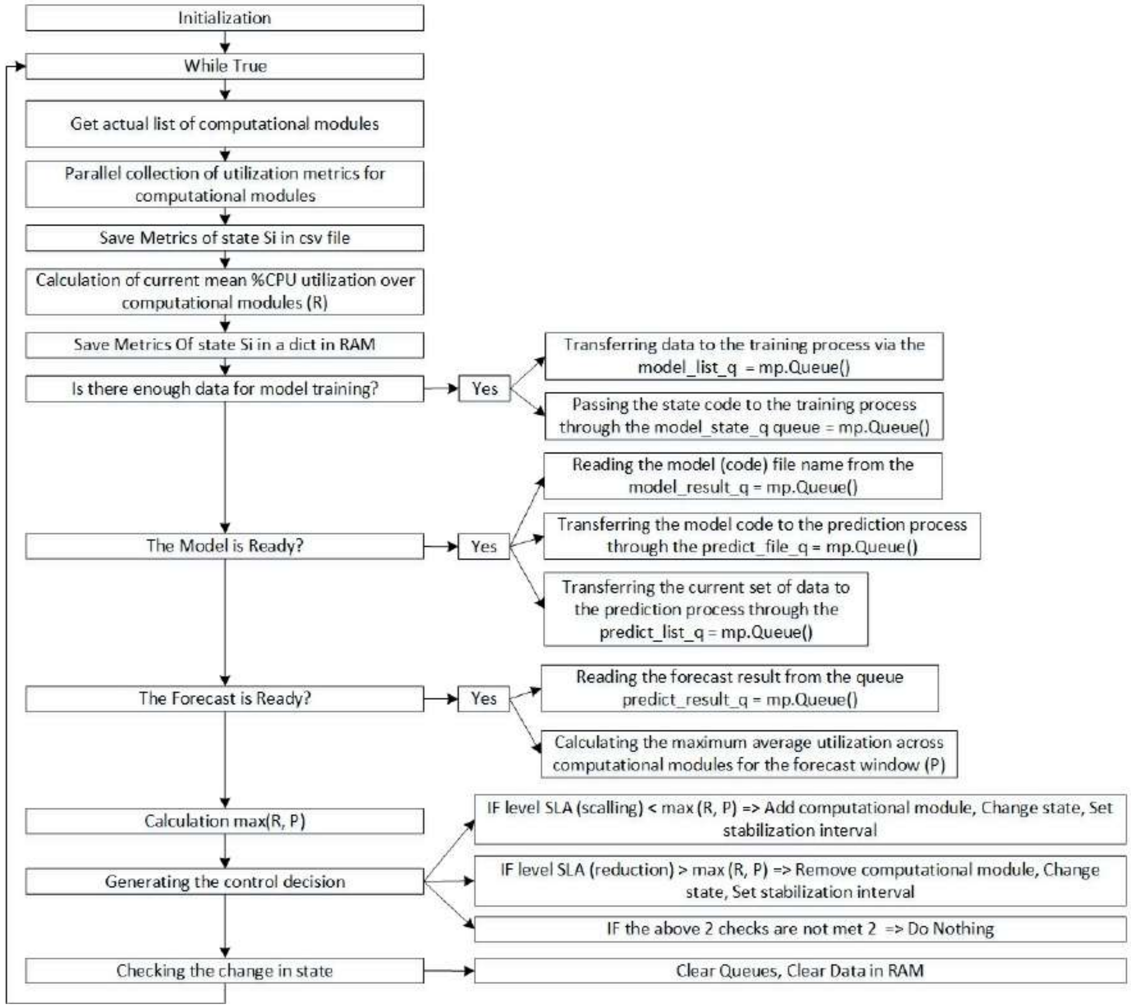


Figure 3. Control system operation algorithm.

(predict_file_q, predict_list_q, predict_result_q) are used for interaction with the prediction process (see Fig. 4).

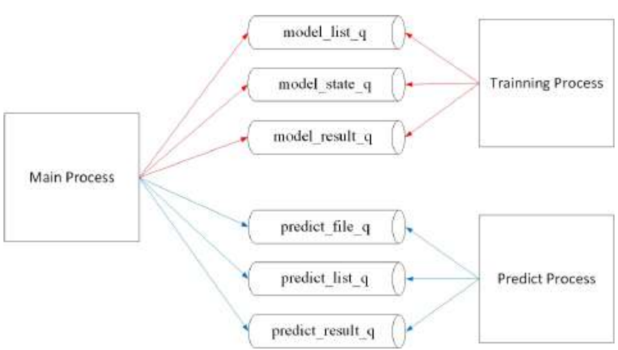


Figure 4. Multiprocessing queues.

D. Training and Prediction Modules

Both modules are implemented in Python, the PyTorch [15] library is used. The training module creates neural network models based on datasets for various states of the managed system and saves the models in the neural network model library. Each model is encoded with state index.

The prediction module loads models by state index from a library of neural network models and performs utilization forecasts with a certain advance for different states of the controlled system.

VII. Parameters of Neural Networks

We assume that each subsequent element in the time series depends on a certain number of previous elements of the series, i. e., lagged values of the original series are used as independent parameters of autoregression. The number of such parameters determines the moving window for the time series. We choose the window size (tw) to be 30 elements. (determined empirically).

We assume that the primary contribution to the approximating function for the next element in the time series comes from a combination of the previous 30 elements of the series. A neural network with one hidden layer and one output neuron is used. The number of neurons in the hidden layer (90) is three times larger than the size of the time window.

In the output layer, there is 1 neuron acting as a summation unit. A fully connected linear layer (nn.Linear) is used with the ReLU activation function. For regularization, neuron dropout is used with a dropout probability of 0.015. The mean squared error (MSE) loss function (nn.MSELoss()) is used. The optimization method used is torch.optim.Adam with a learning rate of 0.002. Preprocessing of the original series with scaling to the interval $[0, 1]$ is not performed because it introduces additional error into the raw data and leads to additional overhead costs for scaling before and after training.

A short time series of 64 samples is used. The interval between samples is $2s$. The duration of the series is $128s$. The original series is divided into two datasets — train (70% - 43 samples) and test (30% — 21 samples). Considering that tw $= 30$, the number of examples for training is 14. Training is conducted for 100 epochs. The batch size is set to 1. The training time for the neural network with the specified architecture on the dataset $\sim 2s$. The prediction execution time, with forecasting future 40 samples $\ll 1s$.

VIII. Description of NN Models Usage Process in the Control System

After obtaining the model for the current state, the main process transfers information about the model to the prediction process. The name of the model file is passed to the predict_file_q queue, and the current utilization data set for forecasting is passed to the predict_list_q queue. The prediction process checks for the presence of data in the specified queues at intervals corresponding to the data collection frequency. After reading the data from the queues, the prediction is executed, and the forecasted data is passed to the model_result_q queue. After receiving the forecast results for the current state, the main process calculates the maximum average utilization (P) across computational modules for the forecast window (Z).

Next, $max(R, P)$ is computed – the maximum between the current utilization value and the maximum forecasted value. This value, together with the code of the current state, the number of computational modules, and the stabilization limit, is used to make a decision about scaling the system. At the same time, if $max(R, P)$ exceeds the addition threshold (A), a decision is made to add resources to the managed system. If $max(R, P)$ is less than the removal threshold (D), a decision is made to remove resources. After the decision is made, a command is sent to the state change block.

IX. Example of the System Under Workload

As an example, Figure 5 illustrates the operation of the system under workload, gradually increasing to 600 users performing various requests to the system over approximately 0.5 hours.

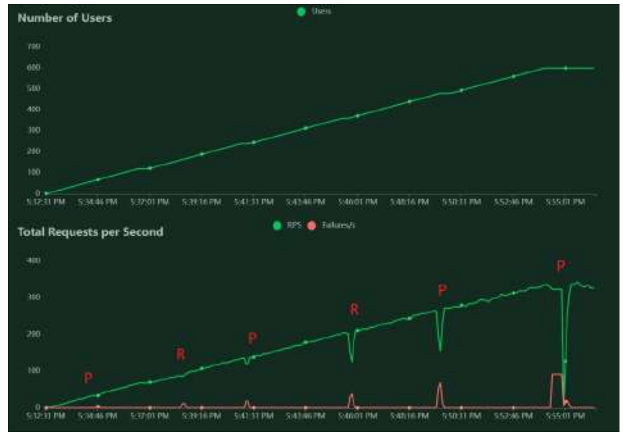


Figure 5. Model System Workload (Number of Users and RPS).

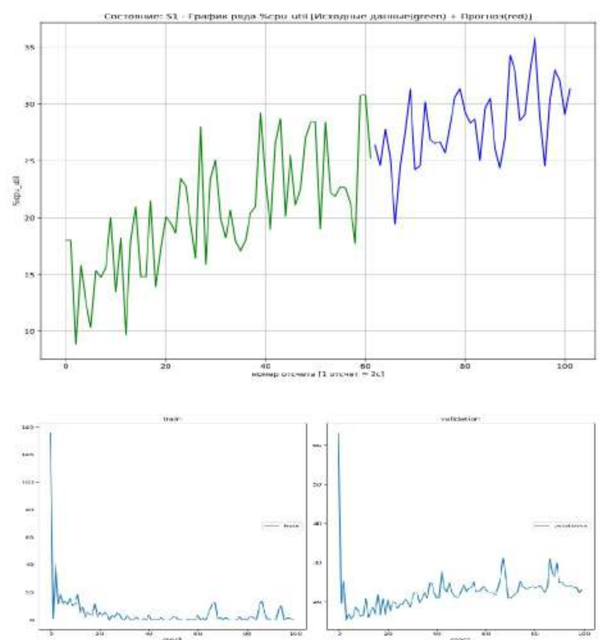


Figure 6. Forecast 1 for state S1.

As a result of the run, the system changed its state 6 times, with 4 changes being proactive and 2 being reactive.

Figure 6 shows an example of utilization forecast for computational modules in state S1 (green color represents the original data, blue color represents the forecast).

Figure 7 illustrates the situation with a reactive state change (sharp peak around the 112$^{\text{th}}$ sample). In this case, the neural network model did not win, although it predicted the state change threshold exceedance (50%).

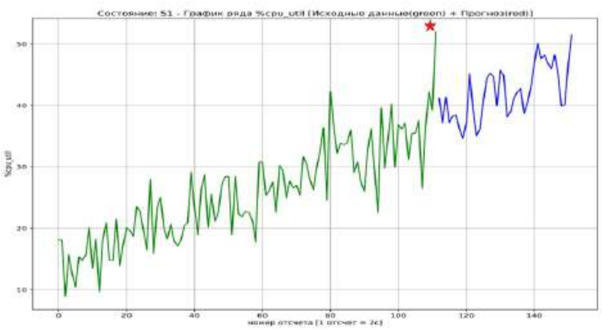


Figure 7. Forecast 51 for state S1.

This situation shows that the complexity of the process increases as the number of users (the number of requests to the system). The neural model does not account for the change in complexity, as the complexity (architecture) of the model itself does not change.

X. Results

The problem related with making real-time decisions in managing the computational resources of a critical IT service under conditions of uncertainty in external load was discussed in the article. As a result of the conducted research, the model system of critical IT service was developed. The architecture of the combined management system are proposed. The technology for real-time decision-making is developed and described, which has been implemented in practice. Experimental researches have been conducted, confirming the workability of the proposed technology.

The original multi-model approach to forecasting problem in case of generating control decisions is proposed, which enhances the adaptive properties and stability of the managed system to external workload. This approach allows to create a library of neural models capable of making forecasts needed parameters for various system states. It becomes possible to further complicate the predictor through the use of ensembles of models. At the same time, a more complex architecture (potentially capable of more accurate predictions) requires increased training time, which reduces the speed of decision making. This situation clearly shows that fast decision-making with preparing a more accurate forecast for a complex signal requires not only an improvement in the algorithm but also more computing resources with high performance.

References

[1] M. Straesser, J. Grohmann, J. von Kistowski, S. Eismann, A. Bauer, S. Kounev. Why Is It Not Solved Yet?: Challenges for Production-Ready Autoscaling // In ICPE '22: ACM/SPEC International Conference on Performance Engineering, Bejing, China, April 9 - 13, 2022, P. 105–115, ACM, 2022.

[2] N. Khan, D. A. Elizondo, L. Deka, M. A. M.–Cabello. Fuzzy Logic applied to System Monitors // IEEE Access, Vol. 9, P. 56523-56538, 2021.

[3] B. M. Nguyen, G. Nguyen, Giang. A Proactive Cloud Scaling Model Based on Fuzzy Time Series and SLA Awareness // Procedia Computer Science (International Conference on Computational Science ICCS 2017), 108, P.365–374, 2017.

[4] V. Persico, D. Grimaldi, A. Pescape, A. Salvi, S. Santini. A Fuzzy Approach Based on Heterogeneous Metrics for Scaling Out Public Clouds. IEEE Transactions on Parallel and Distributed Systems, Vol. 28, No. 8, P. 2117–2130, 2017.

[5] Docker Compose Overview [Elektronnyy resurs]. Rezhim dostupa: https://docs.docker.com/compose/. Data dostupa: 25.03.2024.

[6] Kubernetes Cluster [Elektronnyy resurs].Rezhim dostupa: https://kubernetes.io/. Data dostupa: 25.03.2024.

[7] Minikube Local Kubernetes Cluster [Elektronnyy resurs]. Rezhim dostupa: https://github.com/kubernetes/minikube/. Data dostupa: 25.03.2024.

[8] Docker SDK for Python [Elektronnyy resurs]. Rezhim dostupa: https://github.com/docker/docker-py/. Data dostupa: 25.03.2024.

[9] Echo LabStack High performance extensible minimalist Go web framework [Elektronnyy resurs]. Rezhim dostupa: https://echo.labstack.com/. Data dostupa: 25.03.2024.

[10] Go an open-source programming language supported by Google [Elektronnyy resurs]. Rezhim dostupa: https://go.dev/. Data dostupa: 25.03.2024.

[11] Apache JMeter [Elektronnyy resurs]. Rezhim dostupa: https://jmeter.apache.org/. Data dostupa: 25.03.2024.

[12] Locust [Elektronnyy resurs]. Rezhim dostupa: https://locust.io/. Data dostupa: 25.03.2024.

[13] Joblib Python library [Elektronnyy resurs]. Rezhim dostupa: https://joblib.readthedocs.io/en/latest/parallel.html/. Data dostupa: 25.03.2024.

[14] Multiprocessing Python library [Elektronnyy resurs]. Rezhim dostupa: https://docs.python.org/3/library/multiprocessing.html/. Data dostupa: 25.03.2024.

[15] Pytorch [Elektronnyy resurs]. Rezhim dostupa: https://pytorch.org/. Data dostupa: 25.03.2024.

[16] Starovoytov, A.A. Algoritm proaktivnogo upravleniya vychislitelnymi resursami. 80-ya nauchnaya konferentsiya studentov i aspirantov Belorusskogo gosudarstvennogo universiteta [Elektronnyy resurs. Materialy konf., Minsk, 10–20 marta 2023 g. V. 3 ch. Ch. 1. Belorus. gos. un-t, redkol.: A. V. Blokhin (gl. red.) [i dr.], Minsk, BGU, 2023, 398 s.

НЕЙРОСЕТЕВАЯ ТЕХНОЛОГИЯ ОПЕРАТИВНОГО УПРАВЛЕНИЯ ИТ СЕРВИСОМ

Краснопрошин В. В., Старовойтов А. А.

В работе иссследуется актуальная прикладная проблема, связанная с созданием систем принятия оперативных решений для управления ресурсами критически важных ИТ сервисов. Неопределенность внешней нагрузки является важным фактором, влияющим на оперативное управление. Для улучшения работы систем управления предлагается подход на основе мультимодельного нейросетевого прогнозирования. Описана модельная система критично ИТ сервиса. Предложена оригинальная технология, структура и архитектура системы управления. Проведены эксперименты, которые подтвердили работоспособность указанной технологии.

Received 13.03.2024

Stabilization of Parameters of Technological Operations in the Presence of External Control Actions

Viktor Smorodin
Department of Mathematical Problems of Control and Informatics
Francisk Skorina Gomel State University
Gomel, Belarus
Email: smorodin@gsu.by

Vladislav Prokhorenko
Department of Mathematical Problems of Control and Informatics
Francisk Skorina Gomel State University
Gomel, Belarus
Email: shnysct@proton.me

***Abstract*—A new technique for formalizing probabilistic technological processes when stabilizing the parameters of functioning of technological operations using open semantic technologies for designing intelligent systems is proposed. To stabilize the parameters of technological operations in real time, a procedure for adapting control to variable external control influences is presented based on the neuroregulators algorithms of an intelligent stabilization system that implement control feedback. The principles of development and implementation of intelligent system for stabilizing parameters and a controller for an automated process control system are described.**

***Keywords*—formalization methodology, neuroregulator algorithms, simulation models, stabilization of controlled parameters, controller**

I. Introduction

The operating efficiency of the created automated control systems for real objects largely depends on the quality and adequacy of mathematical models of the research object, which are used at the control system design stage [1]. For this reason, the most significant scientific results in the field of research of the operation of automated production systems and control systems are the development of effective algorithms for adaptive control of automated production based on new methods of neural network modeling of the research object [2]–[4]. The current stage of development of industrial automation means requires ensuring interoperability and semantic compatibility of heterogeneous components of intelligent systems [5].

The paper presents a methodology for synthesizing feedback loops for controlling the technological cycle of automated production based on the use of open semantic technologies for designing intelligent computer systems and algorithms for neural network modeling of control feedback loops.

II. Problems of stabilization of technological cycle parameters

The problem of determining the optimal parameters for controlling technological systems in real time is an important problem of production management in the presence of external control influences during technological operations and random disturbances associated with the design and reliability characteristics of the equipment.

This paper proposes a solution to the problem of stabilizing the parameters of the technological cycle based on the creation of a new generation intelligent computer system capable of stabilizing the parameters of the technological cycle in the presence of external disturbances in real time.

The use of mathematical models of neural networks within the framework of this approach ensures the creation of a new generation of intelligent systems for adapting the control of complex technical complexes, determining optimal control parameters and stabilizing the controlled variables of technological operations in specified ranges of acceptable values depending on external control influences and random disturbances.

The versatility of the proposed approach is determined by a limited set of parameterized procedures that implement algorithms for creating a knowledge base open for expansion based on the ontology of the "probabilistic technological production processes" subject area. The practical and economic significance of the results obtained are determined by the new opportunities provided by the study of existing and design of new complex technological objects.

The construction of a new generation intelligent computer system involves the following development stages:

- methods for formalizing the technological production process based on the use of the ontology of the subject area "technological production processes with probabilistic characteristics";

- schemes for the functioning of a hybrid intelligent computer system, providing the possibility of semantic compatibility and sharing with other solutions at industrial enterprises in the context of the Industry 4.0 concept [6];
- algorithms for the synthesis of control feedback based on the use of neuroregulators;
- method of adaptive control of automated production systems in the presence of external control influences;
- software support for intelligent systems based on adaptation algorithms and the proposed method in real time.

The results provide a new constructive approach to the formalization of the technological cycle, based on the use of open semantic technologies for the design of intelligent systems, and the synthesis of feedback on the management of the production process using neural network modeling, which allows to:

- develop hybrid intelligent computer systems designed to solve problems of adapting the management of complex technical systems, in particular, the technological production cycle;
- ensure integration and semantic compatibility of the components of the control system model with other intelligent systems;
- search for optimal parameters for adapting control and stabilizing controlled variables of technological operations in specified ranges of permissible values.

The described tools make it possible to create flexible intelligent computer control adaptation systems (CCAS), which are a set of semantically compatible and easily replaceable components depending on the range of tasks being solved.

The interaction of the system and its corresponding solvers with the control rack of the automated technological system is carried out using means of software and hardware interface of production.

III. Constructing components of a multi-level control system model in conditions of incomplete information

Optimization of the parameters of the technological cycle of automated production requires the development of effective control adaptation algorithms and methods for constructing neuroregulators that stabilize the parameters of technological operations, taking into account current information about the functioning of the object of study, random disturbances and external control influences, which are recorded during the operation of the control system controller and stored at the control rack automated control system (ACS) for the technological process (TP).

The structural diagram of the interaction of the components of the intelligent control adaptation system is presented in figure 1. A formal description of the control object is carried out based on the use of the ontology of the "probabilistic technological production processes" subject area. When implementing this approach, formalized knowledge is used to describe technological processes with probabilistic characteristics of technological operations and model technological processes.

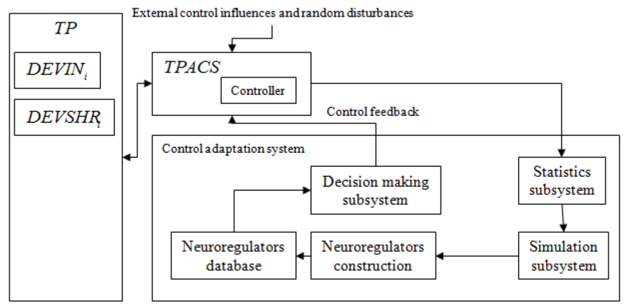


Figure 1. Diagram of interaction of the components of the intelligent control adaptation system

The formalization used is based on the scientific research of the authors in the field of simulation modeling of complex technical systems and implies the use of libraries of simulator units of technological operations integrated into the OSTIS ecosystem on the principles of semantic compatibility [7].

When the characteristics of technological operations of a control object are highly non-stationary, simulation models, neuroregulators and multi-step learning algorithms with better dynamic properties are used to build multi-level mathematical models.

As follows from the interaction diagram of the system components, the formation of control feedback comes down to the search for control adaptation that satisfies user-specified criteria, carried out according to the closed-loop principle, in which neuroregulator models are built on the basis of the collected statistics of the operation of the automated process control system and a collection of simulation models. The decision-making system, operating using the constructed neuroregulators, carries out in real time the formation of corrective influences on the controller of the automated control system of the TP.

The introduction of the Industry 4.0 concept at industrial enterprises is accompanied by the creation of a digital twin of the enterprise and the construction of a unified ontological production model, which is the core of comprehensive information services for the enterprise. One of the stages of building a digital twin model of an enterprise is the embedding of data on low levels of production, such as production processes and equipment [8].

IV. Methodology for formalizing a technological system when stabilizing control parameters

In order to ensure the possibility of using an intelligent control adaptation system, knowledge about the technological process of an enterprise must be recorded in a formal knowledge representation language. The sources of such knowledge can be existing descriptions of the work of enterprises within the framework of accepted international standards (such as ISA 5.1, ISA88) [6], [9]–[11]. Thus, within the framework of the ISA-88 standard, a technological cycle is called a procedure, and a technological operation is called a phase.

If there is a known set of devices and maintenance of the technological cycle, as well as statistical data on their operation, it is possible to move on to simulation modeling of the technological process by replacing devices in the probabilistic network diagram (PND) describing the cycle with units simulating the operation of shared use and individual use devices. The operations present in the model can be implemented as a set of event simulator units and technological operations simulator units. Based on simulation models of the technological cycle constructed using the current operation statistics it is possible to construct neuroregulators that carry out corrections of the controller's control actions [12].

With this approach, the adaptive control system requires a minimum amount of initial information about the incoming signals, and the described formalization makes it possible to synthesize knowledge bases about an industrial enterprise and its technological processes based on the domain ontology within the framework of the Industry 4.0 concept of automation of industrial enterprises.

V. Stabilization of control parameters based on an intelligent computer system

The basis for creating a hybrid intelligent control adaptation system is the idea of developing multi-level simulation models and mathematical models of neural network regulators [12] to solve problems of control optimization, constructing algorithms for synthesizing control feedback for the technological cycle depending on changes in the operating parameters of the controlled object.

The hybrid intelligent control adaptation system includes the following components:

- subsystem for processing and storing statistics of the operation of TP ACS;
- simulation subsystem;
- subsystem for constructing models of neuroregulators;
- database of constructed neuroregulators;
- decision-making subsystem.

TP operation statistics include the values of signals describing the state of TP devices, as well as the values of control signals. The statistics processing and storage subsystem is responsible for saving the historical values of signal data.

The simulation subsystem allows construction and execution of simulation models of a technological process and its control system based on the ontological model of production and the described formalization. Historical values saved by the subsystem for processing and storing performance statistics are used as initial data. Based on them, distribution functions are constructed for resources consumed by technological operations and reliability characteristics of equipment, stored in the knowledge base. Simulator units of the operation of devices for shared and individual use are the basis for creating TP simulation model.

The subsystem for constructing models of neuroregulators implements neural network modeling algorithms to find the optimal control adaptation strategy. Provided that there are known target values of correction signals (for example, in the case of manual data marking, or the presence of an existing high-quality regulator of the system regulator), a neuroregulator can be built using the collected statistics of the operation of the technological process and automated control system [13], [14]. In the absence of a regulator prototype, algorithms are used to search for the optimal policy for selecting actions in an environment built on the basis of a TP simulation subsystem [7], [12].

It should be noted that algorithms of this type allow to automate the search for an adaptation policy, taking into account the criteria specified by the system user for assessing the quality of the policy for selecting actions by the neuroregulator [12]. The target function of the neural network training algorithm may "reward" for reducing production costs and "fine" for equipment downtime, equipment failures or emergences based on the coefficients selected by system user.

Adaptive-critic-based schemes [15]–[20] for searching for an optimal control adaptation strategy have the potential to build an effective regulator of a control system due to the presence of a research element in the process of determining the optimal control policy, which can have a positive effect when solving problems with a complex structure of control decision-making space.

If necessary, a search for the optimal neural network architecture can also be carried out based on grid search or genetic algorithms [21] (Figure 3) using training and validation procedures to evaluate the candidate architectures. A feature of the latter approach is the tracking of genes using historical marks, which allows for the crossing of successful topologies, species division to preserve innovations, consistent movement from simple architectures to more complex ones.

The general scheme of the algorithm for constructing a neuroregulator is shown in Figure 3.

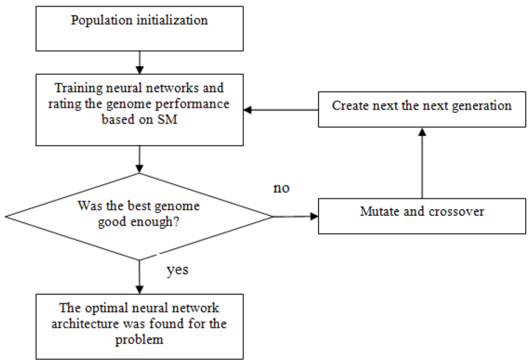


Figure 2. A genetic algorithm scheme to search for the optimal neural network architecture

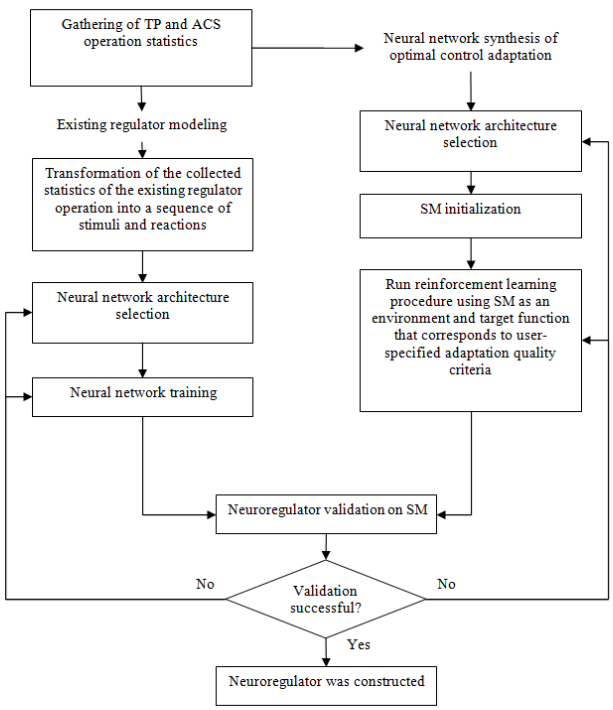


Figure 3. Neuroregulator synthesis algorithm scheme

Simulation modeling allows for model validation after training. Models of neuroregulators are saved to the database for further use or additional training using updated statistics.

The decision-making subsystem uses the constructed models of neuroregulators.

Adaptation of the technological cycle control is carried out on the basis of the operation of the constructed neuroregulator, which forms corrective influences on the process control system to prevent the range of changes in the components U_{fh} of the control vector variables U_h from going outside the allowed range.

Returning the values of control variables within the acceptable intervals is carried out by means of special adjustment means - initiating the launch of MTSO TP, which change the values of the components of the set of control variables U_s or use equipment redundancy schemes.

Provided that appropriate hardware and software interface tools are available, it is possible to implement control adaptation in an automated mode, or to form a recommendation system used by personnel servicing the technological cycle.

Thus, when solving the problem of stabilizing the parameters of technological operations in real time, multi-level mathematical models were used, including neural network and simulation ones, and a new generation hybrid intelligent computer adaptive control system was implemented, created on the basis of open semantic technologies for designing intelligent systems [7].

VI. An example of a simulation model of a probabilistic technological process control system

Simulation modeling of the interaction of control system elements and components of a probabilistic technological process can be carried out on the basis of synchronization schemes of their operation [7].

Simulation modeling of the implementation of a probabilistic technological process includes modeling of microtechnological operations associated with resource requests of a probabilistic nature, as well as possible changes in control variables U, and the operation of devices with possible device failures and their consequences.

The initial parameters for constructing a control system simulation model (CS SM) of an probabilistic technical process (PTP) are a set of parameters characterizing its composition and structure [7], including estimates of the distributions of various parameters of the TP, collected during its operation observation using appropriate software and hardware interface tools. In particular, these include the number of implementations of the simulation model N, the set X_i of PTP resource requests required to run the executive elements of the control system, the parameters of commutation of the executive elements with synchronization elements, input signal generators and the final operating element, configuration vector of the equipment composition and a set of reliability characteristics of the equipment.

Simulation modeling of control system of TP is based on the operation of a set of simulator units. Within the framework of OSTIS technology, the solution to the problem of operation of simulator units can be formalized on the basis of a multi-agent approach [7], when the SM of the PTP control system is implemented as a set of agents corresponding to simulator units that can exchange data through sc-texts.

Within the framework of the problem under consideration, the solver of the problem of operation of a simulator unit can be represented as a decomposition of an abstract non-atomic agent (Figure 4).

abstract non-atomic sc-agent of unit-simulator operation
⇒ *decomposition of abstract sc-agent**:
{• *abstract sc-agent of input signals processing*
• *abstract sc-agent of output signals generation*
}

Figure 4. Decomposition of the abstract non-atomic sc-agent of the operation of the simulator unit of the CS SM of TP

The operation of the simulator unit depends on some set of input signals that may be a result of operation of other simulator units or signal generator units. The simulator unit must process them and generate the appropriate output signals according to certain rules. Thus, in the context of the OSTIS ecosystem, the agent of the simulator unit operation is represented by a set of two agents for processing input signals and generating output signals.

Figure 5 shows a block diagram of a control system for a probabilistic technological process when modeling the problem of stabilizing operating parameters by searching for an optimal strategy for servicing TP devices.

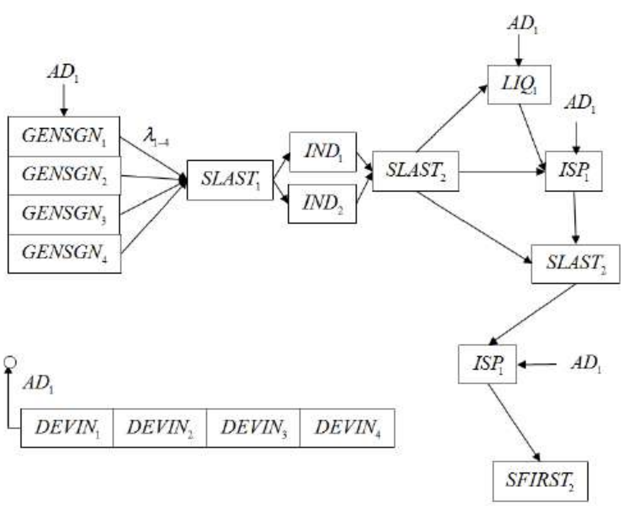


Figure 5. Example of a PTP control system

This control system is responsible for the execution of the technological cycle, during which maintenance of technological cycle equipment devices can occur, as well as the elimination of equipment failures and the elimination of emergences when they occur. The equipment of the technological cycle is represented by 4 devices for individual use, each of which performs a corresponding microtechnological operation.

$GENSGN_{1-4}$ are signal generator units, the intensity of which is set by the corresponding distribution of reliability characteristics of equipment devices. The intensities of the generated signals during the simulation process determine the probability of failures occurring in the corresponding devices.

Synchronizer $SLAST_1$ (type "AND") starts the process of indicating equipment (IND_1) and control variables (IND_2). If an accident is indicated on one of the equipment devices, the consequences of the accident are eliminated (LIQ_1). If the specified intensity value of the control variable (variables) (U) is exceeded, the corresponding device is serviced (ISP_1 - the corresponding operation is performed). If the maintenance was completed or was not performed, microtechnological operations determined by the technological process are performed on each of the devices.

When the $SFIRST_2$ unit is executed, the technological cycle finishes it's operation.

VII. Conclusion

The paper proposes a method for constructing a new generation intelligent computer system for adapting the control of the technological cycle of automated production with control feedback in the presence of random disturbances and external control influences in the integrated environment of open semantic technologies for designing intelligent systems OSTIS. Control feedback is formed on the basis of algorithms for constructing neuroregulators using simulation modeling to find the optimal control adaptation strategy according to specified criteria. An example of a simulation model of a TP control system is given when solving the problem of servicing TP equipment devices and a method for describing its elements within the OSTIS ecosystem is given.

When considering the problem of synthesizing the optimal structure of a technological system with an arbitrary organization of the technological production process and stabilizing the parameters of technological operations, algorithms for constructing control feedback based on genetic algorithms and artificial neural network models were implemented. Stabilization of the parameters of technological operations is carried out within the framework of solving the multi-criteria problem of assessing the quality of the technological process while minimizing the costs of performing a closed technological production cycle based on effective control adaptation algorithms.

References

[1] Maximey, I. V., Demidenko, O. M., Smorodin, V. S. Problems of theory and practice of modeling complex systems. F. Skorina State University, Gomel, 2015, P. 263

[2] Hagan M. T., Demuth H. B. Neural networks for control. Proceedings of the American Control Conference. San Diego, USA, 1999, Vol. 3, P. 1642—1656.
[3] Omidvar O., Elliott D.L. eds. Neural Systems for Control. Academic Press, New York, 1997, P. 358
[4] White D. A., Sofge D. A. Handbook of Intelligent Control. Neural, Fuzzy, and Adaptive Approaches. Van Nostrand Reinhold, 1992.
[5] V. Golenkov, N. Gulyakina, N. Grakova, I. Davydenko, V. Nikulenka, A. Eremeev, V. Tarasov. *From Training Intelligent Systems to Training Their Development Tools. Open Semantic Technologies for Intelligent Systems (OSTIS)*, Minsk, Belarussian State University of Informatics an Radioelectronics Publ., 2018, iss. 2, pp. 81–98.
[6] V. Taberko, D. Ivaniuk, D. Shunkevich, O. Pupena Principles for enhancing the development and use of standards within Industry 4.0, Otkrytye semanticheskie tekhnologii proektirovaniya intellektual'nykh system [Open semantic technologies for intelligent systems], 2020, pp. 167-174.
[7] V. Golenkov, Ed., Tehnologija kompleksnoj podderzhki zhiznennogo cikla semanticheski sovmestimyh intellektual'nyh komp'juternyh sistem novogo pokolenija [Technology of complex life cycle support of semantically compatible intelligent computer systems of new generation]. Bestprint, 2023.
[8] V. Taberko, D. Ivaniuk, N. Zotov, M. Orlov, O. Pupena, N. Lutska Principles of building a system for automating the activities of a process engineer based on an ontological approach within the framework of the Industry 4.0 concept, Otkrytye semanticheskie tekhnologii proektirovaniya intellektual'nykh system [Open semantic technologies for intelligent systems], 2021, pp. 209-218.
[9] "ISA5.1 Standard," https://www.isa.org/standards-and-publications/isa-standards/isa-standards-committees/isa5-1/, (accessed 2024, Jan)
[10] "ISA-88 standard," Available at: https://www.isa.org/isa88/, (accessed 2024, Jan).
[11] "ISA-95 standard," https://www.isa.org/standards-and-publications/isa-standards/isa-standards-committees/isa95/, (accessed 2024, Jan)
[12] V. Smorodin, V. Prokhorenko Software-Technological Complex For Adaptive Control Of A Production Cycle Of Robotic Manufacturing, Otkrytye semanticheskie tekhnologii proektirovaniya intellektual'nykh system [Open semantic technologies for intelligent systems], 2022, pp. 401-404.
[13] V. Smorodin, V. Prokhorenko Application of Neuro-Controller Models for Adaptive Control. Chertov O., Mylovanov T., Kondratenko Y., Kacprzyk J., Kreinovich V., Stefanuk V. (eds) Recent Developments in Data Science and Intelligent Analysis of Information. ICDSIAI 2018. Advances in Intelligent Systems and Computing, 2019, vol 836, pp. 30-38
[14] V. Smorodin, V. Prokhorenko Control Of A Technological Cycle Of Production Process Based On A Neuro-Controller Model, Otkrytye semanticheskie tekhnologii proektirovaniya intellektual'nykh system [Open semantic technologies for intelligent systems], 2019, pp. 251-256.
[15] Sutton, R. S., Barto, A. G. *Reinforcement Learning: An Introduction*, Cambridge, The MIT Press, 1998
[16] Watkins, C. J. C. H., and Dayan, P. Q-learning. *Machine Learning*, 1992, vol. 8, no 3-4, pp. 279–292
[17] Sutton, R. S., McAllester, D., Singh S., Mansour Y. *Policy Gradient Methods for Reinforcement Learning with Function Approximation, Advances in Neural Information Processing Systems 1*, NIPS 1999
[18] Bakker, B. Reinforcement Learning with Long Short-Term Memory, *NIPS*, 2001
[19] Mnih V., Kavukcuoglu K., Silver D., Rusu A., Veness J., Bellemare M., Graves A., Riedmiller M., Fidjeland A., Ostrovski G., Petersen S., Beattie C., Sadik A., Antonoglou I., King H., Kumaran D., Wiestra D., Legg S., Hassabis D. Human-level control through deep reinforcement learning. *Nature*, 2015, vol. 518, no 7540, pp.529–533
[20] Hausknecht M., Stone P. Deep recurrent q-learning for partially observable mdps. *2015 AAAI Fall Symposium Series*, 2015
[21] K. Stanley, R. Miikkulainen Evolving Neural Networks through Augmenting Topologies, Evolutionary computation, 2002, pp.99-127

СТАБИЛИЗАЦИЯ ПАРАМЕТРОВ ТЕХНОЛОГИЧЕСКИХ ОПЕРАЦИЙ ПРИ НАЛИЧИИ ВНЕШНИХ УПРАВЛЯЮЩИХ ВОЗДЕЙСТВИЙ

Смородин В. С., Прохоренко В. А.

Предложена новая методика формализации вероятностных технологических процессов при стабилизации параметров функционирования технологических операций с использованием открытых семантических технологий проектирования интеллектуальных систем. Для стабилизации параметров технологических операций в режиме реального времени представлена процедура адаптации управления под изменяющиеся внешние управляющие воздействия на основе разработки алгоритмов нейрорегуляторов интеллектуальной системы стабилизации, реализующих обратные связи по управлению. Изложены принципы создания и разработки системы стабилизации параметров и контроллера автоматизированной системы управления технологическим процессом.

Received 13.03.2024

Crowd Attention Estimation Automatisation Based on Semi-Automatic Image Semantic Segmentation by Using UNet and CRF Networks

Stanislav Sholtanyuk, Yakov Malionkin
The Department of Computer Applications and Systems
Belarusian State University
Minsk, Belarus
ssholtanyuk@bsu.by, malionkinjr@gmail.com

Bin Lei
Nanjing Research Institute of Electronics Engineering
Nanjing City, China
40220720@qq.com

Alexander Nedzved
The Department of Information Management Systems
Belarusian State University
Minsk, Belarus
Anedzved@bsu.by

***Abstract*—Semantic segmentation of crowd images plays a pivotal role in various applications such as crowd management, surveillance, and urban planning. In this paper, we propose an approach for dense and sparse crowd image semantic segmentation based on semi-automatic labeling by employing a combination of UNet and Conditional Random Field (CRF).**

We introduce a technique for generating segmentation maps for crowd images. We utilize UNet for initial rough segmentation followed by refinement using CRF. Experimental results demonstrate the model performs better in binary segmentation (crowd and on-crowded regions) rather than ternary segmentation (dense crowds, sparse crowds, and non-crowded areas). However the latter shows better results in terms of crowd detection (regardless of its type). Besides, we show the CRF refinement is significant in ternary segmentation.

Also, we highlight some crowd behavior patterns based on the proposed segmentation model. They differ in people's attention types, connections within and between crowds, and possibilities of emergencies.

***Keywords*—artificial neural networks, computer vision, crowd images, crowd detection, crowd behavior, image analysis, machine learning, semantic segmentation**

I. Introduction

Segmentation is the process of breaking the image into distinct segments or regions that represent objects of interest or their structure. Image segmentation is one of the pivotal stages in computer vision and image analysis. The main purpose of segmentation is to highlight key objects and their features for a more detailed understanding of the content represented in the image. More rigorously, during image segmentation, a label representing a certain class is assigned to each image pixel so pixels in the same class stand for a joint object and demonstrate some shared characteristics, and pixels with different labels somehow differ from each other.

Image segmentation can be compared with a classification task with some initial classes given. However, the image classification task implies a label for the whole image as the result, whereas in semantic segmentation, a label is assigned to each image pixel. Thus, unlike classification tasks, not only does segmentation have a purpose to determine the main object of interest but also to examine its optical and morphological characteristics like its edges, position on the image, and its position relative to other objects (if any).

In computer vision, there are two main types of image segmentation: semantic and instance segmentation. In semantic segmentation, each pixel must be associated with one of the predefined classes (e. g. background, person, vehicle, building, etc.), and such classes are represented by their colors. In instance segmentation, each pixel is also classified based on some given classes, but distinct objects within a class are highlighted by different colors.

Semantic segmentation of crowd images is an effective tool for analyzing and understanding crowded scenes, crowd structure, and behavior. Crowd semantic segmentation can be used in the following applications:

1) Counting and analyzing the crowd. Segmentation can facilitate such tasks by decomposing the image into clusters, and some basic techniques could be applied to them to analyze the crowd structure within them. It is useful for monitoring crowded places like stadiums, markets, malls, fairs, social events, etc. [1]–[4].
2) Security and surveillance. Some segmentation techniques allow highlighting single persons which is important for detecting abnormal, troublesome,

The research was supported by the project "Technology Development Agreement of developing of algorithms of remote sensing image processing", agreement number 22CETC19-ICN1785

or even potentially dangerous situations, e.g. lost things, suspicious behavior, civil unrest, etc. [5], [6].

3) Marketing and analytics. In marketing, crowd segmentation can be used for analyzing customers' behavior when they wander, seek something, or stop near a merchant's place in malls and fairs. In such context, segmentation can be a helping tool to improve goods placing, estimate marketing strategies efficiency, and improve customer service.
4) Transport management. In urban planning and traffic management, crowd segmentation can be used for pedestrian traffic optimization, crowd prevention, as well as planning more efficient pedestrian and transport routes.
5) Social behavior research. Nowadays, crowd segmentation is effectively used to research social behavior, e. g. by analyzing the dynamics of interaction between people in different scenes.

However, semantic segmentation of crowd images is a challenging task due to the complex and dynamic nature of crowd scenes, which often exhibit variations in density, scale, occlusions, and illumination conditions. Accurate segmentation of individual objects within a crowd is crucial for the above-mentioned applications. Traditional methods for crowd segmentation rely on handcrafted features and manual annotation, which are labor-intensive and often fail to capture the diverse characteristics of crowd scenes.

In recent years, semantic technologies are widely used in various computer vision applications such as knowledge based computer vision [7]–[9], video and images annotation [10], and video retrieval [11], [12]. In fact, they have the potential to significantly enhance the capabilities of crowd segmentation and attention estimation systems by incorporating semantic understanding into the analysis process. By leveraging semantic technologies, such as ontologies and knowledge graphs [13], [14], as well as semantic reasoning mechanisms [9], [15], researchers and practitioners can improve the accuracy, efficiency, and interpretability of crowd segmentation and attention estimation algorithms.

One key aspect of applying semantic technologies to crowd segmentation is the incorporation of domain-specific knowledge about crowd behavior, scene context, and environmental factors [16], [17]. By encoding this knowledge into formal ontologies or knowledge graphs, segmentation algorithms can better understand the semantics of crowd scenes, leading to more robust and context-aware segmentation results [18], [19]. Furthermore, semantic reasoning mechanisms can be used to infer higher-level semantic concepts from low-level segmentation outputs, enabling the identification of complex crowd behaviors and interactions [9].

In context of OSTIS systems, deep learning techniques have shown remarkable success in various computer vision tasks, including semantic segmentation. Convolutional Neural Networks (CNNs) have emerged as powerful tools for learning discriminative features directly from data, enabling end-to-end training for semantic segmentation tasks [21]–[24].

II. Main Semantic Segmentation Techniques Survey

Semantic segmentation, the task of assigning semantic labels to each pixel in an image, has witnessed significant advancements in recent years driven by deep learning techniques. In this subsection, we provide an overview of the main semantic segmentation techniques, focusing on classical and deep learning-based approaches.

A. Sliding Window

A simple semantic segmentation method using a sliding window involves sequentially applying the window of different sizes to the entire image [25], [26]. This process consists of several stages like setting the size of the window, applying it to the image, classifying the extracted features, and building the semantic map. Such an approach, however, possesses multiple cons some of them being:

1) High computational complexity. When working with high-resolution images, step-by-step window displacements lead to excessive iterations, during which several time-consuming operations are performed.
2) The lack of a global context. As far as each region is processed independently, such a technique grasps little to no connection between regions. It might result in a fragmented representation of the object and a lack of understanding of the whole picture.
3) Different size objects predicament. If the image depicts multiple objects of interest with different sizes, then the fixed-sized window might struggle with extracting features from some of them. Dynamic resizing of the window is likely to pose extra computational difficulties.
4) Objects overlapping. When there are some overlapping objects on the image, the sliding window method is likely to give poorly highlighted edges, especially if the objects are close to each other or have the same size.
5) Sensitivity to the parameters. Several parameters like window size or the step value should be fine-tuned precisely. Otherwise, the result might get worse dramatically.

The method can be used in remote monitoring when an observer is so distant that the scene can be considered as infinitely distanced from them. Another suitable condition to use the approach is the equality of sizes of interesting objects so one could predetermine the window size.

B. Fully Convolutional Networks

Fully convolutional networks (FCN) are the type of neural network designed for semantic segmentation tasks. Instead of using fully connected layers, FCNs use convolutional layers. It allows processing input images of arbitrary size and generating segmentation maps with the same size [24]. The main concept of FCN is replacing fully connected layers with convolutional ones to obtain a segmentation map of the same size as the initial image. Besides, some intermediate layers and skip connections between them could be used to improve the segmentation and get more detailed information.

Fully convolutional networks possess the following drawbacks:

1) Ineffectiveness with objects having different sizes. FCNs might concentrate more on larger objects, neglecting smaller ones.
2) The spatial information loss. Using maximal pooling layers and image upscaling might lead to spatial information loss, especially if some features are neglected in the convolutional layers.
   Furthermore, FCNs are characterized by a vast amount of parameters. As far as FCNs use convolutional layers, the number of parameters might significantly exceed compared to simple models. This issue raises even more drawbacks:
3) Computational complexity. Estimating the parameters and fine-tuning the FCN requires vast time and computational resources.
4) Training data requirements. A large number of parameters arise need in big training data which must be well-prepared to avoid network overfitting which causes poor ability of the model to make general results.
5) FCNs are prone to overfitting.

C. Convolutional Neural Networks

Convolutional neural networks (CNN) are one of the key frameworks in image processing and computer vision. They combine such tools as image convolution, image pooling, feature extracting, and classification based on those features [27]. The main idea is to use multiple layers of different types: convolutional (to extract features), pooling (to solve image size-related issues), and fully connected layers (to classify the image based on the extracted features). CNNs have established their place in computer vision and image processing thanks to many advantages like effectiveness in extracting semantic, morphological, and spatial features, their ability to process images of different sizes, as well as their ability to identify the image context.

D. Conditional Random Fields

Conditional random fields (CRF) is a statistical model effectively used with CNNs to refine semantic segmentation maps. A CRF takes part in postprocessing the results of a CNN prediction to refine and improve the segmentation spatial structure [28], [29]. The common way to use CRF in image semantic segmentation features the next stages:

1) Receiving predictions from the CNN. The prediction includes a segmentation map with probabilities for each pixel belonging to each considered class.
2) Preparing the features for the CRF. The probabilities from the segmentation map are used to form features given to the input of the CRF. Spatial coordinates of separate pixels or objects may also be such features.
3) Applying the CRF to refine the segmentation. The CRF uses context and spatial data from the CNN to refine the segmentation. CRF usually models interconnections between neighbor pixels and implements that information into the semantic map.
4) MAP optimization. The CRF uses the MAP method (Maximum A Posteriori) to tune its parameters to maximize the *a posteriori* probability for each pixel to fit in the appropriate class.

CRFs provide context information based on the information on interconnections between pixels. That allows us to improve edge detection and highlighting object details. Besides, CRF might reduce noise and smooth predictions which is extremely important in applications where high-quality object separation is crucial.

III. Methodology

In the research, we consider the following task of crowd semantic segmentation. Based on various characteristics (e. g. crowd density, people's spatial distribution, their visual texture), crowds can be classified as dense and sparse. Different approaches can be employed to determine if the given crowd is dense or sparse, e.g. manual annotation, crowd density maps, computer vision methods considering the texture of the image, as well as social force models [30], [31]. In this paper, we use a semi-automatic approach to generate ground truth maps for binary (non-crowded and crowd regions) and ternary (non-crowded, sparsely crowded, and densely crowded areas) semantic segmentation. This features the following steps:

- An annotated crowd images dataset is used (Fig. 1a). For each image, the annotations present the locations of the labels assigned to each individual's head.
- Based on the labels' locations, a 2D binary array is assigned to each image where 0 indicates the absence of a person's head, and 1 stands for a label.
- The array is Gaussian blurred to obtain density maps (Fig. 1b).

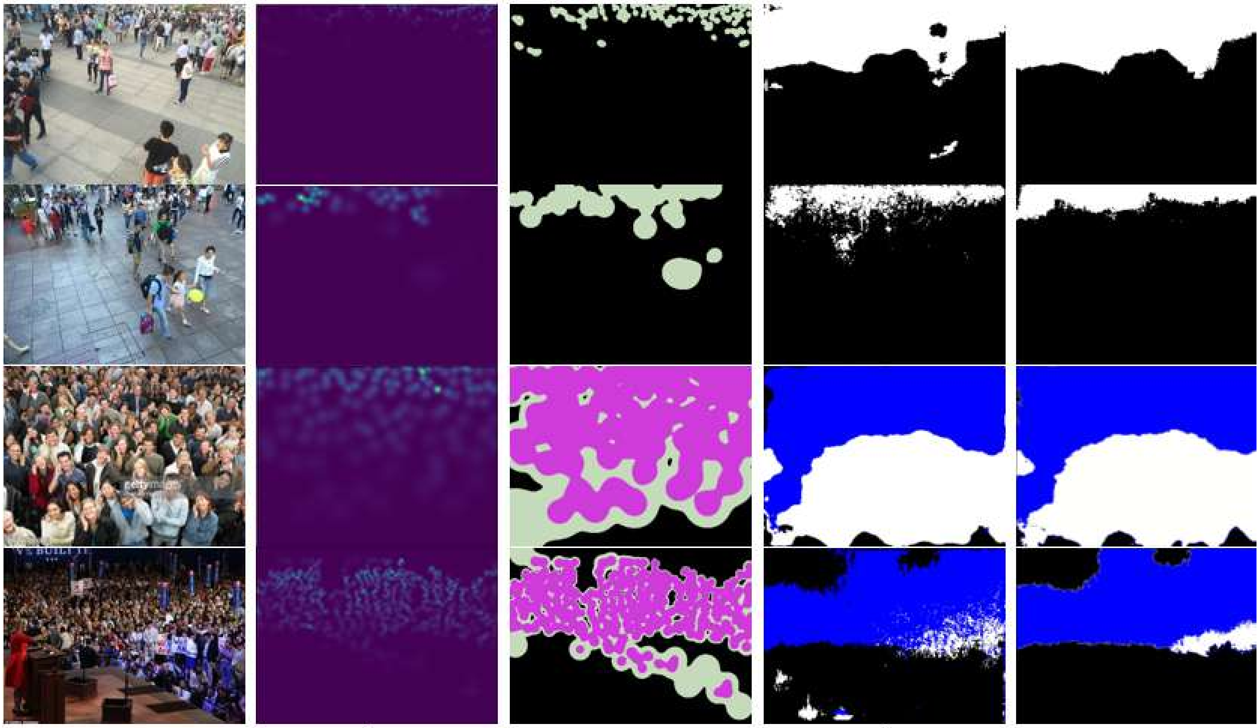


Figure 1. A crowd image (a), the corresponding density map (b), ground truth segmentation (purple is dense crowd, green is sparse crowd, and black is non-crowded areas) (c), predicted segmentation after employing UNet (blue is dense crowd, white is sparse crowd, and black is non-crowded areas) (d), and the segmentation refined after using CRF (e). Each row presents an example of binary and ternary prediction (first two rows and last two rows respectively) using either dice (first and third lines) or focal (second and fourth) loss function

- The resulting density maps are segmented into two or three areas based on thresholding values (Fig. 1c).

Based on the analysis given above, we decided to use a CNN + CRF model for the task. We use UNet as the network to calculate initial predictions. The initial images and the corresponding ground truth segmentation maps are divided into training, validating, and testing samples to train the UNet neural network which effectively takes advantage of the depicted objects' semantic, morphological, and spatial characteristics. The neural network gives an initial segmentation map (Fig. 1d). After obtaining the initial segmentation map, a CRF is used to refine it. As a result, we get the final crowd segmentation based on the individuals' head location (Fig. 1e).

After obtaining the final segmentation maps, some metrics based on the relations between ground truth and obtained maps are calculated to evaluate the final prediction accuracy. Based on such metrics, we can compare the impact of various parameters on the final semantic segmentation.

A. *ShanghaiTech Dataset*

For the experiment, we use a highly recognized ShanghaiTech dataset [32]. It consists of two parts. Part A features 482 crowd images taken from the Internet. In each image, there are from 33 to 3138 individuals, and the majority of the images represent dense crowds. Part B consists of 716 images. The images contain mainly sparse crowds from 9 to 576 people. Hence, we decided to use the B part to train the model for binary segmentation (crowd and non-crowded regions), and the A part for ternary segmentation (dense crowd, sparse crowd, no crowd). In both parts, 100 images form the training sample, 100 images — the validating sample and other ones are used to test the trained model.

To generate ground truth segmentation, we build density maps first. We do so by using Gaussian blurring (Fig. 1b). After that, we segment density maps based on thresholding values. We use two thresholds:

$$\mu_1 = 0.001M,$$
$$\mu_2 = 0.01M,$$

where M stands for the maximal value in the considered density map. For binary segmentation, only μ_2 is used (Fig. 1c).

B. *UNet*

UNet, one of the deep learning networks with an encoder-decoder architecture, is a popular neural network architecture designed for semantic segmentation tasks, particularly in biomedical image segmentation [21]. It makes maximal use of feature maps in full scales for accurate segmentation and efficient network architecture with fewer parameters.

The architecture is characterized by its U-shaped design, consisting of a contracting path (encoder) followed by an expansive path (decoder) allowing for precise localization while capturing contextual information (Fig. 2). The EfficientNetB3 was used as the encoder [33], [34]. It consists of convolutional (Fig. 2, blue blocks) and bottleneck layers (Fig. 2, yellow blocks) [35]. In the middle column of Fig. 2, dashed arrows denote skip connections between layers in the encoder and the decoder which help the decoder part recover spatial information lost during downsampling in the encoder part. In the research, we used Segmentation Models, a Python library with Neural Networks for Image Segmentation based on well-known Keras and TensorFlow frameworks [36].

The input images must have 3 channels (e. g. RGB) and have the same size. To effectively use the network, we resize initial images and the corresponding ground truth segmentations to 256x256. This allows us to get the results quickly without significant quality loss. However, other sizes could also be used, e. g. 224x224 which is a standard input size for EfficientNetB3, as well as bigger sizes. Also, the following parameters were used during the training: batch size is 5, optimizer function is Adam, the learning rate is 0.0001, the activation function on the last layer is softmax, and the number of epochs is 50.

Besides, we consider two loss functions that take into consideration the spatial characteristics of segments. Dice loss is widely used as a metric showing how much two images are similar to each other [23], [37], [38]:

$$GDL = 1 - 2\frac{\sum\limits_{l=1}^{C} w_l \sum\limits_{n=1}^{N} r_{ln}p_{ln} + \varepsilon}{\sum\limits_{l=1}^{C} w_l \sum\limits_{n=1}^{N} (r_{ln} + p_{ln}) + \varepsilon},$$

where N is the number of pixels on each of two compared images, C is the number of classes, $p_l n$ and $r_l n$ are probabilities for the n-th pixel from both images to be in the l-th class, w_l is a normalizing coefficient, ε is a term to avoid division by zero. In the paper, value $\varepsilon = 10^{-7}$ is used.

Focal loss is the metric widely used in image classification and segmentation tasks [37], [39]. It derives from the cross entropy concept and addresses the one-stage object detection scenario in which there is an extreme imbalance between foreground and background classes during training. The loss function is calculated according to the formula:

$$FL(p) = -(1-p)^{\gamma}\log(p),$$

where p is the model's estimated probability for a pixel to belong to a certain class, and γ is the focusing parameter to down-weight easy examples and focus on training on hard ones. We used the default value which is $\gamma = 2$.

Both functions are suitable for binary and ternary semantic segmentation. Besides, they require only a few parameters to define, so the models don't become too hard to tune.

C. CRF

In the research, we use PyDenseCRF, a Cython-based Python wrapper for a fully connected CRF with a highly efficient approximate inference algorithm implemented in which the pairwise edge potentials are defined by a linear combination of Gaussian kernels [28], [40]. Concepts of appearance and smoothness are used in the network to calculate *a posteriori* probabilities for each pixel belonging to each class. Appearance is the property of a segmentation map to have nearby pixels of the same color likely belonging to the same class. In smooth models, large classes must absorb small isolated regions nearby. The formalization of both concepts can be expressed by the formula:

$$k(\vec{f_i}, \vec{f_j}) = w^{(1)}\exp\left(-\frac{|p_i - p_j|^2}{2\theta_\alpha^2} - \frac{|I_i - I_j|^2}{2\theta_\beta^2}\right) + \\ + w^{(2)}\exp\left(-\frac{|p_i - p_j|^2}{2\theta_\gamma^2}\right),$$

where p_i and p_j – positions of two pixels, I_i and I_j – their colors, $\vec{f_i}$ and $\vec{f_j}$ – their features vectors, k – the similarity function to be maximized, $w^{(1)}$ and $w^{(2)}$ – linear combination weights, θ_α, θ_β, θ_γ are initially defined parameters. In the research, we used the following values: $\theta_\alpha = 10$, $\theta_\beta = 20$, $\theta_\gamma = 1$.

D. The Results Processing

After evaluating the final segmentation maps, some metrics functions are calculated to perform pixel-wise comparison ground truth maps with obtained results.

For binary segmentation, we calculated four values for each pair:

$$\begin{aligned} TP &= |\{(i,j) : p_{ij} = 1 \wedge \hat{p}_{ij} = 1\}|, \\ FP &= |\{(i,j) : p_{ij} = 1 \wedge \hat{p}_{ij} = 0\}|, \\ TN &= |\{(i,j) : p_{ij} = 0 \wedge \hat{p}_{ij} = 0\}|, \\ FN &= |\{(i,j) : p_{ij} = 0 \wedge \hat{p}_{ij} = 1\}|, \end{aligned}$$

where (i, j) stands for the position of two pixels to be compared, $\hat{p}_{ij}$ is the value for the ground truth pixel (1 stands for crowded region, and 0 means the pixel belongs to non-crowded area), and p_{ij} is the value for the pixel on predicted map. After that, accuracy, crowd predictive value, and non-crowded predictive value are calculated:

$$\begin{aligned} acc &= \tfrac{TP+TN}{TP+FP+TN+FN}, \\ cpv &= \tfrac{TP}{TP+FP}, \\ npv &= \tfrac{TN}{TN+FN}. \end{aligned}$$

Besides, for each image, we calculated the number of annotation labels that fell into each class (according to ground truth and predicted segmentation maps):

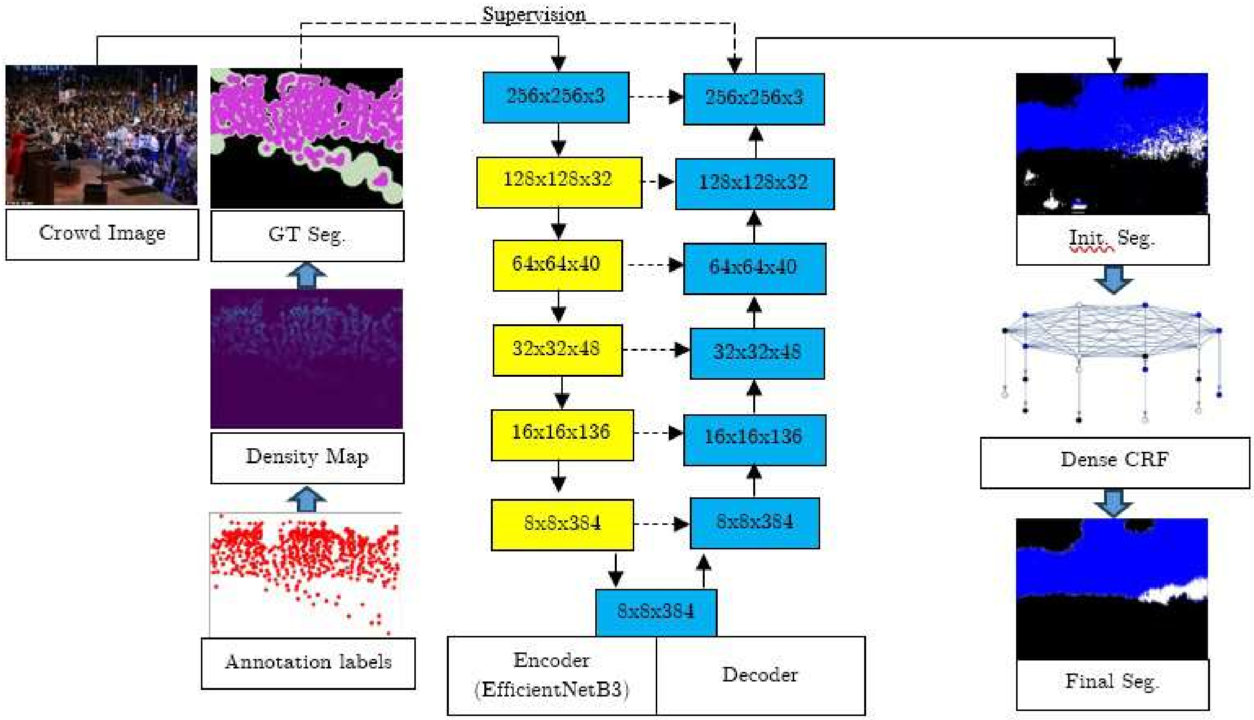


Figure 2. The crowd segmentation prediction framework consisting of image preprocessing (left column), the UNet network (middle column), and dense CRF to refine the predicted segmentation (right column)

$$
\begin{aligned}
N_0 &= |\{i : p_{h_i} = 0\}| \\
\hat{N}_0 &= |\{i : \hat{p}_{h_i} = 0\}|, \\
N_1 &= |\{i : p_{h_i} = 1\}|, \\
\hat{N}_1 &= |\{i : \hat{p}_{h_i} = 1\}|,
\end{aligned}
$$

where N_i stands for the number of annotation labels felt in the class labeled by the number i, and h_i is the location of the i-th annotation label. Then, to estimate the rate of the crowd detection, we calculated the $N_1/\hat{N}_1$.

For the ternary segmentation, the following characteristics are calculated:

$$
\begin{aligned}
DD &= |\{(i, j) : p_{ij} = 2 \wedge \hat{p}_{ij} = 2\}|, \\
DS &= |\{(i, j) : p_{ij} = 2 \wedge \hat{p}_{ij} = 1\}|, \\
DN &= |\{(i, j) : p_{ij} = 2 \wedge \hat{p}_{ij} = 0\}|, \\
SD &= |\{(i, j) : p_{ij} = 1 \wedge \hat{p}_{ij} = 2\}|, \\
SS &= |\{(i, j) : p_{ij} = 1 \wedge \hat{p}_{ij} = 1\}|, \\
SN &= |\{(i, j) : p_{ij} = 1 \wedge \hat{p}_{ij} = 0\}|, \\
ND &= |\{(i, j) : p_{ij} = 0 \wedge \hat{p}_{ij} = 2\}|, \\
NS &= |\{(i, j) : p_{ij} = 0 \wedge \hat{p}_{ij} = 1\}|, \\
NN &= |\{(i, j) : p_{ij} = 0 \wedge \hat{p}_{ij} = 0\}|,
\end{aligned}
$$

where 0 stands for non-crowded pixels, 1 is sparse crowd, and 2 is dense crowd. Based on these values, we calculate metrics that we call accuracy, dense crowd predictive value, crowd predictive value, and non-crowded predictive value:

$$
\begin{aligned}
acc &= \frac{DD+SS+NN}{DD+DS+DN+SD+SS+SN+ND+NS+NN}, \\
dpv &= \frac{DD}{DD+DS+DN}, \\
cpv &= \frac{DD+DS+SD+SS}{DD+DS+DN+SD+SS+SN}, \\
npv &= \frac{NN}{ND+NS+NN}.
\end{aligned}
$$

In the same manner, as for binary segmentation, we calculate the number of annotation labels corresponding to all three predicted classes:

$$
\begin{aligned}
N_0 &= |\{i : \hat{p}_{h_i} = 0\}|, \\
\hat{N}_0 &= |\{i : \hat{p}_{h_i} = 0\}|, \\
N_1 &= |\{i : \hat{p}_{h_i} = 1\}|, \\
\hat{N}_1 &= |\{i : \hat{p}_{h_i} = 1\}|, \\
N_2 &= |\{i : \hat{p}_{h_i} = 2\}|. \\
\hat{N}_2 &= |\{i : \hat{p}_{h_i} = 2\}|.
\end{aligned}
$$

Based on such characteristics, we can calculate the rate of dense crowd detection equalling $N_2/\hat{N}_2$.

E. Crowd Dense Semantics

After obtaining and refining the segmentation maps (Fig. 3a), we clusterize the people in crowd images based on their positions provided as annotations for the considered dataset. For this, we divide all annotation points according to the corresponding connected areas of resultant segmentations (Fig. 3b) and clusterize each group separately (Fig. 3c). We don't clusterize points falling into non-crowded areas. Any clustering method for 2D points can be used. We decided to use DBSCAN as it demonstrated its benefit in various researches in logistics, spatial analysis, and behavior patterns detection [41], [42].

After the clustering, all points are divided into multiple clusters represented as convex hulls of the points inside them (Fig. 3c). Each cluster is characterized by its location, density, people count, as well as spatial connections

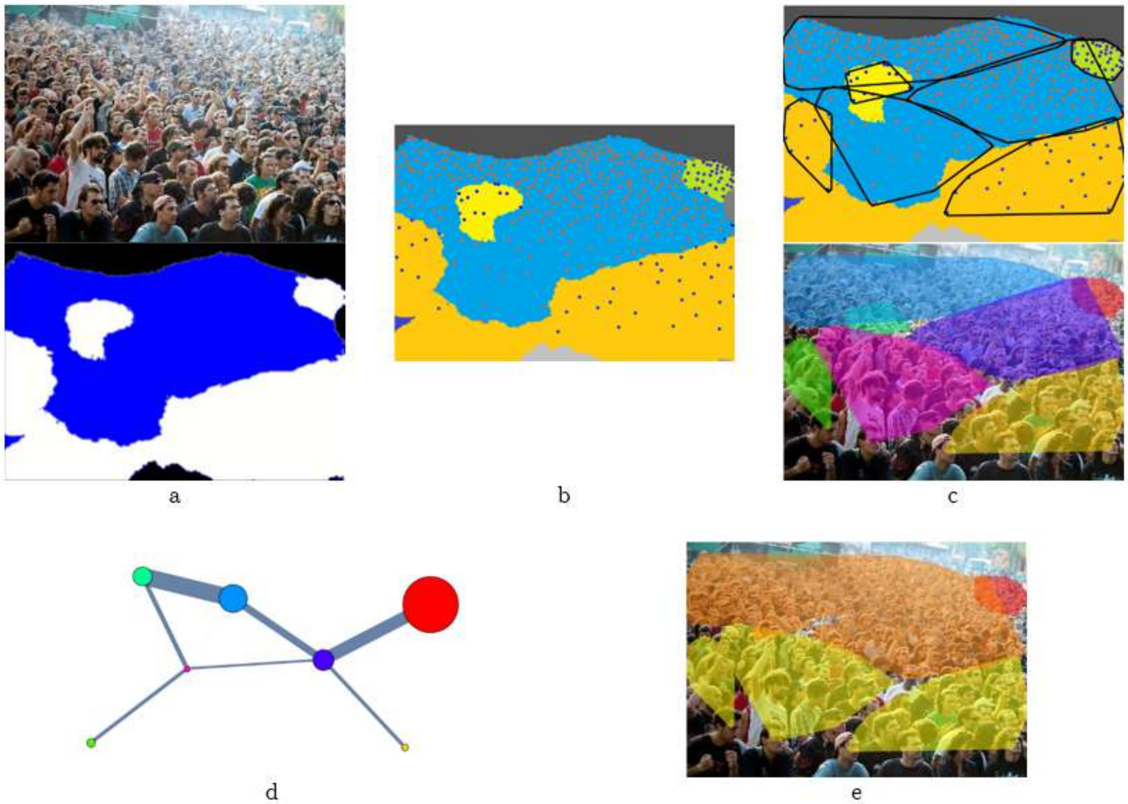


Figure 3. Semantics in a crowd: the initial image and the segmentation map for it (a), connected areas in the segmentation map (b), clusterization of annotation labels within connected areas (c), the clusters connectivity graph (d), crowd semantic clusters (e)

with other clusters. All clusters and such connections are presented as the graph where each vertex is assigned to a cluster, and two vertices are adjacent if only two clusters overlap or share a border (Fig. 3d). The size of a vertex is proportional to the density of the crowd in the corresponding cluster, and the thickness of each edge is proportional to the area shared between two adjacent clusters. Based on the separate clusters' properties and the graph's connectivity, we can interpret semantics for the depicted crowd in the image. In Fig. 3e, the red cluster is a dense crowd (more than 1 individual per 1000 pixels), the orange one is a regular crowd (0.5-1.5 individuals), and the yellow clusters present a sparse crowd (less than 0.7 individuals). Different clusters can share their borders, overlap, or even be a part of another cluster. Considering such facts, we can evaluate the crowd semantics.

IV. Results and Discussion

A. Segmentation Evaluation

After training neural networks and obtaining predicted segmentation maps for images in testing samples (Fig. 1, 3a), we received the results presented in Table I. Statistics on calculated accuracies and predictive values through the testing samples (from ShanghaiTech dataset's part B for binary segmentation and part A for ternary segmentation) are presented there.

As we can see, the CRF refinement significantly improves the ternary segmentations in terms of overall quality (acc.) and dense crowd detection (DPV). In other cases, it didn't demonstrate better results. Comparing dice and focal loss functions' results, we can conclude that the neural network with focal loss can predict crowd regions slightly better (according to the CPV). Nevertheless, dice loss allows us to predict dense crowds slightly better (DPV). Also, binary segmentation gave better results than ternary one in terms of overall performance. However, it takes place due to its ability to detect non-crowded areas (NPV) whereas the ternary segmentation model is more successful in detecting crowded areas (CPV).

Statistics on crowd detection rates are presented in Table II. Here we can see the poor performance of the model using the focal loss function in detecting sparse crowds (low values of $N_1/\hat{N}_1$), but for dense crowds detection, both dice and focal functions results are approximately equal (dice function demonstrates slightly better results though). On average, the model is prone to overlook some parts of sparse and dense crowds (from 3 % to 22% according to average and median results).

Table I
Accuracy and Predictive Values

	UNet Segmentation				CRF Segmentation			
	Acc.	CPV	DPV	NPV	Acc.	CPV	DPV	NPV
Binary Segmentation, Dice Lose								
Minimal	0.2629	0.0108	-	0.3144	0.2556	0.0086	-	0.3141
Average	0.7427	0.3962	-	0.8984	0.7440	0.3978	-	0.8987
Median	0.7580	0.3933	-	0.9464	0.7589	0.3939	-	0.9479
Maximal	0.9243	0.8968	-	1	0.9217	0.9001	-	1
Binary Segmentation, Focal Lose								
Minimal	0.3379	0	-	0.3330	0.3389	0	-	0.3322
Average	0.8308	0.4023	-	0.8513	0.8098	0.4185	-	0.8292
Median	0.8671	0.3960	-	0.8891	0.8412	0.4110	-	0.8635
Maximal	0.9785	0.9897	-	0.9965	0.9780	1	-	0.9950
Ternary Segmentation, Dice Lose								
Minimal	0.0685	0.1310	0	0	0.0622	0.1271	0	0
Average	**0.3175**	0.6830	**0.3746**	0.5127	**0.4498**	0.6819	**0.5169**	0.5132
Median	**0.3067**	0.7412	**0.3245**	0.5017	**0.4386**	0.7447	**0.5616**	0.5132
Maximal	**0.7331**	0.9992	0.9992	1	**0.9168**	0.9983	0.9947	1
Ternary Segmentation, Focal Lose								
Minimal	0.0053	0.0466	0	0	0.0534	0.0899	0.0002	0
Average	**0.3178**	0.7028	**0.2437**	0.4929	**0.5026**	0.7014	**0.4890**	0.4765
Median	**0.3012**	0.8016	**0.0771**	0.4913	**0.4952**	0.7785	**0.4911**	0.4693
Maximal	**0.7775**	0.9998	1	1	**0.9381**	1	0.9911	1

Sometimes the rates $N_1/\hat{N}_1$ and $N_2/\hat{N}_2$ are greater than 1. It may indicate situations where separate persons are detected as crowds and sparse crowds are recognized as dense crowd regions. According to our calculations, from 15% to 25% images have such values. Hence, the model might both under- and overestimate the crowd density. Also, data from Table II prove again using the CRF is crucial for ternary segmentation prediction.

Table II
Crowd Detection Rates

	UNet segmentation		CRF segmentation	
	Dice	Focal	Dice	Focal
Binary Segmentation ($N_1/\hat{N}_1$)				
Minimal	0.0097	0	0	0
Average	0.7896	0.1719	0.7909	0.1687
Median	0.8454	0.1189	0.8473	0.0872
Maximal	1.3235	0.8868	1.3235	0.9057
Ternary Segmentation ($N_2/\hat{N}_2$)				
Minimal	0	0	0.0402	0
Average	**0.5298**	**0.0199**	**0.8785**	**0.7896**
Median	**0.5215**	**0**	**0.9704**	**0.8517**
Maximal	1.2971	1	1.2971	1.2971

B. Crowd Semantics

Most of the ShanghaiTech datasets images are captured by a camera observing a nearby scene from above. Hence, the typical clusterization consists of distant dense clusters, closer regular clusters, and near sparse clusters (Fig. 4a). However, the crowd in an image can be divided vertically if there is a tall object like a pole or a flag (Fig. 4b).

Sometimes the pattern doesn't hold, which could indicate a group with interest. Some examples where we can detect a people's interest include:

- Multiple clusters with equal density spanning most of the image (Fig. 4c). Those usually present a uniform crowd with regular attention.
- A small cluster within or near a bigger one of the different type (Fig. 4, d-e). Those situations usually present a concentration of interest in particular groups within or near the crowd.
- A significant overlapping between clusters of different types. (Fig. 4f). This one can indicate a spreading interest or joining the people groups. Real-time surveillance systems must detect such actions to prevent any dire situations.
- Elongated clusters presenting regulate or dense crowds may indicate the presence of a queue in the region (Fig. 4g, the bottom orange cluster). If the density is high enough, some extraordinary situations might take place like queue crushes or evacuation panic which must be dealt with immediately.
- A wide sparse cluster at the bottom of the image might indicate a group of people that is very close to the observer (Fig. 4, d, h). Overlapping between the close cluster and other, distant ones is another feature of such a situation. Depending on the people's behavior, such a close group might be considered an outlier or an interest group, especially when it grows or approaches the observer.

V. Conclusions

This paper presents an approach for semantic segmentation of dense and sparse crowd images, addressing the critical

Figure 4. Semantic segmentation of crowds of different types: regular crowds with no attention (a, b), uniform crowd with regular attention (c), diverse crowd containing groups with increased interest (d, e), diverse crowd with a spreading group with interest (f), crowd with a queue (g), crowd with a close cluster (h)

need for accurate crowd analysis in various applications such as crowd management, surveillance, and urban planning. Our proposed method leverages a combination of UNet and CRF networks, augmented by a semi-automatic labeling technique based on Gaussian blur and thresholding methods to generate ground truth maps. Furthermore, we highlight some typical crowd behavior patterns based on clustering the people groups by their density and interconnections between them. Indicating those patterns is important for understanding crowd structures and dynamics as well as establishing crowd management and safety.

Through extensive experimentation and evaluation, we have demonstrated the effectiveness of our approach in accurately segmenting crowd images, particularly in binary segmentation tasks distinguishing crowded from non-crowded regions. While our model excels in binary segmentation, we acknowledge the challenges encountered in ternary segmentation tasks involving dense crowds, sparse crowds, and non-crowded areas. Despite this, our model shows promising results in crowd detection regardless of crowd density. Besides, we prove the necessity of CRF refinement to get better results in ternary segmentation.

References

[1] K. Khan et al., "Crowd Counting Using End-to-End Semantic Image Segmentation," *Electronics*, 2021, vol. 10, no. 11, #1293, doi: 10.3390/electronics10111293.

[2] L. Liu, Z. Qiu, G. Li, S. Liu, W. Ouyang, L. Lin, "Crowd counting with deep structured scale integration network," in *2019 IEEE International Conference on Computer Vision (CVF)*, pp. 1774-1783.

[3] S. Sholtanyuk, A. Leunikau, "Lightweight Deep Neural Networks for Dense Crowd Counting Estimation," in *Pattern Recognition and Information Processing (PRIP'2021).* United Institute of Informatics Problems of the National Academy of Sciences of Belarus, Minsk, 2021, pp. 61–64.

[4] S. Sholtanyuk, "Finding The Optimal Segmentation of a Crowd Image with Watershed Method," in *Information Systems and Technologies (CSIST'22).* Part 2, Belarusian State University, Minsk, 2022, pp. 217-223. Available at: https://elib.bsu.by/handle/123456789/288544.

[5] F. Abdullah, A. Jalal, "Semantic Segmentation Based Crowd Tracking and Anomaly Detection via Neuro-Fuzzy Classifier in Smart Surveillance System". *Arabian Journal for Science and Engineering*, 2023, vol. 48, no. 2, pp. 2173-2190.

[6] M. Gruosso, N. Capece, U. Erra, "Human Segmentation in Surveillance Video with Deep Learning", *Multimedia Tools and Applications*, 2021, vol. 80, no. 1, pp. 1175-1199.

[7] A. Kroshchanka, E. Mikhno, M. Kovalev, V. Zahariev, A. Zagorskij, "Semantic Analysis of the Video Stream Based on Neuro-Symbolic Artificial Intelligence," in *Open Semantic Technologies for Intelligent Systems*

(OSTIS-2021). Belarusian State University of Informatics and Radioelectronics, Minsk, 2021, pp. 193-204.

[8] A. Kroshchanka, V. Golovko, E. Mikhno, M. Kovalev, V. Zahariev, and A. Zagorskij, "A Neural-Symbolic Approach to Computer Vision," in *Open Semantic Technologies for Intelligent Systems*, eds.: V. Golenkov, V. Krasnoproshin, V. Golovko, and D. Shunkevich, Cham: Springer International Publishing, 2022, pp. 282–309, doi: 10.1007/978-3-031-15882-7_15.

[9] L. Greco, P. Ritrovato, M. Vento, "On the use of semantic technologies for video analytics," *J Ambient Intell Human Comput*, 2021, vol. 12, pp. 567–587, doi: 10.1007/s12652-020-02021-y.

[10] P. Anderson, B. Fernando, M. Johnson, S. Gould, "SPICE: Semantic Propositional Image Caption Evaluation," in *Computer Vision–ECCV 2016*, eds.: B. Leibe, J. Matas, N. Sebe, M. Welling, Cham: Springer International Publishing, 2016, pp. 382-398, doi: 10.1007/978-3-319-46454-1_24.

[11] M. H. T. De Boer, Y.-J. Lu, H. Zhang, K. Schutte, C.-W. Ngo, W. Kraaij, "Semantic Reasoning in Zero Example Video Event Retrieval," *ACM Trans. Multimedia Comput. Commun. Appl.*, 2017, vol. 13, no. 4, Article 60, 17 p., doi: 10.1145/3131288.

[12] Z. Feng, Z. Zeng, C. Guo, Z. Li, "Exploiting Visual Semantic Reasoning for Video-Text Retrieval," *arXiv preprint, arXiv:2006.08889,* 2020.

[13] S. Munir, S.I. Jami, S. Wasi, "Towards the Modelling of Veillance based Citizen Profiling using Knowledge Graphs," *Open Computer Science*, 2021, vol. 11, no. 1, pp. 294-304, doi: 10.1515/comp-2020-0209.

[14] X. Guo, M. Gao, G. Zou, A. Bruno, A. Chehri, G. Jeon, "Object Counting via Group and Graph Attention Network," in *IEEE Transactions on Neural Networks and Learning Systems*, 2023, pp. 1-12, doi: 10.1109/TNNLS.2023.3336894.

[15] L. Greco, P. Ritrovato, A. Saggese, M. Vento, "Improving reliability of people tracking by adding semantic reasoning," in *IEEE conference on advanced video and signal based surveillance (AVSS)*, IEEE, 2016, pp. 194-199.

[16] K. Humphrey, G. Underwood, G., "Domain knowledge moderates the influence of visual saliency in scene recognition," *British Journal of Psychology*, 2009, vol. 100, no. 2, pp. 377-398, doi: 10.1348/000712608X344780.

[17] B. Chen, Z. Yan, K. Li, P. Li, B. Wang, W. Zuo, L. Zhang, "Variational Attention: Propagating Domain-Specific Knowledge for Multi-Domain Learning in Crowd Counting," in *IEEE/CVF International Conference on Computer Vision*, 2021, pp. 16065-16075.

[18] V. A. Sindagi, V. M. Patel, "HA-CCN: Hierarchical Attention-Based Crowd Counting Network," in *IEEE Transactions on Image Processing*, 2020, vol. 29, pp. 323-335, doi: 10.1109/TIP.2019.2928634.

[19] W. Liu, M. Salzmann, P. Fua, "Context-Aware Crowd Counting," in *IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 5099-5108.

[20] J. Wang, Z. Chen, Y. Wu, "Action recognition with multiscale spatio-temporal contexts," in *CVPR 2011*. IEEE, June 2011, pp. 3185-3192.

[21] O. Ronneberger, P. Fischer, T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015, pp. 234-241.

[22] V. Badrinarayanan, A. Kendall, R. Cipolla, "SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017, vol. 39, no. 12, pp. 2481-2495.

[23] F.I. Diakogiannis, F. Waldner, P. Caccetta, C. Wu, "ResUNet-a: a Deep Learning Framework for Semantic Segmentation of Remotely Sensed Data," *ISPRS Journal of Photogrammetry and Remote Sensing*, 2020, vol. 162, pp. 94-114.

[24] J. Long, E. Shelhamer, T. Darrell, "Fully Convolutional Networks for Semantic Segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3431-3440.

[25] P. Arbeláez, B. Hariharan, C. Gu, S. Gupta, L. Bourdev, J. Malik, "Semantic Segmentation Using Regions and Parts," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, June 2012, pp. 3378-3385.

[26] R. Girshick, J. Donahue, T. Darrell, J. Malik, "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 580-587.

[27] K. Simonyan, A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *arXiv preprint, arXiv:1409.1556*, 2014.

[28] P. Krähenbühl, V. Koltun, "Efficient Inference in Fully Connected CRFs with Gaussian Edge Potentials," *Advances in Neural Information Processing Systems*, 2011, vol. 24, pp. 109-117.

[29] X. He, R.S. Zemel, M.A. Carreira-Perpinán, "Multiscale Conditional Random Fields for Image Labeling," in *2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2004, doi: 10.1109/CVPR.2004.1315232.

[30] N. Dalal, B. Triggs, "Histograms of Oriented Gradients for Human Detection," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, 2005, vol. 1, pp. 886-893.

[31] D. Helbing, P. Molnár, "Social force model for pedestrian dynamics," *Physical Review E*, 1995, vol. 51, no. 5, pp. 4282-4286.

[32] Y. Zhang, D. Zhou, S. Chen, S. Gao, Y. Ma, "Single-Image Crowd Counting via Multi-Column Convolutional Neural Network," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 589-597.

[33] M. Tan, Q. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," in *International conference on machine learning*. PMLR, May 2019, pp. 6105-6114.

[34] S. He et al., "An Image Inpainting-Based Data Augmentation Method for Improved Sclerosed Glomerular Identification Performance with The Segmentation Model EfficientNetB3-UNet," *Scientific Reports*, 2024, vol. 14, no. 1, #1033.

[35] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, L.C. Chen, "MobileNetV2: Inverted Residuals and Linear Bottlenecks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4510-4520.

[36] P. Iakubovskii, "Segmentation Models," *GitHub repository*. Available at: https://github.com/qubvel/segmentation_models (accessed 2024, Mar)

[37] S. Jadon, "A Survey of Loss Functions for Semantic Segmentation," in *2020 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*. IEEE, Oct. 2020, pp. 1-7.

[38] F. Milletari, N. Navab, S.A. Ahmadi, "V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation," in *2016 International Conference on 3D Vision (3DV)*. IEEE, Oct. 2016, pp. 565-571.

[39] M.S. Hossain, J.M. Betts, A.P. Paplinski, "Dual Focal Loss to Address Class Imbalance in Semantic Segmentation," *Neurocomputing*, 2021, vol. 462, pp. 69-87.

[40] L. Beyer, "PyDenseCRF," *GitHub repository*. Available at: https://github.com/lucasb-eyer/pydensecrf (accessed 2024, Mar).

[41] M. Chen, S. Banitaan, M. Maleki, Y. Li, "Pedestrian Group Detection with K-Means and DBSCAN Clustering Methods," in *2022 IEEE International Conference on Electro Information Technology (eIT)*. Mankato, MN, USA, 2022, pp. 1-6, doi: 10.1109/eIT53891.2022.9813918.

[42] A. Boumchich, J. Picaut, E. Bocher, "Using a Clustering Method to Detect Spatial Events in a Smartphone-Based Crowd-Sourced Database for Environmental Noise Assessment," *Sensors*, 2022, vol. 22, #8832, doi: 10.3390/s22228832.

АВТОМАТИЗАЦИЯ ОЦЕНКИ ВНИМАНИЯ СКОПЛЕНИЙ ЛЮДЕЙ НА ОСНОВЕ ПОЛУАВТОМАТИЧЕСКОЙ СЕМАНТИЧЕСКОЙ СЕГМЕНТАЦИИ ИЗОБРАЖЕНИЙ С ИСПОЛЬЗОВАНИЕМ СЕТЕЙ UNET И CRF

Шолтанюк С. В., Малёнкин Я. О.,
Лэй Б., Недзьведь А. М.

Семантическая сегментация изображений скоплений людей играет ключевую роль в различных приложениях, таких как управление толпой, наблюдение и городское планирование. В данной статье предложен подход к семантической сегментаций изображений с плотной и разреженной толпой на основе полуавтоматической разметки, использующий комбинацию UNet и условных случайных полей (CRF).

Представлена методика генерации карт сегментации для изображений скоплений людей. Сеть UNet используется для первоначальной, грубой сегментации, после которой следует её уточнение с использованием CRF. Результаты экспериментов показали, что модель лучше выполняет бинарную сегментацию (области, занятые толпой, и области, свободные от толпы), нежели тернарную сегментацию (области плотной толпы, разреженной толпы, и области, свободные от толпы). Однако, в ходе тернарной сегментации получились лучшие результаты по сегментации толпы в целом (без учёта типа толпы). Кроме того, показана значимость уточнения сегментации при помощи CRF в задаче тернарной сегментации толпы.

Также на основе предложенной модели сегментации выделены некоторые закономерности поведения скоплений людей. Они различаются по типу внимания людей, связями внутри скоплений людей и между ними, а также вероятностью возникновения чрезвычайных ситуаций.

Received 25.03.2024

Clustering of Multispectral Images by Terrain Classes Using a Semantic Approach

Xi Zhou
Nanjing Research Institute of Electronics Engineering
Nanjing, China.
zhouxi_nju@126.com

Artsiom Nedzved
Belarusian State University
Minsk, Belarus
artiom.nedzved@gmail.com

Alexei Belotserkovsky
United Institute of Informatics Problems of the NAS of Belarus
Minsk, Belarus
alex.belot@gmail.com

***Abstract*—This article conducted a study on methods for clustering multispectral images by terrain classes using semantic technologies. Using Word2Vek technology, a semantic form of the image is constructed, which is used to determine the class of the image.**

***Keywords*—multispectral image, covariance, vector difference, Word2vec**

I. Introduction

Semantic scene segmentation based on images is one of the most important tasks in computer vision. Although we have made enormous progress in recent years using sophisticated image descriptors [1], [2] and more advanced machine learning techniques [3], [4], segmentation remains a challenging task. Although humans can easily interpret images semantically, this remains challenging for computer vision systems, primarily due to the ambiguity of the effect of light and surface reflections on a given pixel value. For example, dark pixels may be the result of reflections from dark surfaces under normal lighting conditions or reflections from light surfaces in shadows. Deciphering the contribution of light and reflection to images is a challenging task [5]. To solve this problem, we either need to make assumptions about the world or get more information.

In this work, we explore semantic segmentation of multispectral images using the latter approach. In particular, multispectral images from the Sentinel-2 satellite will be used.

The spectral imaging is imaging using multiple bands in the electromagnetic spectrum. While a conventional camera captures light in three wavelength ranges in the visible spectrum, RGB, spectral imaging involves a wide range of techniques beyond RGB. Spectral imaging can use infrared, visible light, ultraviolet, x-rays, or any combination of the above. This may involve acquiring image data simultaneously in the visible and non-visible ranges, illuminating beyond the visible range or using optical filters to capture a specific spectral range, and the ability to capture hundreds of wavelength ranges for each image pixel.

Multispectral remote sensing includes visible, near-infrared, and short-wave infrared imaging. These images were obtained over several broad wavelength ranges. Thus, multispectral imaging captures image data in a specific range of wavelengths across the entire electromagnetic spectrum. Different materials caught in the frame reflect and absorb rays at different wavelengths differently. In remote sensing, materials can be distinguished by their spectral reflectance signatures observed in remote sensing (Earth remote sensing) images. In this case, it is very difficult to make a direct identification, as described in [6].

A higher level of spectral detail in multispectral images provides a better ability to recognize subtle differences. For example, multispectral remote sensing makes it possible to distinguish three minerals due to the high spectral resolution. At the same time, the multispectral Landsat Thematic Mapper system cannot distinguish between these three minerals.

II. Multispectral satellite imaging

One of the priority areas for processing multispectral information and deciphering remote sensing data is theoretical and applied research aimed at increasing the efficiency of multispectral information processing. In theoretical and practical terms, the creation of systems that support the information processing process requires the development of new and improvement of existing methods and algorithms for information analysis, as well as the development of special mathematical, algorithmic and software for information processing and decision-making systems, which is explained by the following reasons. Firstly, the algorithms used to decipher remote sensing data do not provide the required accuracy and reliability of the results. Secondly, the use of clustering algorithms for multispecies data is not qualitatively satisfactory for expert assessment by specifying reference areas. Thirdly, the development of fundamentally new clustering algorithms is often not effective compared to improving existing algorithms, in terms of increasing processing speed and reducing the number of iterations.

In addition, measuring the degree of similarity of sensing objects is much simpler than forming feature descriptions.

One of the priority areas for processing multispectral information and deciphering remote sensing data is theoretical and applied research aimed at increasing the efficiency of multispectral information processing. In theoretical and practical terms, the creation of systems that support the information processing process requires the development of new and improvement of existing methods and algorithms for information analysis, as well as the development of special mathematical, algorithmic and software for information processing and decision-making systems, which is explained by the following reasons. Firstly, the algorithms used to decipher remote sensing data do not provide the required accuracy and reliability of the results. Secondly, the use of clustering algorithms for multispecies data is not qualitatively satisfactory for expert assessment by specifying reference areas. Thirdly, the development of fundamentally new clustering algorithms is often not effective compared to improving existing algorithms, in terms of increasing processing speed and reducing the number of iterations. In addition, measuring the degree of similarity of sensing objects is much simpler than forming feature descriptions.

In addition, remote sensing systems are currently being constantly improved, which makes it possible to obtain images of increasingly higher spectral and spatial resolution. There are systems that allow shooting in hundreds of spectral ranges. The spatial resolution of images is also constantly being improved. If in the 70s each pixel of a space image corresponded to 80 meters of the Earth, now there are images with a resolution of 1 meter, and sometimes better. Older techniques developed for lower-resolution imagery do not extract all the useful information contained in modern imagery. Therefore, there is a need for new methods for interpreting images of area objects that take into account the advantages of modern multispectral imaging.

As a rule, color satellite images are formed based on the absorption or reflection of certain radiation waves of the spectrum. As a result, in most cases, the color is formed based on filling the areas with different dyes with different concentrations. The color is formed by decomposing the absorption values of dye mixtures into the absorption values of individual spots, in this case, a simple decomposition according to the coordinates of color systems. Such decomposition does not allow to obtain a linear relationship between the dye concentration and absorption, which corresponds to the spectral line under monochromatic conditions. Most of the tasks of monitoring the Earth's surface are focused on polychromatic conditions, for which it is impossible to obtain accurate spectral values.

A. *Landsat-8*

One example of a multispectral sensor is Landsat-8. For example, Landsat-8 produces 11 images using the following bands:

1) Coastal aerosol (COASTAL AEROSOL) in range 1 (0.43-0.45 µm).
2) Blue (BLUE) in the range 2 (0.45-0.51 µm).
3) Green (GREEN) in the range 3 (0.53-0.59 µm).
4) Red (RED) in the range 4 (0.64-0.67 µm).
5) Near infrared (NIR) in the range 5 (0.85-0.88 microns).
6) Shortwave infrared 1 (SWIR 1) in the range 6 (1.57-1.65 µm).
7) Shortwave infrared 2 (SWIR 2) in the range 7 (2.11-2.29 µm).
8) Panchromatic (PANCHROMATIC) in the range 8 (0.50-0.68 microns).
9) Cirrus (CIRRUS) in the range 9 (1.36-1.38 µm).
10) Thermal Infrared 1 (TIRS 1) in the 10 range (10.60-11.19 µm).
11) Thermal Infrared 2 (TIRS 2) in the 11 range (11.50-12.51 µm).

Each band has a spatial resolution of 30 meters, with the exception of bands 8, 10 and 11. Band 8 has a spatial resolution of 15 meters, and bands 10 and 11 have a pixel size of 100 meters. In this case, there is no range between 0.88 and 1.36 µm because the atmosphere absorbs light at these wavelengths, as in Figure 1.

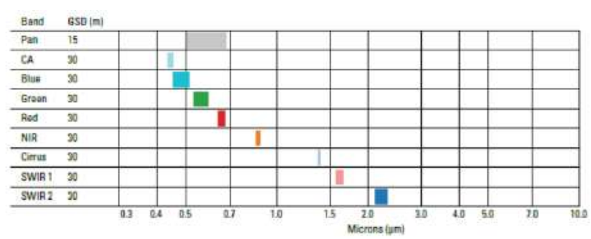


Figure 1. Ranges of spectral channels of Landsat satellites.

III. Description of the dataset

In [7] proposes a set of multispectral images captured by the Sentinel-2 satellite imagery. Sentinel-2 satellite imagery is publicly and freely available through the Copernicus Earth observation program. In [8], we present a new dataset based on Sentinel-2 satellite imagery, covering 13 spectral bands and consisting of 10 classes, containing a total of 27,000 tagged and georeferenced images. An example of one Industrial class image with the displayed spectral bands Red, Green, Blue can be seen in Figure 2. Of these, we have selected 8 of the most informative strips.

The dataset has the following classes:

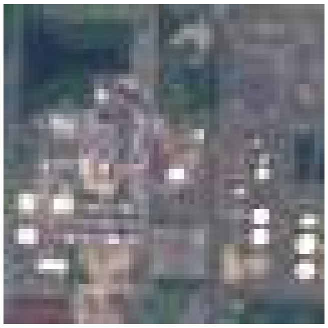


Figure 2. Industrial-1011 in RGB spectral bands.

A. Annual Crop class

The Annual Crop class is a satellite image of an agricultural area represented by fields of annual crops. A feature of the class is the presence of both large areas of immature vegetation and large areas of mature vegetation, brightly illuminated by infrared radiation.

B. Forest class

The Forest class is a collection of satellite images of forested areas. A feature of this class is the overwhelming presence of vegetation, brightly illuminated by infrared radiation.

C. HerbaceousVegetation class

The HerbaceousVegetation class represents satellite imagery of hilly or steep terrain. A feature of the class is the rare presence of vegetation in the image.

D. Highway class

The Highway class is satellite imagery of the area in which a major highway passes. They are characterized by the fact that after image processing the highway should be clearly visible.

E. Industrial class

The Industrial class presents satellite images of industrial areas. A special feature of the class is a large number of buildings and a small area of terrain with vegetation. Also, after processing, the outlines of buildings are often lost, but they can be restored to construct an outline using the difference in the images.

F. Industrial class

The Pasture class contains satellite images of flat terrain. A feature of the class is the abundance of vegetation, brightly illuminated by infrared radiation.

G. PermanentCrop class

The PermanentCrop class is a satellite image of an agricultural area represented by fields of permanent crops. A special feature of the class is the presence of large areas of mature vegetation, brightly illuminated by infrared radiation.

H. Residential class

The Residential class represents satellite images of populated areas. A special feature of the class is the large number of buildings throughout the image.

I. River class

The River class is satellite imagery of an area where there is a river. They are characterized by the fact that after image processing the river should clearly stand out.

J. SeaLake class

The SeaLake class contains satellite images of the seascape. They are characterized by the overwhelming presence of water.

IV. Calculating the covariance matrix

To classify the image class represented by 8 spectral channels, it is proposed to use an algorithm to calculate the covariance matrix, where each cell i,j will denote the covariance of the i-th spectral layer and the j-th layer, as in Figure 3.

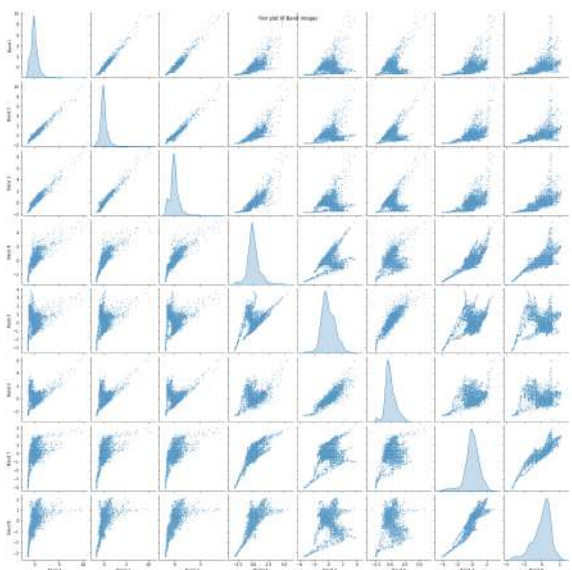


Figure 3. Visualization of the image covariance matrix with 8 spectral bands.

One such algorithm is the principle of principal components. He decides:

1) Approximate the data by linear manifolds of lower dimension: find a linear manifold of a given dimension $k < d$, the sum of squared distances to which is minimal.
2) Find a subspace of a given dimension, in the orthogonal projection onto which the spread (dispersion) (sample variance for $k = 1$) is maximum.
3) Find a subspace of a given dimension, in the orthogonal projection onto which the root-mean-square distance between each pair of points is maximum.

The principal component method consists of calculating eigenvectors and eigenvalues of the covariance matrix of the data space, then constructing projections in such a way that the direction of the maximum dispersion of the projection always coincides with the eigenvector having the maximum eigenvalue equal to the value of this dispersion. The covariance matrix after PCA processing can be seen in Figure 4.

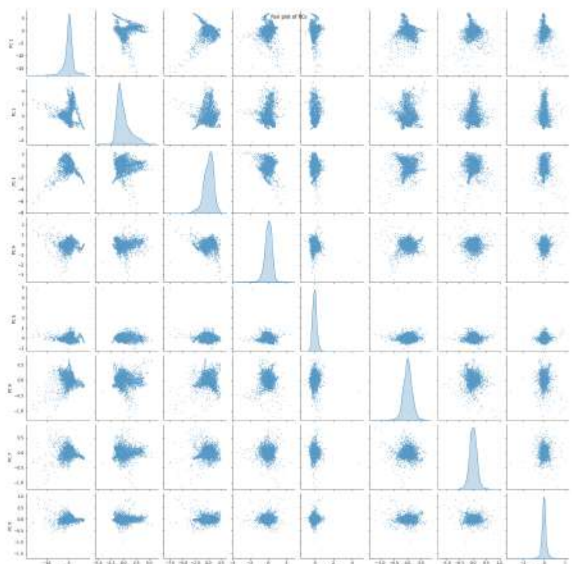


Figure 4. Visualization of the image covariance matrix with 8 spectral bands.

The next step of the algorithm is to reduce the dimension of the data space, but in the case of multispectral images this is not a necessary step: each projection is a new image layer that stores the necessary data. Instead, in [9] it is proposed to work only with the most informative image from the resulting projections.

Different images of the same class will have almost identical spectral data ratios. Indeed, for Forest class images the frequencies of 560 nm will prevail. and 842 nm., when for the SeaLake class 490 nm. and 945 nm. As a consequence, since each cell of the covariance matrix denotes either the variance of some layer of the multispectral image if that cell lies on the diagonal, or the covariance of two specific layers if it does not lie on the diagonal, the covariance matrices for each class will have the same patterns of dominant relationships.

V. Construction of a semantic form of the area

If we consider the matrix as a vector, then using the Word2vec technology from [10], from the previous statements we can conclude that the image is converted into a word that has metric characteristics, as shown in Figure 5.

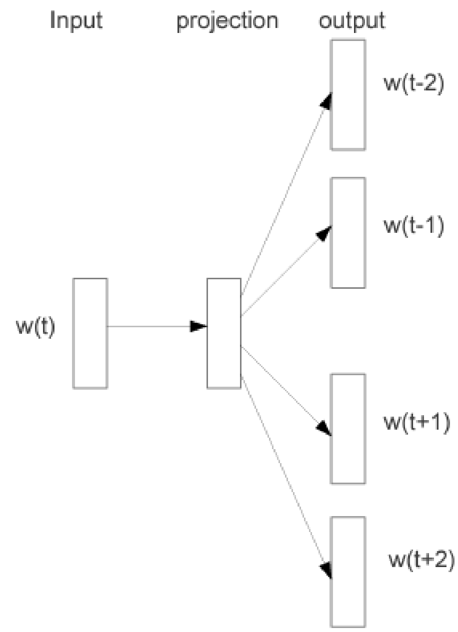


Figure 5. Illustration of how Word2vec technology works.

Following the Word2vec principle, we can build a semantic model to determine the semantic image of an image. Thus, the image is converted into a covariance ratio vector, where each cell uniquely defines the covariance of the two spectral layers. Consequently, if defining vectors are selected for each class, then by the difference between the vectors of the vector space, the dimension of which is n*n, where n is the number of spectral bands in the image, it will be possible to predict which class the image belongs to by the semantic difference of the vectors.

Using this semantic approach, we convert the image into a covariance ratio vector. Each element in this vector encapsulates the covariance between the two spectral layers, encoding not only spectral information but also semantic nuances. By selecting definition vectors for each terrain category, we create a semantic reference frame. Subsequently, using the semantic differences between these vectors within a vector space (which is expanded according to the number of spectral bands in the image), images can be accurately classified based on their semantic properties. By leveraging semantic techniques, we overcome the limitations of traditional pixel-based analysis and gain a deeper understanding of the semantic landscape of multispectral images.

VI. Demonstration of the method

To demonstrate how the method works, consider matrices for three different classes: Industrial, Forest, SeaLake. All three classes have different spectral characteristics that will uniquely determine the covariance matrix for the images representing the class. By calculating the average value of the covariance matrix for classes whose sample included 1000 images, as well as the vector of eigenvalues of these matrices.

A. Calculation of class matrices

Matrix for class Industrial:
[[1 0.981 0.971 0.839 0.328 0.33 0.583 0.676]
[0.981 1. 0.976 0.876 0.414 0.426 0.64 0.696]
[0.971 0.976 1. 0.877 0.347 0.348 0.653 0.738]
[0.839 0.876 0.877 1. 0.549 0.45 0.823 0.823]
[0.328 0.414 0.347 0.549 1. 0.886 0.59 0.323]
[0.33 0.426 0.348 0.45 0.886 1. 0.504 0.255]
[0.583 0.64 0.653 0.823 0.59 0.504 1. 0.903]
[0.676 0.696 0.738 0.823 0.323 0.255 0.903 1.]]

Eigenvalues:
[5.566 1.449 0.706 0.145 0.059 0.012 0.024 0.041]

Matrix for class Forest:
[[1. 0.878 0.907 0.853 0.724 0.75 0.841 0.854]
[0.878 1. 0.907 0.888 0.805 0.856 0.861 0.863]
[0.907 0.907 1. 0.917 0.738 0.731 0.882 0.907]
[0.853 0.888 0.917 1. 0.888 0.793 0.972 0.976]
[0.724 0.805 0.738 0.888 1. 0.898 0.901 0.857]
[0.75 0.856 0.731 0.793 0.898 1. 0.794 0.755]
[0.841 0.861 0.882 0.972 0.901 0.794 1. 0.992]
[0.854 0.863 0.907 0.976 0.857 0.755 0.992 1.]]

Eigenvalues:
[5.568 1.259 0.627 0.218 0.015 0.066 0.136 0.112]

Matrix for class SeaLake:
[[1. 0.492 0.446 0.366 0.27 0.23 0.295 0.289]
[0.492 1. 0.477 0.408 0.32 0.26 0.329 0.296]
[0.446 0.477 1. 0.532 0.418 0.394 0.402 0.373]
[0.366 0.408 0.532 1. 0.583 0.552 0.537 0.501]
[0.27 0.32 0.418 0.583 1. 0.582 0.585 0.519]
[0.23 0.26 0.394 0.552 0.582 1. 0.54 0.482]
[0.295 0.329 0.402 0.537 0.585 0.54 1. 0.561]
[0.289 0.296 0.373 0.501 0.519 0.482 0.561 1.]]

Eigenvalues:
[4.047 1.147 0.589 0.506 0.492 0.383 0.413 0.425]

And also consider two images from the dataset, Industrial-1011, shown in Figure 2, and SeaLake-1016, shown in Figure 6.

B. Calculation of image matrices

The first proposed method for determining the class of an image will be the difference in the metrics of the matrix space L0. This method allows you to quickly and even visually determine whether an image belongs to a class. The main problem of this method is reducing the dimension of space from eight stripes to one number, as a result of which collisions arise when semantically different vectors return a measure that is close in value. Because of this, a significant number of errors arise when determining the class of an image.

Figure 6. SeaLake-1016 in RGB spectral bands.

The second method for determining the class membership of an image is the nearest neighbor search algorithm. This algorithm consists of three steps:

1) The distance between each image and the eigenvectors of each class is calculated. The distance is taken to be the quadratic difference of vectors.
2) Find the minimum distance for each image.
3) The class to which the image belongs is determined by comparing the minimum distances.

This method requires a little more calculations, but it takes into account the ratio of the spectral bands of the matrices in a certain order, as well as the difference between the corresponding bands.

The image Industrial-1011 obtained the following values of the covariance matrices and eigenvectors:

Matrix for class Industrial-1011:
[[1. 0.982 0.957 0.841 0.136 0.195 0.564 0.774]
[0.982 1. 0.975 0.873 0.214 0.275 0.604 0.778]
[0.957 0.975 1. 0.899 0.18 0.216 0.63 0.818]
[0.841 0.873 0.899 1. 0.383 0.321 0.816 0.929]
[0.136 0.214 0.18 0.383 1. 0.925 0.69 0.36]
[0.195 0.275 0.216 0.321 0.925 1. 0.608 0.302]
[0.564 0.604 0.63 0.816 0.69 0.608 1. 0.894]
[0.774 0.778 0.818 0.929 0.36 0.302 0.894 1.]]

Eigenvalues:
[5.47 1.862 0.487 0.089 0.042 0.032 0.006 0.014]

Based on the calculation results, image Industrial-1011, the quadratic distance between the image vector and the Industrial class vector is 0.482, the Forest class vector is 0.662, and the SeaLake class vector is 1.839. For clarity, distances are rounded to the third decimal place. The probability of an image belonging to the Industrial class is 84%, to the Forest class is 78%, and to the SeaLake class is 38%. Thus, we can conclude that the image belongs to the Industrial class, but it is worth noting the presence of local vegetation in the image.

The SeaLake-1016 image obtained the following values of covariance matrices and matrix space norms:

Matrix for class SeaLake-1016 :
[[1. 0.465 0.383 0.433 0.409 0.306 0.316 0.124]
[0.465 1. 0.56 0.621 0.605 0.459 0.477 0.137]
[0.383 0.56 1. 0.546 0.512 0.439 0.386 0.172]
[0.433 0.621 0.546 1. 0.647 0.51 0.481 0.202]
[0.409 0.605 0.512 0.647 1. 0.553 0.562 0.238]
[0.306 0.459 0.439 0.51 0.553 1. 0.42 0.174]
[0.316 0.477 0.386 0.481 0.562 0.42 1. 0.095]
[0.124 0.137 0.172 0.202 0.238 0.174 0.095 1.]]

Eigenvalues:
[3.981 0.955 0.746 0.611 0.553 0.453 0.371 0.331]

Based on the calculation results, image Industrial-1011, the quadratic distance between the image vector and the Industrial class vector is 1.902, the Forest class vector is 1.823, and the SeaLake class vector is 0.310. For clarity, distances are rounded to the third decimal place. The probability of an image belonging to the Industrial class is 47%, to the Forest class is 49%, and to the SeaLake class is 92%. Thus, we can conclude that the image belongs to the SeaLake class, and unambiguously.

Diagram of the algorithm

The general diagram of the algorithm is presented in Figure 7.

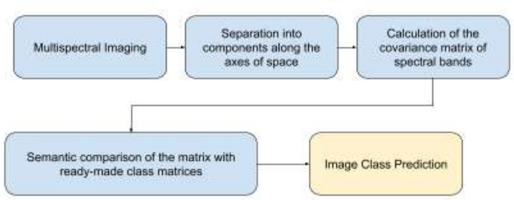


Figure 7. General diagram of the algorithm.

Acknowledgment

Financial support for the project "Agreement on the development of technology for developing algorithms for processing images of remote sensing of the Earth", agreement number: 22CETC19-ICN1785.

References

[1] J. Shotton, J.M. Winn, C. Rother, A. Criminisi, "TextonBoost: Joint Appearance, Shape and Context Modeling for Multi-class Object Recognition and Segmentation," A. Leonardis, H. Bischof, A. Pinz, ECCV 2006, Part I. LNCS, vol. 3951, pp. 1–15. Springer, Heidelberg, 2006.
[2] G. Csurka, F. Perronnin, "An efficient approach to semantic segmentation," IJCV 95, 2011.
[3] J. Verbeek, B. Triggs, "Region classification with markov field aspects models," CVPR, 2007.
[4] L. Ladicky, C. Russell, P. Kohli, P.H.S. Torr, "Graph Cut Based Inference with Co-occurrence Statistics," K. Daniilidis, P. Maragos, N. Paragios: ECCV 2010, Part V. LNCS, vol. 6315, pp. 239–253. Springer, Heidelberg, 2010.
[5] G.D. Finlayson, M.S. Drew, B.V. Funt, "Color constancy: generalized diagonal transforms suffice," Journal of the Optical Society of America 11, 3011–3019, 1994.
[6] G. Eason, B. Noble, and I. N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955.
[7] Patrick Helber, Benjamin Bischke, Andreas Dengel, Damian Borth. "Eurosat: a new dataset and deep learning benchmark for land use and land cover classification." IEEE Journal on Selected Topics in Applied Earth Observation and Remote Sensing, 2019.
[8] Patrick Helber, Benjamin Bischke, Andreas Dengel. "Introducing EuroSAT: a new dataset and deep learning benchmark for land use and land cover classification." 2018 IEEE International Symposium on Geosciences and Remote Sensing.
[9] Santosh Kumar R., "Principal Component Analysis: Drilling Insight through Image Visualization," 2020.
[10] Mikolov T., Chen K., Corrado G., Dean J. Efficient Estimation of Word Representations in Vector Space. In Proceedings of Workshop at ICLR, 2013.

ОПРЕДЕЛЕНИЕ КЛАССА МУЛЬТИСПЕКТРАЛЬНОГО ИЗОБРАЖЕНИЯ ПО СЕМАНТИЧЕСКОЙ РАЗНОСТИ КОВАРИАЦИОННЫХ МАТРИЦ

Бу Цин, Недзьведь А. А., Белоцерковский А.

Аннотация: В данной статье проведено исследование спутниковых мультиспектральных изображений, спектральных данных местности, а также представлен метод определения принадлежности изображений к классам местности с использованием семантического анализа вектора собственных значений ковариационной матрицы спутникового мультиспектрального изображения.

Received 27.03.2024

Thyroid Gland Ultrasonography Automation Through Intelligent Analysis

Alena Cherkas
Belarusian State University of Informatics and Radioelectronics
Minsk, Belarus
Email: alena_cherkas_13@mail.ru

***Abstract*—This article proposes an algorithm for automating the process of medical ultrasound diagnostics using intelligent analysis. The actions are described using the example of a thyroid gland study. Additional verification of the result by the artificial intelligence allows novice doctors to feel more confident and minimize the influence of the human factor on the quality of diagnosis.**

***Keywords*—Ultrasonography automation; thyroid gland ultrasonography; artificial intelligence in medicine; intelligent image analysis; neural networks; deep learning; convolutional neural network; OSTIS; OSTIS Technology integration;**

I. Introduction

Currently, thyroid problems are widespread in the population of the Republic of Belarus. This is due to the disaster at the Chernobyl nuclear power plant in 1986. The Gomel and Mogilev regions of the country were the most affected. In the first 10 days after the accident, the concentration of radioactive iodine was increased in some territories of the republic, which led to an increase in cases of thyroid pathology.

According to the Ministry of Health for 2021, 3.8% of the population of Belarus has pathology of this organ. There is an increase in the incidence every year.

Therefore, improving the technique of ultrasound diagnostics of thyroid pathologies is an urgent issue of our time.

Neural networks are gaining more and more popularity. They are often used in medical diagnostics. For example, to process test results, improve the quality of magnetic resonance imaging, analyze large amounts of data, and even perform surgical interventions.

By connecting artificial intelligence to the research, it is possible to reduce the influence of the human factor on the quality of the diagnosis. The result of ultrasound often depends on the doctor's experience. After all, this type of diagnosis involves processing the results directly at the time of the study. In this regard, there remains a high risk of missing an important feature of the organ structure.

II. Domain analysis

Diagnostic ultrasound is a safe, non-invasive diagnostic technique used to image inside the body. Ultrasound probes, called transducers, produce sound waves that have frequencies above the threshold of human hearing (above 20KHz), but most transducers in current use operate at much higher frequencies (in the megahertz (MHz) range).

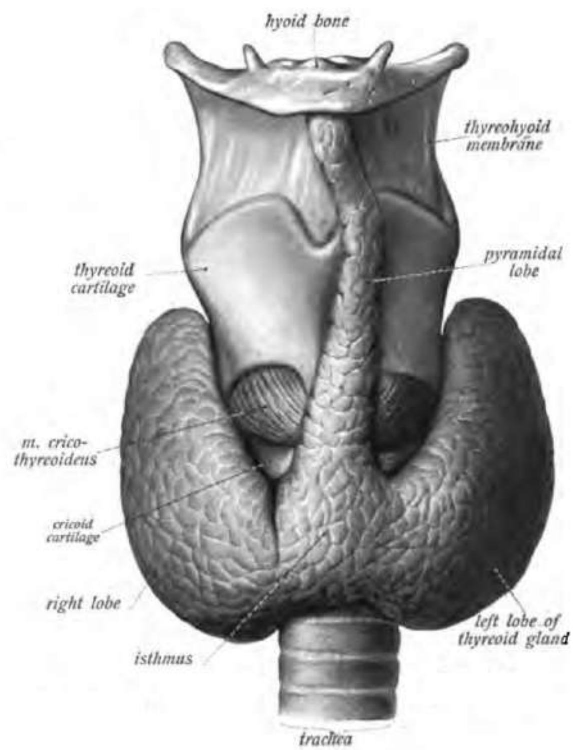


Figure 1. Thyroid gland scheme [1]

Ultrasound waves are produced by a transducer, which can both emit ultrasound waves, as well as detect the ultrasound echoes reflected back. In most cases, the active elements in ultrasound transducers are made of special ceramic crystal materials called piezoelectrics. These materials are able to produce sound waves when an electric field is applied to them, but can also work in reverse, producing an electric field when a sound wave hits them. When used in an ultrasound scanner, the transducer sends out a beam of sound waves into the body.

The sound waves are reflected back to the transducer by boundaries between tissues in the path of the beam (e.g. the boundary between fluid and soft tissue or tissue and bone). When these echoes hit the transducer, they generate electrical signals that are sent to the ultrasound scanner. Using the speed of sound and the time of each echo's return, the scanner calculates the distance from the transducer to the tissue boundary. These distances are then used to generate two-dimensional images of tissues and organs.

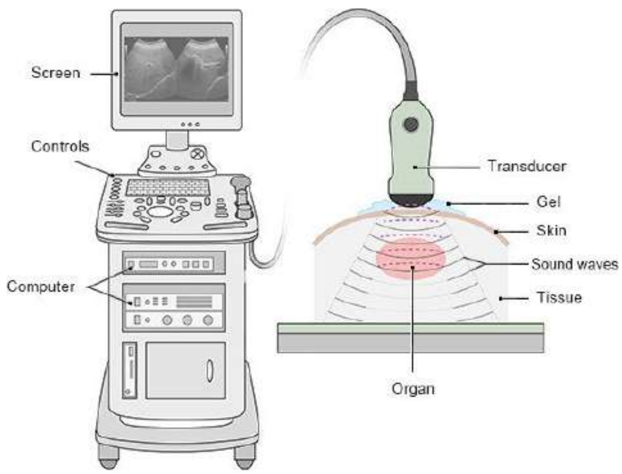


Figure 2. Ultrasound system scheme [2]

During an ultrasound exam, the technician will apply a gel to the skin. This keeps air pockets from forming between the transducer and the skin, which can block ultrasound waves from passing into the body. [3]

Ultrasound is a method of shadows. This study does not show a volumetric model of human organs, however, it allows to estimate the size, volume, density, correct location of the organ relative to the entire body, the presence of fluid in the area under study, as well as cysts, tumors and other formations. Moreover, minimally invasive operations are also performed under the control of an ultrasonic sensor. This allows the doctor to monitor the progress of the study without having to make large incisions on the human body.

During the ultrasound examination, the signal passes through the tissues of the human body and returns back. Solid dense organs reflect the sound signal well. Therefore, areas such as bones and stones look white on ultrasound.

Internal organs and soft tissues are usually represented by different shades of gray depending on the density of the organ.

Voids and liquid are shown in black on the screen, because in this case there are no obstacles in the signal path, and therefore it is practically not reflected at all.

Depending on the diagnostic areas, different types of sensors are used. The linear sensor has a rectangular image. The 2D sensor has a wide aperture, and its central frequency is in the range of 2.5-12 MHz (3D-4D is in the range of 7.5-11 MHz). Piezoelectric crystals in a linear sensor are located in the same plane, so such a sensor provides good visibility at close range. It is used for ultrasound of blood vessels, muscles, performing anesthesia under ultrasound control, examining mammary glands, thyroid gland and other superficial organs.

Figure 3. An example of an ultrasound made by a linear sensor [4]

In a convex sensor, piezoelectric crystals are arranged curvilinearly. Therefore, such a sensor visualizes deeply located structures well. The convexic 2D sensor has a wide aperture, and its central frequency is 2.5-7.5 MHz (3D, 4D — 3.5-6.5 MHz). With its help, ultrasound of the fetus, pelvic organs, and abdominal cavity is performed.

The sector phased array sensor is so named after the type of piezoelectric element device, which is called a phased array. The phased array sensor has a small aperture and a low frequency (the central frequency is 2-7.5 MHz). The shape of the scanning area is almost triangular. These sensors have poor resolution in the near field but give a good view at depth. They allow to observe structures through a narrow intercostal gap. With its help, ultrasound of the heart, abdominal organs, and brain is performed.

For ultrasound diagnosis of the thyroid gland, only a linear sensor is often used. However, it is possible to see a trapezoidal image on the screen of the device. This is due to the fact that the linear sensor has the function of a virtual convection, which allows to make the viewing plane wider and accommodate the entire organ there.

The convex sensor is used only when the thyroid gland is enlarged and the patient's body weight is too large.

Also ultrasound information can be displayed in several ways:

A-mode: As spikes on a graph (used to scan the eye).

B-mode: As a 2-dimensional anatomic images (used during pregnancy to evaluate the developing fetus or to evaluate internal organs).

Figure 4. An example of an ultrasound made by a linear sensor with virtual convex mode [5]

M-mode: As waves displayed continuously to show moving structures (used to evaluate the fetus's heartbeat or to evaluate heart valve disorders).

B-mode ultrasonography is most commonly done.

Sonography can be enhanced with Doppler measurements, which employ the Doppler effect to assess whether structures (usually blood) are moving towards or away from the probe, and its relative velocity. By calculating the frequency shift of a particular sample volume, for example a jet of blood flow over a heart valve, its speed and direction can be determined and visualised. This is particularly useful in cardiovascular studies (sonography of the vasculature system and heart) and essential in many areas such as determining reverse blood flow in the liver vasculature in portal hypertension. The Doppler information is displayed graphically using spectral Doppler, or as an image using color Doppler (directional Doppler) or power Doppler (non directional Doppler). This Doppler shift falls in the audible range and is often presented audibly using stereo speakers: this produces a very distinctive, although synthetic, pulsing sound.

Doppler ultrasonography uses changes that occur in the frequency of sound waves when they are reflected from a moving object (called the Doppler effect). In medical imaging, the moving objects are red blood cells in the blood. Thus, Doppler ultrasonography can be used to evaluate.

It is used to evaluate how well the heart is functioning (as part of echocardiography), to detect blocked blood vessels, especially in leg veins, as in deep vein thrombosis, when veins are blocked by a blood clot. To detect narrowed arteries, especially the carotid arteries in the neck, which carry blood to the brain.

Strictly speaking, most modern sonographic machines do not use the Doppler effect to measure velocity, as they rely on pulsed wave Doppler (PW). Pulsed wave machines transmit pulses of ultrasound, and then switch to receive mode. As such, the reflected pulse that they receive is not subject to a frequency shift, as the insonation is not continuous. However, by making several measurements, the phase change in subsequent measurements can be used to obtain the frequency shift (since frequency is the rate of change of phase). To obtain the phase shift between the received and transmitted signals, one of two algorithms is typically used: the Kasai algorithm or cross-correlation. Older machines, that use continuous wave (CW) Doppler, exhibit the Doppler effect as described above. To do this, they must have separate transmission and reception transducers. The major drawback of CW machines, is that no distance information can be obtained (this is the major advantage of PW systems - the time between the transmitted and received pulses can be converted into a distance with knowledge of the speed of sound).

In the sonographic community (although not in the signal processing community), the terminology "Doppler" ultrasound, has been accepted to apply to both PW and CW Doppler systems despite the different mechanisms by which the velocity is measured.

Spectral Doppler ultrasonography shows blood flow information as a graph. It can be used to assess how much of a blood vessel is blocked.

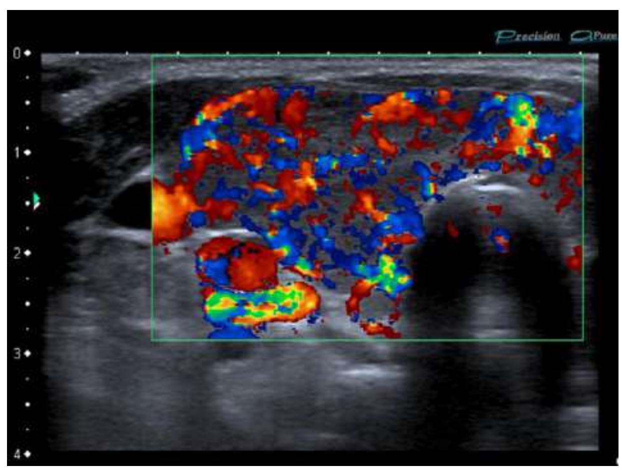


Figure 5. An example of thyroid dopplerography [6]

III. Overview of existing approaches

There are already a lot of scientific articles on the processing of ultrasound results using artificial intelligence. Some of them are even applied in practice. Moreover, there is a S-Detect (Samsung RS80A ultrasound system, Seoul, Korea). It is the first commercially available ultrasound CAD based on deep learning technology for thyroid imaging. [7]

S-Detect is a computer-aided detection (CAD) software developed by Samsung Electronics for use with their RS80A ultrasound system. It is designed to assist radiologists and clinicians in the detection and characterization of breast lesions during ultrasound examinations.

Figure 6. An example of Samsung S-Detect-system interface [8]

This system has a lot of pros: it provides real-time feedback to the clinician during the examination, enabling immediate assessment and decision-making regarding lesion characterization and management. More than that, the software provides standardized reporting templates that facilitate structured documentation of lesion characteristics, including malignancy probability scores and recommended management options. This promotes consistency and completeness in reporting. S-Detect seamlessly integrates with the RS80A ultrasound system's workflow, allowing for efficient and streamlined use within clinical practice. It is user-friendly and does not significantly disrupt the examination process.

But S-Detect is primarily designed for breast lesion characterization and may not be suitable for other types of lesions or organs. Its utility is limited to breast ultrasound examinations and may not address the full spectrum of diagnostic challenges encountered in clinical practice. While S-Detect is user-friendly, clinicians may require some time to familiarize themselves with the software's features and functionality. Training and ongoing education may be necessary to optimize its use and interpretation. Moreover, The implementation of S-Detect may incur additional costs associated with software licensing, training, and maintenance. Clinics and healthcare facilities must consider the financial implications before adopting the technology.

There are another people who have been automating ultrasound diagnostics of the thyroid gland using artificial intelligence. For example in 2021 scientists from Romania published their article "Intelligent Diagnosis of Thyroid Ultrasound Imaging Using an Ensemble of Deep Learning Methods".

They developed a CNN-VGG ensemble fused from two models: a pre-trained fined tuned model VGG-19 and an efficient lightweight CNN model. The proposed ensemble method proved to be an excellent and stable classifier with a good performance in terms of overall sensitivity (95.75%), specificity (98.43%), accuracy (97.35%), AUC (0.96), positive predictive value (95.41%) and negative predictive value (98.05%). [9]

Also there are scientists from China who published an "Artificial intelligence in thyroid ultrasound" article in 2023. Their research is more focused on the prevention and early detection of the thyroid cancer. They also used deep learning algorithms to achieve this goal. They tests different types of DL-based neural networks. [7]

The research of the above-mentioned scientists has been very successful. Their authors placed great emphasis on training the neural network to make diagnoses and look for pathology in ultrasound diagnostic images.

In the current work, a simpler and more global approach is considered: the neural network does not diagnose, but only assists the doctor. Their joint work makes it possible to minimize the errors of both the doctor and the software. The approach is described using the example of thyroid gland examination, but it can also be used in ultrasound diagnostics of other organs.

Also in this article, it is proposed to analyze not individual images, but a video recording of the entire research process.

Based on the approach described below, it is planned to develop a software product in the future and implement it into the work of a medical institution in a test mode.

Also, in the future, it is planned to develop the idea in such a way as to process not the final product of the work of some software: a visual representation of the ultrasound process, but the initial product, that is, ultrasonic signals. This will make the processing process faster.

IV. Proposed approach

Currently, artificial intelligence has been used in medicine for a long time. Integrating artificial intelligence into the ultrasound diagnostic process is not the easiest task. After all, software needs time to analyze. Many studies allow to process the result later: for example, MRI, X-ray and others. But a standard ultrasound examination involves the interpretation of the result by a doctor right at the time of the study. In this regard, the quality of the study directly depends on the experience and attentiveness of the doctor.

To do intelligent processing of MRI results, it is enough to simply install the appropriate program on your computer. Because the MRI is first fully performed and then interpreted. And due to the fact that the ultrasound examination is simultaneously performed and interpreted, a third-party computer is rarely used by a doctor for it. But connecting the software directly to the ultrasound machine is almost impossible, for two reasons. Firstly, devices from different manufacturers with different software are used for diagnostics, which is written in low-level languages and can be difficult to integrate with other more modern technologies. Secondly, as mentioned earlier, the software needs time to process the data.

After numerous consultations with specialists in the medical field, analyzing the situation and finding the best way to introduce artificial intelligence into the ultrasound diagnostic process, it was decided to record the research process in a video format file. Then the data is transferred to the computer. The video is divided into frames of 0.5 seconds of research. It is this time interval that will allow not to process the same images several times, but at the same time not to miss important changes. The frames are then processed by a neural network. At the end of processing, the software generates its own, it will highlight a suspicious area and comment on it. In this case, the doctor can either ignore the prompts of artificial intelligence, if he has already paid attention to this pathology, or put a sensor and review the moment of interest again.

Training a neural network for the automated analysis of thyroid gland ultrasonography images involves several key steps.

The first step is to gather a large dataset of thyroid ultrasound images. These images should cover a wide range of thyroid conditions, including cysts, tumors, nodules, and other pathologies. The dataset should be diverse and representative of the population being analyzed.

Once the dataset is collected, it needs to be preprocessed to ensure consistency and quality. This may involve resizing the images, standardizing the brightness and contrast, and removing noise or artifacts. Each image in the dataset needs to be labeled with the corresponding thyroid pathology, such as cyst, tumor, or normal. This step is crucial for supervised learning, where the neural network learns from labeled examples.

Then it is need to choose an appropriate neural network architecture for the task. Convolutional Neural Networks (CNNs) are commonly used for image classification tasks due to their ability to capture spatial hierarchies in data.

A Convolutional Neural Network (CNN) is a type of deep learning algorithm specifically designed for processing and analyzing visual data, such as images. It is inspired by the structure and function of the human visual cortex and is well-suited for tasks such as image classification, object detection, and image segmentation.

CNNs can not only classify images but also localize the regions within the image that contain abnormalities. This is crucial in medical imaging tasks, as it allows clinicians to pinpoint the location of cysts or tumors within the thyroid gland. CNNs can be trained to output bounding boxes or segmentation masks that delineate the boundaries of detected abnormalities.

CNNs consist of multiple layers, including convolutional layers, pooling layers, and fully connected layers.

Convolutional layers are the building blocks of CNNs. They apply convolution operations to input images, using learnable filters (also called kernels) to extract features

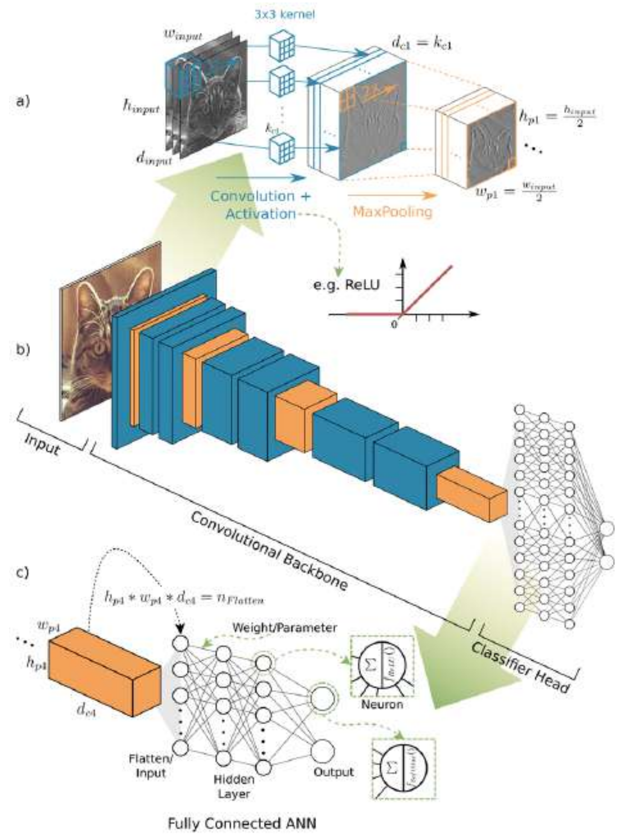


Figure 7. Overview and details of a convolutional neural network (CNN) architecture for image recognition [10]

such as edges, textures, and patterns. These filters slide over the input image, computing dot products between the filter weights and the input pixels to produce feature maps. Multiple filters are used in each convolutional layer to capture different features.

Pooling layers downsample the feature maps produced by the convolutional layers, reducing their spatial dimensions while retaining the most important information. The most common type of pooling operation is max pooling, where the maximum value within each region of the feature map is retained, effectively reducing the size of the feature maps.

Fully connected layers, also known as dense layers, are traditional neural network layers where every neuron is connected to every neuron in the previous and subsequent layers. These layers are typically used at the end of the CNN to map the extracted features to the output classes or labels.

As for the size of the training sample, it depends on various factors such as the complexity of the task, the diversity of the dataset, and the chosen neural network architecture. In general, larger datasets tend to yield better-performing models, especially for deep learning tasks. However, the minimum size of the training sample required for effective model training can vary significantly depending on the specific problem being addressed. It is essential to strike a balance between dataset size, computational resources, and model performance when designing the training pipeline. In the case of medical imaging tasks like thyroid ultrasound analysis, larger datasets with thousands to tens of thousands of labeled images are typically required to train accurate and robust models.

NNs can be trained effectively even with limited labeled data by employing data augmentation techniques. These techniques involve applying transformations such as rotation, scaling, flipping, and cropping to the input images, thereby augmenting the training dataset and improving the model's generalization ability.

Pre-trained CNN models, which have been trained on large-scale datasets such as ImageNet, can be fine-tuned for medical image analysis tasks with relatively small datasets. By leveraging the feature representations learned from generic image data, transfer learning enables CNNs to achieve better performance and faster convergence when applied to medical image datasets, including thyroid ultrasound images.

CNNs can provide insights into the decision-making process by generating heatmaps or saliency maps that highlight the regions of the image that contribute most to the model's predictions. This interpretability is valuable for clinicians, as it helps them understand why a particular diagnosis or classification was made by the CNN.

In summary, Convolutional Neural Networks offer powerful capabilities for automatically detecting and localizing cysts, tumors, and other abnormalities on thyroid ultrasound images. By learning complex patterns and structures from labeled data, CNNs can assist radiologists and clinicians in diagnosing thyroid pathologies more accurately and efficiently, leading to improved patient outcomes.

The expected result of the implementation of the approach should be a web application with artificial intelligence inside. Between the desktop and the web application, the choice fell on the second option. This is due to the fact that the neural network is able to independently learn additionally in the course of its work. To do this, she needs to have access to the results of working with the application of other users. It is more convenient to do this in a web format. It is also necessary to be able to refine the application. By updating web applications, the added changes will quickly appear to all users, unlike the desktop application, where each user will have to update it.

V. Overcoming obstacles

However, there are some serious pitfalls here.

With the web approach, a single server will have access to all application data. This violates the privacy policy and the protection of the user's personal data. The solution to this problem was found in having a separate server for each medical facility. And also not to transfer user data to the application. Based on the specifics of this software, it can process anonymous data and this will not affect the result.

The second difficulty encountered along the way is the presence of artifacts in the research process. The neural network must learn how to process them.

Artifacts in ultrasonic diagnostic imaging refer to misleading features or distortions present in the ultrasound image that do not accurately represent the anatomical structures being examined. These artifacts can arise due to various factors, including the properties of the ultrasound beam, the interaction of ultrasound waves with tissues, equipment settings, patient characteristics, and operator technique. Understanding and mitigating artifacts are essential for ensuring the accuracy and reliability of ultrasound diagnoses.

There are different types of artifacts. Reverberation Artifacts occurs when sound waves bounce back and forth between two strong reflectors, creating multiple, evenly spaced echoes on the image. It can give the appearance of additional structures or false boundaries within tissues.

Shadowing occurs when sound waves are attenuated by highly reflective or dense structures, resulting in a hypoechoic or anechoic region behind the structure. This can obscure underlying structures and limit visualization.

Edge artifacts occur at the interfaces between tissues with different acoustic properties. They manifest as bright or dark lines along tissue boundaries and can distort the appearance of adjacent structures.

Noise in ultrasound images can result from electronic interference, acoustic reverberations, or random fluctuations in signal intensity. It can degrade image quality and reduce diagnostic accuracy.

Motion artifacts occur when there is movement of the patient or probe during image acquisition. This can lead to blurring or ghosting of structures and compromise image clarity.

Teaching a CNN to process artifacts in ultrasound images is not an easy task. But there are some ways to overcome it.

Adversarial training involves training the CNN simultaneously with a generator network that generates realistic artifacts and a discriminator network that distinguishes between real images and artifacts. This helps the CNN learn to discriminate between artifacts and true structures.

Constructing a dataset that includes annotated examples of various artifacts encountered in clinical practice can facilitate CNN training. Annotating images to identify regions affected by artifacts allows the CNN to learn to ignore or compensate for them during analysis.

Pre-trained CNN models trained on general image datasets can be fine-tuned using ultrasound images containing artifacts. By leveraging the feature representations learned from diverse datasets, transfer learning enables the CNN to adapt to artifact-rich ultrasound images more effectively.

By training CNNs to recognize and process artifacts in ultrasound images, it can enhance the robustness and reliability of automated diagnostic systems, ultimately improving patient care and outcomes in ultrasonic diagnostic imaging.

Another difficulty is the structure of the organ itself. Although the thyroid gland was chosen as an example for research in this article as one of the easiest organs to analyze, it has its own characteristics. The thyroid gland consists of lobes. On both sides of the organ there are carotid arteries, in which there is an active blood flow, sometimes it looks pulsating on ultrasound. This may prevent the neural network from performing a high-quality analysis. Moreover, in the middle of the organ is the larynx, which also needs to be distinguished from pathology.

However, the thyroid gland is still an easy organ to diagnose. In comparison, for example, with abdominal organs, thyroid ultrasound rarely shows a nebula associated with a large amount of subcutaneous fat in the patient.

Figure 8. Thyroid gland ultrasonography example [11]

Due to the fact that every person has a larynx and carotid arteries, the neural network will learn to isolate them and accept them as normal thanks to a large training sample.

In the picture 6, the round blackouts on the sides of the thyroid gland are the carotid arteries. And the round gray area in the middle is the larynx.

VI. OSTIS Technology integration

Working with artificial intelligence is not limited to neural networks alone. One of the strong representatives

Figure 9. Thyroid gland ultrasonography in longitudinal projection example [12]

of symbolic artificial intelligence is OSTIS Technology. [13]

By integrating this technology into the described project, the following results can be achieved:

1) Thanks to the implementation of OSTIS, it is possible to additionally train the neural network not only on ongoing research, but also on feedback from medical experts.

2) An intelligent assistant system can be integrated into the application, which will determine not only the presence or absence of pathology, but will also be able to analyze the general state of the patient's health and draw conclusions about what a particular problem in the body is related to.

3) The treatment regimen for some pathologies is described by protocols and is similar in different patients. Thus, the system integrated with OSTIS will be able not only to check for problems in the organ, but also to offer appropriate treatment. Thus, the doctor will not have to write it himself. It will only be enough to edit a ready-made treatment regimen.

4) OSTIS is based on a knowledge base. Therefore, the system takes all the information from there and draws conclusions based on it. Although neural networks are a fairly productive tool, they have a large percentage of error. OSTIS will help to minimize the number of incorrect answers and reduce the reliability of the system analysis to 99%.

VII. Conclusion

Medical ultrasonography, a widely-used diagnostic imaging modality, plays a pivotal role in healthcare by providing real-time images of internal organs and tissues. However, the manual interpretation of ultrasound images can be challenging and time-consuming, often requiring specialized expertise. In recent years, significant advancements in artificial intelligence (AI) and image analysis techniques have revolutionized medical imaging, paving the way for the automation of ultrasonography interpretation through intelligent image analysis.

This article provides a comprehensive overview of the application of AI in medical ultrasonography and its potential to enhance diagnostic accuracy, efficiency, and patient care. It proposes one of the solutions which can help to minimize the number of errors associated with the human factor. After all, an ultrasound diagnostic doctor should be extremely attentive and focused throughout the entire work shift. However, the study may take place at night, the person may be in poor health, there may be too many patients, the doctor may not have enough experience. These factors directly affect the quality of the study and the timeliness of diagnosis of life-threatening pathologies.

At the moment, artificial intelligence is rarely used on a large scale due to the fact that it cannot completely replace humans. Especially in such an area as medicine. This sphere doesn't forgive mistakes. The option proposed here is a compromise between using only artificial intelligence and only human power.

The article describes an algorithm for creating an intelligent system for determining thyroid pathologies using image analysis. During the work, the advantages and disadvantages of this approach were considered, and options for overcoming the difficulties that will have to be faced during the implementation of the project were proposed. The subject area was also analyzed, the process of ultrasound examination, the principle of operation of the ultrasound machine were described. Moreover, an analysis of existing publications and projects on related topics was carried out.

Integration of the system with OSTIS technology was also proposed.

The automation of medical ultrasonography through intelligent image analysis holds great promise for improving diagnostic accuracy, efficiency, and patient outcomes. By harnessing the power of AI and deep learning techniques, clinicians can leverage advanced tools to enhance their diagnostic capabilities and provide better patient care. However, further research, validation, and collaboration between clinicians, researchers, and technologists are essential to overcome challenges and realize the full potential of AI-driven automatization in medical imaging.

References

[1] Thyroid. [Online]. Available: https://en.wikipedia.org/wiki/Thyroid

[2] (2020, Sep) How do ultrasound examinations work? [Online]. Available: https://www.informedhealth.org/how-do-ultrasound-examinations-work.html

[3] (2023, Dec) ultrasound. [Online]. Available: https://www.nibib.nih.gov/science-education/science-topics/ultrasound

[4] (2017) Clinical ultrasound diagnosis of the thyroid gland. [Online]. Available: https://ultrasonicthyroid.ru/part-1

[5] (2004) Dopplerography — doppler ultrasound examination of the thyroid gland. [Online]. Available: https://www.dr-md.ru/doppler.html

[6] Doppler. [Online]. Available: https://radiopaedia.org/articles/thyroid-inferno

[7] C.-L. Cao, Q.-L. Li, J. Tong, L.-N. Shi, Wen-Xiao, L. Y. Xu, J. Cheng, T.-T. Du, J. li, and X.-W. Cui, “Artificial intelligence in thyroid ultrasound,” *Frontiers in Oncology*, May 2023.

[8] Samsung ultrasound system. [Online]. Available: https://www.samsunghealthcare.com/en/products/UltrasoundSystem

[9] C. M. Vasile, A. L. Udriștoiu, A. E. Ghenea, M. Popescu, C. Gheonea, C. E. Niculescu, A. M. Ungureanu, Ștefan Udriștoiu, A. I. Drocaş, L. G. Gruionu, G. Gruionu, A. V. Iacob, , and D. O. Alexandru, “Intelligent diagnosis of thyroid ultrasound imaging using an ensemble of deep learning methods,” *Medicina (Kaunas)*, p. 395, Apr 2021.

[10] T. Hoeser and C. Kuenzer, “Object detection and image segmentation with deep learning on earth observation data: A review-part i: Evolution and recent trends,” *Remote Sensing*, vol. 12, May 2020.

[11] Thyroid gland. [Online]. Available: https://gormonia.com/ru/services

[12] Interpretation of ultrasound of the thyroid gland. 4th lesson for patients. [Online]. Available: https://dzen.ru/a/XvA7FQxUFi9_RiGv

[13] Ostis project. [Online]. Available: https://github.com/ostis-ai

АВТОМАТИЗАЦИЯ УЛЬТРАЗВУКОВОГО ИССЛЕДОВАНИЯ ЩИТОВИДНОЙ ЖЕЛЕЗЫ С ПОМОЩЬЮ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА

Черкас Е. О.

Эта статья предлагает алгоритм автоматизации процесса медицинской ультразвуковой диагностики с помощью интеллектуального анализа. Действия описаны на примере исследования щитовидной железы. Дополнительная проверка результата со стороны нейронной сети позволяет начинающим докторам чувствовать себя более уверенно и минимизировать влияние человеческого фактора на качество диагностики.

Received 02.04.2024

Chest X-ray Image Processing Based on Radiologists' Textual Annotations

Aleksandra Kosareva, Dzmitry Paulenka, Eduard Snezhko
The United Institute of Informatics Problems of the National Academy of Sciences of Belarus
Biomedical Image Analysis Department, https://image.org.by
Minsk, Belarus
Email: kosarevaaleksandra4317@gmail.com, dmitri.pavlenko@gmail.com,
eduard.snezhko@gmail.com

***Abstract*—More than 11,000 chest x-ray images and their corresponding text annotations were analyzed, and the first pilot studies on image processing tailored to text annotations of radiology specialists were conducted. An image processing pipeline for a database and for a neural network has been developed. The prediction of the parameter "Overall percent of abnormal volume" was performed and the mean absolute error (MAE) for the InceptionResNet50V2 neural network model was 11.073.**

***Keywords*—medical image processing, medical image analysis, deep learning, computer-aided diagnosis, chest x-ray, textual annotations of lung lesions**

I. Introduction

In this article the main efforts are made to analyze and prepare Chest X-ray (CXR) images and corresponding *text* data annotations. A total of *11,493* non-empty CXR images were downloaded (in fact there are 13,521 instances, however 2,028 of them were empty and not downloaded from TB Portals [1] website).

On the CASE BROWSER [2] website CXR text annotation can be viewed as in Fig. 1 along the following path:

Patient: № → Case → View Imaging Study → Diagnostic Report.

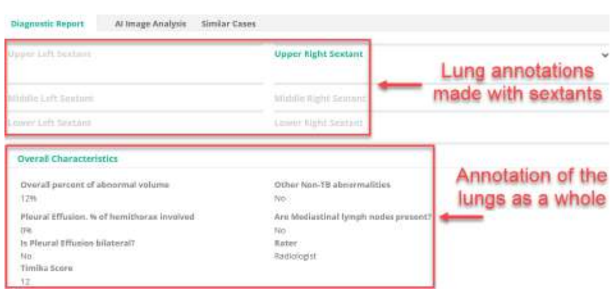


Figure 1. CXR textual annotation on the CASE BROWSER [2] website.

All CXR textual data in corresponding JSON files as well as on the CASE BROWSER [2] and TB DEPOT [3] websites are divided into three blocks of information (Fig. 2):

- anonymized patient information (gender, age, country, diagnosis, etc.);
- CXR radiologists' textual annotations, *which is currently being analyzed* in this article;
- treatment history, which includes medications and treatment days.

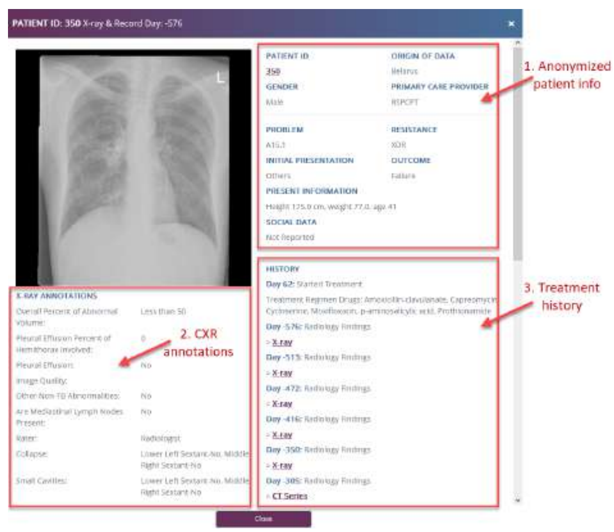


Figure 2. CXR case information on the TB DEPOT [3] website.

Second data block in Fig. 2 or "CXR annotations" can also be roughly divided into three groups:

- CXR annotations, as if it were a computed tomography (CT) scan, in total 106 images;
- CXR annotations of the six lobes of the lung (sextants) with twenty parameters for each sextant, in total 546,364 non-empty text annotations in 9,154 CXR images;
- CXR annotations of the lung as a whole using six parameters, see "Overall Characteristics" tab in Fig. 1, in total 11,387 CXR images.

For clarity, the image categories are shown in the diagram below, Fig. 3.

The results of the analyses of the annotations of these groups are summarized below. Further, after the data

This work was carried out with the financial support of the ISTC-PR150 "Belarus TB Database and TB Portal" project.

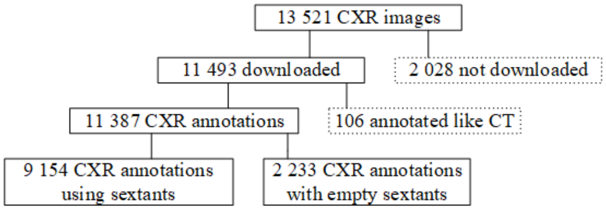


Figure 3. CXR image categories from TB Portals [1] website.

analysis there is a report on the use of the obtained data in solving the prediction of the parameter "Overall percent of abnormal volume" is given.

II. CXR annotations, as if it were a CT

There are 106 images described like a CT. All from Georgia. These images have 19 parameters with disease descriptions, like a CT scan. Annotation parameters for CXR are absent.

For example, in Fig. 4, patient ID 1534 has one annotated image with a study modality CR (Computed Radiography), but the actual description is as if it were a CT (Computed Tomography).

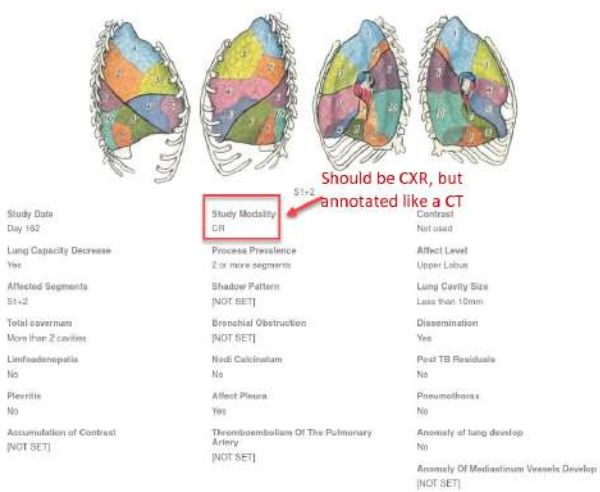


Figure 4. Patient ID 1534 with CXR annotation, as if it were a CT.

It was decided to exclude these images from the further investigations for the following reasons:

- their annotation is different than most other files, meaning that these images cannot participate in neural network training together with everyone else;
- 106 files can participate in separate neural network training, but there is more valuable data to explore;
- it is unclear why these images are described differently, and until this is clarified, it is better to exclude them from further study.

III. CXR annotations of sextants

A. *Lesion names to describe the lungs lobes*

In the CXR image, the lungs are divided into six lobes or sextants:

- Upper Right Sextant;
- Upper Left Sextant;
- Middle Right Sextant;
- Middle Left Sextant;
- Lower Right Sextant;
- Lower Left Sextant.

Each sextant has twenty identical lesion names, which can be found in the corresponding JSON files for each CXR image:

1) Small Cavities
2) Medium Cavities
3) Large Cavities
4) Is any Large cavity belong to a multi-sextant cavity?
5) Can Multiple cavities be seen?
6) Small Nodules
7) Medium Nodules
8) Large Nodules
9) Huge Nodules
10) Is any calcified or partially-calcified Nodule exists?
11) Is any non-calcified Nodule exists?
12) Is any clustered Nodule exists
13) Are multiple Nodules exist?
14) Low/ground glass Density (active fresh nodules)
15) Medium Density (stabilized fibrotic nodules)
16) High Density (calcified nodules, typically sequella)
17) Low/ground glass Density
18) Medium Density
19) High Density
20) Collapse

The data displayed on the CASE BROWSER [2] and TB DEPOT [3] websites and the tag names in JSON files are slightly different, but a mutually unambiguous match can be made even by a non-specialist. The only non-obvious correspondence is that the field "Infiltrate" on the website should correspond to the tag "High Density" in JSON files. This match between the "Infiltrate" field and "High Density" tag was found by the method of exceptions as the last possible match.

Mutually unambiguous correspondence between the data on the websites and in JSON files is given in Tab. I and Tab. III.

The sequential arrangement of the twenty lesions in Tab. I is the same as on the CASE BROWSER [2] website.

The number of non-empty lesions in the lung sextants is listed in Tab. II.

For convenience and to save space, in Tab. II, the word "Sextant" is omitted in table header. The total number of

Table I
Twenty lesions to describe the lungs lobes in CXR images

№	Tags in **JSON** files "info-all.json" [a]	**CASE BROWSER** [2] website	**TB DEPOT** [3] website
		Cavity	
1	Small Cavities	Small Cavities (less than 3 cm)	Small Cavities
2	Medium Cavities	Medium Cavities (3-5 cm)	Medium Cavities
3	Large Cavities	Large Cavities (more than 5 cm)	Large Cavities
4	Is any Large cavity belong to a multi-sextant cavity?	Does any Large cavity belong to a multi-sextant cavity?	Large Cavity Multi Sextant
5	Can Multiple cavities be seen?	Can Multiple cavities be seen?	Multiple Cavities Seen
		Nodules	
6	Small Nodules	Small Nodules (less than 3 mm)	Small Nodules
7	Medium Nodules	Medium Nodules (5-15 mm)	Medium Nodules
8	Large Nodules	Large Nodules (15-30 mm)	Large Nodules
9	Huge Nodules	Huge Nodules (more than 30 mm, tuberculoma)	Huge Nodules
10	Is any calcified or partially-calcified Nodule exists?	Any calcified or partially-calcified Nodules?	Partially Calcified Nodule Exists
11	Is any non-calcified Nodule exists?	Any non-calcified Nodules?	Non Calcified Nodule Exists
12	Is any clustered Nodule exists	Any clustered Nodules (nodules 2-5 mm apart)?	Any Clustered Nodule Exists
13	Are multiple Nodules exist?	Can Multiple Nodules be seen?	Multiple Nodule Exists
14	Low/ground glass Density (active fresh nodules)	Low/ground glass Density (active fresh nodules)	Low Ground Glass Density Active Fresh Nodules
15	Medium Density (stabilized fibrotic nodules)	Medium Density (stabilized fibrotic nodules)	Medium Density Stabalized Fibrotic Nodules
16	High Density (calcified nodules, typically sequella)	High Density (calcified nodules, typically sequella)	High Density Calcified Typically Sequella
		Infiltrate	
17	Low/ground glass Density	Low/ground glass Density	Infiltrate Low/Ground Density
18	Medium Density	Medium Density	Infiltrate Medium Density
19	High Density	High Density	Infiltrate
		Collapse	
20	Collapse	Collapse	Collapse

[a]Lesion names have been left unchanged for accuracy of description.

Table II
The number of non-empty lesions in the lung sextants

Lesion name	Upper Right	Upper Left	Middle Right	Middle Left	Lower Right	Lower Left
Small Cavities	6723	5856	4654	4717	2756	2596
Medium Cavities	6720	5856	4653	4714	2755	2596
Large Cavities	6719	5855	4652	4714	2755	2596
Is any Large cavity belong to a multi-sextant cavity?	6732	5868	4663	4727	2768	2608
Can Multiple cavities be seen?	6732	5867	4661	4727	2768	2606
Small Nodules	6724	5856	4656	4719	2757	2599
Medium Nodules	6724	5855	4653	4715	2755	2596
Large Nodules	6720	5853	4652	4714	2755	2596
Huge Nodules	6720	5854	4652	4714	2755	2596
Is any calcified or partially-calcified Nodule exists?	6733	5868	4666	4727	2768	2609
Is any non-calcified Nodule exists?	6731	5868	4665	4728	2768	2609
Is any clustered Nodule exists	6731	5864	4665	4727	2769	2608
Are multiple Nodules exist?	6730	5862	4661	4723	2768	2608
Low/ground glass Density (active fresh nodules)	6726	5856	4656	4718	2757	2599
Medium Density (stabilized fibrotic nodules)	6724	5857	4654	4716	2755	2596
High Density (calcified nodules, typically sequella)	6720	5853	4652	4714	2754	2596
Low/ground glass Density	6724	5856	4657	4715	2760	2598
Medium Density	6724	5858	4655	4716	2757	2597
High Density	6720	5855	4652	4713	2755	2596
Collapse	6720	5855	4652	4714	2754	2598

text annotations is 546,364. For the right lung: 282,817. For the left lung: 263,547.

The division of the lungs into sextants is made by a radiologist according to the following rules.

For CXR:

- Upper sextants are above the lower edge of the aortic arch;
- Middle sextants are between the lower edge of the aortic arch and the right inferior pulmonary vein);
- Lower sextants are below the right inferior pulmonary vein.

For CT:

- Upper sextants are above keel level;
- Middle sextants are between keel and right inferior pulmonary vein;
- Lower sextants are below right inferior pulmonary vein.

Approximately, *with some reservations*, the lung may be considered to be divided into three equal lobes vertically and into two equal lobes horizontally. It was decided to determine the boundaries of the sextants by calculating the third part of the total area of the lungs. With this choice of boundaries, the sizes of the sextants will be approximately the same in area. In this case the boundaries of the sextant are approximate and not related to lung anatomy. The left lung is where the heart and left arm are located. The right lung is where the right arm is located.

Tab. II shows that the highest number of lung diseases was registered in the upper lung lobes and the lowest number of lung diseases was registered in the lower lung lobes. At the same time, on average, more diseases were registered in the right lung than in the left lung.

According to Radiopaedia "post-primary infections have a strong predilection for the upper zones" [4]. Right versus left lesions asymmetry is not so obvious as top versus bottom and depends on the type (class) of the lesion. Statistical Atlas of Lung Lesions [5], [6], which was made by our team, is still in research. The difficulty in studying the statistical distribution of lesions is that they are of different nature and different biological substrates.

B. *Analysis of sextant lesion names*

According to the CASE BROWSER [2] website 20 lesion names are divided on the four groups (classes): "Cavity", "Nodules", "Infiltrate" and "Collapse" (see Tab. I).

Additional analysis of the sextant lesions list is shown in Fig. 5.

In Fig. 5, equally significant sections of text are highlighted with the same color. The 20 parameters are specially grouped in such a way that the same significant values are next to each other. Here are the definitions for these diseases.

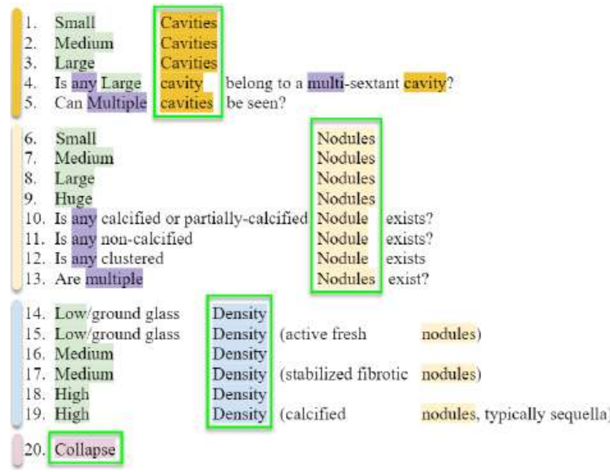


Figure 5. Graphical result of lesion names analysis for lung lobes.

A lung *cavity* or pulmonary cavity is an abnormal, thick-walled, air-filled space within the lung. Cavities in the lung can be caused by infections, cancer, autoimmune conditions, trauma, congenital defects, or pulmonary embolism. The most common cause of a single lung cavity is lung cancer. Bacterial, mycobacterial, and fungal infections are common causes of lung cavities [7].

According to the glossary of terms for chest imaging proposed by the Fleischner Society, a lung *nodule* is defined as an approximately rounded opacity more or less well-defined measuring up to 3 cm in diameter [8]. Rounded lesions measuring more than 3 cm in diameter are termed *lung masses* and should be considered indicative of lung cancer until histologically proven otherwise. Lung mass approach differs from that of nodules [9].

The *density* of the lung reflects the total mass of fluid, air, and dry lung tissue per unit volume of the lung. Lung density can be measured by evaluation of attenuation of an electron beam with CT. This technique has been shown to be sufficiently reliable and sensitive to distinguish normal from abnormal lung water [10].

A pulmonary *infiltrate* is a substance denser than air, such as pus, blood, or protein, which lingers within the parenchyma of the lungs. Pulmonary infiltrates are associated with pneumonia, tuberculosis, and sarcoidosis [11].

A *collapsed* lung (pneumothorax) occurs when air escapes from the lung. The air then fills the space outside of the lung between the lung and chest wall. This buildup of air puts pressure on the lung, so it cannot expand as much as it normally does when you take a breath. The medical name of this condition is pneumothorax [12].

Four general categories (classes) of disease can be distinguished from the definitions above:

1) class "Cavity" in the presence of one or more abnormal thick-walled spaces in the lung filled with air (gas);

2) class "Nodule" in the presence of one or more small up to 3 cm opacities;
3) class "Density" or "Infiltrate" in the presence of abnormal lung thickenings;
4) class "Collapse" or pneumothorax, when the lung cannot fully expand when breathe in.

Lesions analysis showed gradation in density, with the exception of the "Collapse" class. The "Cavity" class has the lowest density. Abnormal fluids and other seals are next that fall into the "Density" ("Infiltrate") class. And the densest is the "Nodule" class.

Nodules, abnormal density and cavities can be of different sizes (small, low, medium, large, high, huge), in different numbers (any, multiple, belong to a multi-sextant), in different qualities, especially nodules (calcified, partially-calcified, non-calcified, clustered, stabilized, active).

IV. CXR annotations of overall characteristics

Parameters used to describe the lungs as a whole (without sextants) are in Tab. III.

The seventh parameter "Timika Score" from Tab. III is not present in JSON files. The *Timika CXR score* is a machine learning tool for diagnosing tuberculosis, which was developed in 2010 by investigators at the Menzies School of Health Research in Darwin, Australia. The score in a scale of 1 to 140 was designed for physicians in underserved clinical settings and is based on the overall abnormal percent of volume of the lungs on CXR, plus the presence of cavitation [13], [14]. Most likely, the Timika CXR score is calculated automatically after the radiologist annotation and is not stored in the TB database.

The "Rater" parameter can have three values: "General practitioner", "Radiologist" or "Other", and if it is set, it means that the CXR image has been annotated. In this case, if the image is annotated and the sextants are blank, the patient is free of lung lesions.

V. Description of the top catalogues for further research

At this stage of the project, *the main effort* was focused on building the CXR image databases for further investigations. The catalogue tree is represented as follows.

level0 Original data from TB Portals [1] in DICOM format ("*.dcm") containing CXR, CT, JSON description files and other auxiliary files.

level1 All CXR and CT images have been converted to NIFTI format ("*.nii.gz").

level2 CXR images were manually reviewed and only those images that could be used in further research were selected into this catalogue.

- *cxr* Selected CXR images.
- *cxr_annotations* Excel tables with text annotations, paths in *level1* directory and other auxiliary information for each CXR file.
- *cxr_scripts* Python scripts to prepare *level2* directory.
- *cxr_masks* Lung masks obtained via LungExpert API [15].
- *cxr_thumbnails* Images preview in PNG format and size 512×512 pixels.
- *cxr_data* Preprocessed CXR images. For example, CXR images cropped by lung mask, normalized to the range [-1, +1] and then resized via lanczos method to 256×256 or 512×512 pixels.
- *temp* Intermediate files that will be deleted after *level2* is finished.

level3 Directories with investigations. Each subdirectory here is the separate investigation.

- *cxr_abnormal_volume* Investigation of the lung lesion percentage.
- *cxr_sextants* Investigation of the lung lobes.

All directories named "*cxr_*" mean that CXR files are processed.

All directories named "*ct_*" mean that CT files are processed.

CXR images processing pipeline:

- manual review and screening out defective images with saving data in the "*cxr*" directory;
- verification and correction for orientation, inversion, etc. with modification of the data in the "*cxr*" directory;
- getting mask via LungExpert API [15] with saving data in the "*cxr_masks*" directory;
- normalization and resizing with saving data in the "*cxr_data*" directory;
- slicing into sextants and saving data in the "*cxr_sextants*" directory.

The processing pipeline for the neural network is shown in Fig. 6:

- applying modality, inversion and orientation checks to the input image;
- obtaining lung mask via LungExpert API [15];
- cropping the image along the borders of the lungs, normalization and resizing;
- slice into sextants for lung lobes investigation;
- application of neural networks to determine overall characteristics or to determine sextant lesion characteristics;
- comparing the results obtained with the radiologist's annotations.

VI. Prediction of the parameter "Overall percent of abnormal volume"

A. Task description

The first parameter to research was the "Overall percent of abnormal volume" parameter. It can be an integer between 0 and 100 %. Zero percent means that the lungs are healthy and have no abnormal volume and 100 %

Table III
Parameters to describe the lungs as a whole

№	Tags in **JSON** files "info-all.json"	**CASE BROWSER** [2] website	**TB DEPOT** [3] website
1	Overall percent of abnormal volume	Overall percent of abnormal volume	Overall Percent of Abnormal Volume
2	Pleural Effusion. % of hemithorax involved	Pleural Effusion. % of hemithorax involved	Pleural Effusion Percent of Hemithorax Involved
3	Is Pleural Effusion bilateral?	Is Pleural Effusion bilateral?	Pleural Effusion
4	Other Non-TB abnormalities	Other Non-TB abnormalities	Other Non-TB Abnormalities
5	Are Mediastinal lymph nodes present?	Are Mediastinal lymph nodes present?	Are Mediastinal Lymph Nodes Present
6	Rater	Rater	Rater
7	—	Timika Score	—

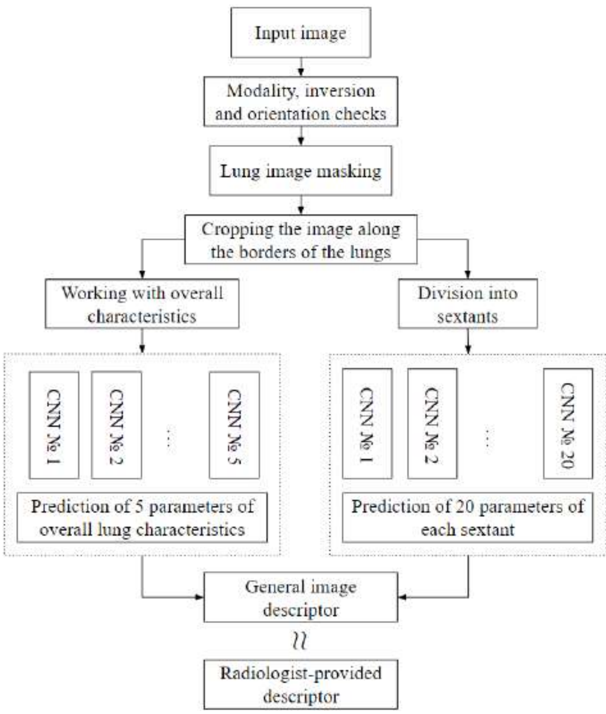


Figure 6. The processing pipeline for the neural network.

means that the entire lung volume is affected. From the TB DEPOT data dictionary [16]: "Overall percent of abnormal volume. Pleural effusion should be excluded. This is a professional judgment number in addition to the volume that can be calculated".

The *InceptionResNet50V2* neural network is used to predict the parameter "Overall percent of abnormal volume" based on the input CXR image. Among several tested architectures, this neural network showed the best results. It has also been suggested that it is not necessary to use a neural network to predict an abnormal percentage of lung volume. It was assumed that the "classical" machine learning method based on regression analysis would suffice.

Thus, three phases have been identified to fulfill this task:

1) preparation of the primary dataset;
2) application of the Support Vector Machine (SVM) regression method;
3) conducting experiments and comparing results.

This study was carried out in order to create a correct pipeline for further investigations of other CXR image parameters.

B. Preparation of the primary dataset

To prepare the dataset, all images were reviewed. Unsuitable images have been excluded. An example of such excluded images is shown in Fig. 7.

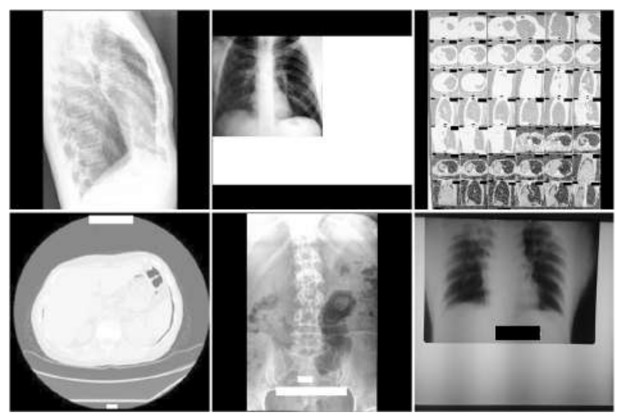


Figure 7. Examples of excluded images.

Excluded images are images with lateral patient orientation, containing large white/black frames, incorrect modality, other body parts, etc.

After review the primary dataset contains in total 8,875 images. It was noticed that some remained images are flipped horizontally (the heart is on the right side). For example, a patient with ID 8620 in Fig. 8 has heart on the right side of the body, the image is flipped.

A simple three-layer convolutional neural network (CNN) was implemented to find such incorrectly oriented images (Fig. 9).

To train this three-layer CNN an additional dataset was used. Each class of the training dataset contains 1,260 flipped (heart on the right) and 1,260 correctly oriented images. Test dataset contains 314 CXR images in each class.

F1-score on the test dataset was 0.9936. The resulting model was enough to optimize the process of preparing a

Figure 8. Incorrectly orientated image, the heart is on the right side.

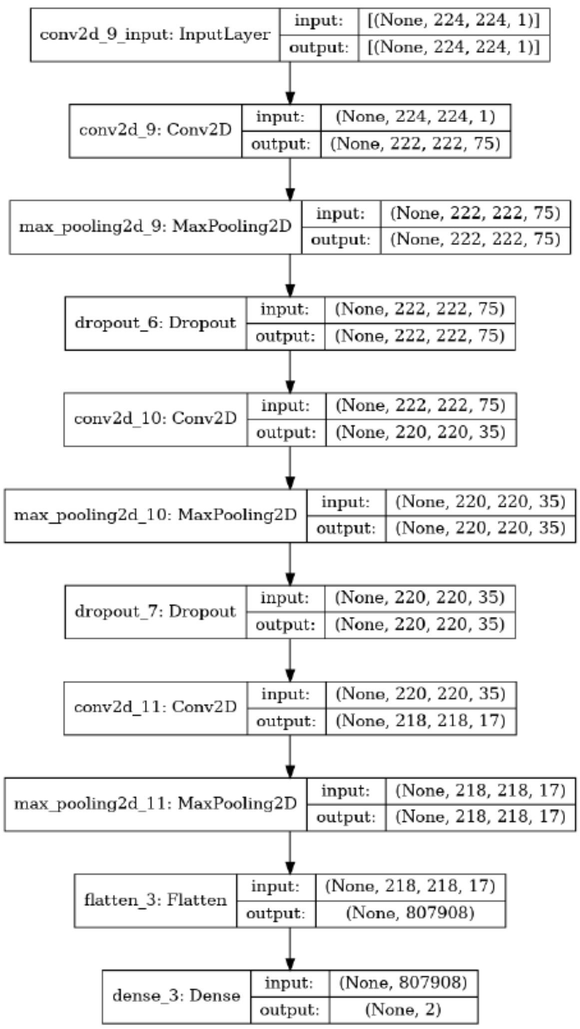


Figure 9. Three-layer CNN architecture to find horizontally flipped images.

general dataset. Solving the image orientation problem is important for proper slicing of CXR images into sextants.

As a result, the primary dataset obtained consists of 6,000 CXR images in the train set and 2,875 CXR images in the test set.

C. Application of the Support Vector Machine regression method

It was hypothesized that a "classical" machine learning method without the use of AI approaches would be sufficient to predict the value of parameter "Overall percent of abnormal volume" from the CXR input image.

To test this hypothesis SVM regression method using image histograms was used and compared with neural network method. A pre-trained on the ImageNet [17] CNN InceptionResNet50V2 was chosen as a neural network method.

Both methods are compared on the same dataset (6,000 CXR images in the train set and 2,875 CXR images in the test set).

The mean absolute error (MAE) was calculated as a metric for the analysis.

D. Conducting experiments and comparing results

MAE for the SVM method was 17.6494.

MAE for the InceptionResNet50V2 was 11.0730.

Undoubtedly, the margin of error is smaller when using InceptionResNet50V2. Accordingly, the SVM method did not perform well and cannot be used to predict "Overall percent of abnormal volume" parameter from the CXR input image.

The Grad-CAM [18] algorithm to visualize class activation maps was used to analyze the performance of the InceptionResNet50V2 neural network.

Examples of correctly predicted CXR images with their prediction heatmaps are shown in Fig. 10.

A comparison of the neural network prediction and the radiologist's annotation showed that prediction heatmaps are partially cover the sextants marked by the radiologist. This is the result for only one of the annotated parameters, for a combination of a group of parameters the results can be significantly improved.

Two examples of incorrectly predicted "Overall percent of abnormal volume" with their prediction heatmaps are shown in Fig. 11.

Both images in Fig. 11 have an "Overall percent of abnormal volume" parameter equal to 5 %, but the neural network predicts values of 25 % and 14 % respectively. On the first example the greatest activation of the heatmap occurred outside the lungs or in the background. On the second example the greatest activation of the heatmap revealed the artifact: protective lead apron.

Some possible errors in CXR textual annotations have also been discovered. For example, the patient with ID 426 in Fig. 12 obviously has a damaged right lung,

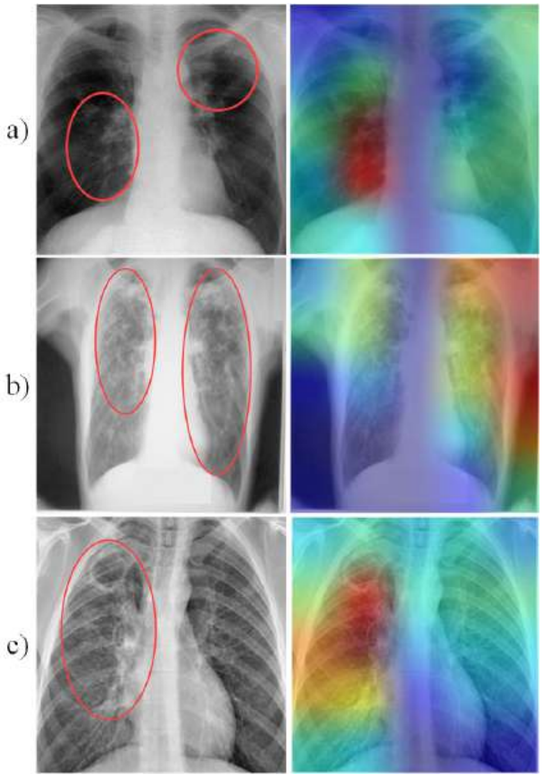


Figure 10. Examples of correctly predicted images with their prediction heatmaps for "Overall percent of abnormal volume" parameter:
a) patient ID 13676, actual value 14 %, predicted value 14 %;
b) patient ID 2641, actual value 30 %, predicted value 28 %;
c) patient ID 19415, actual value 34 %, predicted value 33 %.

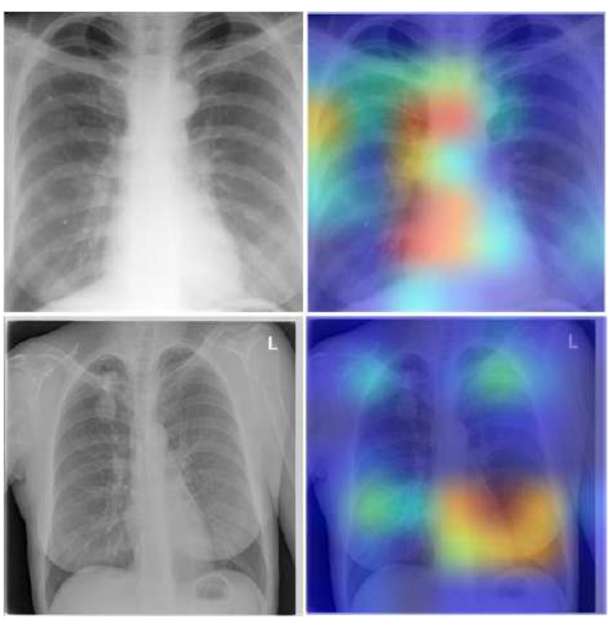


Figure 11. Examples of incorrectly predicted "Overall percent of abnormal volume" with corresponding prediction heatmaps.

but "Overall percent of abnormal volume" parameter is set to zero.

Figure 12. The patient's right lung has been damaged, but "Overall percent of abnormal volume" parameter is set to zero.

A study of patient ID 426 in the CASE BROWSER [2] found that the other parameter "Pleural Effusion. % of hemithorax involved" is equal to 50 %. Also, four sextants were annotated and have some lesions: Upper, Middle, Lower Right and Lower Left Sextants. The right lung is particularly badly damaged.

At least 44 such images with markup inconsistencies were found. It is necessary to exclude them from the training samples.

Simple consistency checks between different CXR annotation markup parameters should be developed for future research. It is necessary to pay attention not only to the images, but also to the various textual descriptions provided.

As a result, it was concluded that the high MAE value is due to the following reasons:

- background noise (outside the lungs) that requires mask cropping to remove it;
- the presence of artifacts in the lungs that requires more data to train the neural network to correctly distinguish these artifacts;
- some errors in CXR image annotations that should be excluded from the dataset.

VII. Discussion of the application of semantic technologies in the context of the tasks considered in this paper

In 1982, Japanese scientists developed a program for the fifth generation of electronic computing machines. Despite the fact that more than 42 years have passed, the fifth generation computers have not been fully realized. The main difficulty in creating fifth-generation computers or future computers is to create a machine with artificial intelligence (AI) that will be able to draw logical conclusions from the facts presented.

To interact with a fifth-generation computer, a person (user) will not need to develop software for the machine to solve the task at hand. In addition, fifth generation computers will solve the problem of data formalization in the interaction between computer and computer and human and computer. Commands for the machine can be formulated in ordinary spoken language without knowledge of formal programming languages as well as input and output data formats.

Thus, fifth-generation computers are intelligent semantic systems with extremely high interoperability. Interoperability is the ability of a product or system, whose interfaces are completely open, to interact and function with other products or systems without any access or implementation restrictions.

Although fifth generation computers have not yet been realized, active work is being done to remove the barriers between humans and computers, and between computers and computers.

One of the directions for the development of interoperable intelligent computer systems is automatic data transformation between different modalities: image to text, text to image, speech to text, text to speech, image to speech, speech to image, image to music, 3D reconstruction of an object from a set of images, CT scan to X-ray image, satellite image to geographical map, geographical map to satellite image and other modality transformations [19], [20].

This research project is developing an intelligent semantic system that converts a textual description into an image, namely the textual markup of a radiologist into a heatmap of lesion foci in the corresponding CXR medical image.

A neural network trained to solve such a problem will be a decision support system for population screening. The trained neural network model will produce a textual description of lung lesions based on the patient's chest radiograph, as well as a heatmap corresponding to these lesions in the form of a graphical representation Fig. 13.

Figure 13. Operation schematic of the CXR-based decision support system under development.

The problem of generating a heatmap from a textual description and the subsequent problem of generating a heatmap and corresponding textual description from an input CXR image is one of the challenges of semantic image segmentation.

Semantic image segmentation is the task of dividing parts of an image into subgroups of pixels belonging to corresponding objects, with subsequent classification of these objects. Unlike classification and object detection tasks, the task of semantic segmentation is more complex both in terms of solution methods and computational resources [21].

Analyzing the literature to identify different semantic methods for medical image processing revealed the following approaches:

- probabilistic latent semantic analysis (PLSA), which, in conjunction with neural network, is able to mining the hidden semantics of an image [22];
- implementation of Semantic Similarity Graph Embedding (SSGE) framework, which explicitly explores the semantic similarities among images [23];
- investigation and development of the concept of a personal intellectual assistant (secretary, referent) [24].

Unfortunately, this paper does not apply the found semantic methods to medical image processing. However, the application of such semantic techniques in the context of the tasks considered in this paper is very relevant in future investigations.

As correctly noted in [25], "Currently, decision support systems in radiation mammology focus on the detection and classification of neoplasms, despite the fact that *the real work of a radiologist does not imply a diagnosing*. Computer vision systems use *a black box model* and do not explain the results of work, which is unacceptable in medicine".

This study also focuses on a deeper evaluation of the behavior of such a "black box" (neural network model) by studying the activation heatmaps on different convolutional layers of the neural network, which are obtained using the Gradient-weighted Class Activation Mapping (Grad-CAM) method [18]. The investigation of heatmaps on different convolutional layers provides a better understanding of the decision-making logic of the neural network based on the input data.

Fig. 5 shows the result of the semantic analysis of lesion names. As shown in Fig. 5, there are four classes of lesions: "Cavity", "Density", "Nodule" and "Collapse". Meanwhile, the two classes "Density" and "Nodule" have meaning overlap in the three lesion names. In the future, semantic analysis methods will be applied to better understand the disease asymmetry of different lung lobes (see Tab. II).

After the decision support system for population screening is completed, it is planned to be implemented into the already existing software "AI-based software for computer-assisted diagnosis of lung diseases using chest X-Ray and CT images" (LungExpert, https://lungs.org.by).

VIII. Conclusions

In this article, text annotations to CXR images were analyzed. At this stage of the project, the main efforts of the authors were focused on the formation of databases for further research. A catalogue tree with new datasets is described and constantly developing.

Two main tasks based on radiologists' textual annotations were planned and the corresponding pipeline for the neural networks training was described: pulmonary disease study using sextants and the pulmonary disease study using overall characteristics.

The task of predicting the parameter "Overall percent of abnormal volume" showed that it is necessary to develop additional simple consistency checks between different CXR textual annotations. Also using lung masks is a good idea to improve the quality of neural networks.

Further research is planned to create neural network attention heatmaps based on the textual descriptions of radiologists.

Acknowledgment

This work was carried out with the financial support of the ISTC-PR150 "Belarus TB Database and TB Portal" project.

References

[1] TB Portals. Available at: https://tbportals.niaid.nih.gov (accessed 2024, March)

[2] CASE BROWSER. Available at: https://data.tbportals.niaid.nih.gov (accessed 2024, March)

[3] TB DEPOT. Available at: https://depot.tbportals.niaid.nih.gov/#/cohort-creation?tab=4 (accessed 2024, March)

[4] Tuberculosis (pulmonary manifestations). Available at: https://radiopaedia.org/articles/tuberculosis-pulmonary-manifestations-1 (accessed 2024, March)

[5] Statistical Atlas of Lung Lesions. Available at: https://image.org.by/lesionAtlas (accessed 2024, March)

[6] V. Kovalev, V. Liauchuk, A. Gabrielian, A. Rosenthal, "Towards Statistical Atlas of Lung Lesions," *International Journal of Computer Assisted Radiology and Surgery*, 21-25 June, Munich, Germany, 2020, Vol. 15, Suppl. 1, pp. s31–s32. [Online]. Available: https://www.researchgate.net/publication/344217581_Towards_Statistical_Atlas_of_Lung_Lesions

[7] Lung cavity. Available at: https://en.wikipedia.org/wiki/Lung_cavity (accessed 2024, March)

[8] D.M. Hansell, A.A. Bankier, H. MacMahon, T.C. McLoud, N.L. Müller, J. Remy, et al., "Fleischner Society: Glossary of terms for thoracic imaging," *Radiology*, 2008, Vol. 246, No. 3, pp. 697–722. [Online]. Available: https://doi.org/10.1148/radiol.2462070712

[9] K. Loverdos, A. Fotiadis, C. Kontogianni, M. Iliopoulou, M. Gaga, "Lung nodules: A comprehensive review on current approach and management," *Annals of Thoracic Medicine*, 2019, Vol. 14, Issue 4, pp. 226–238. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6784443

[10] G. Metry, G. Wegenius, B. Wikström, V. Källskog, P. Hansell, P.G. Lindgren, H. Hedenström, B.G. Danielson, "Lung density for assessment of hydration status in hemodialysis patients using the computed tomographic densitometry technique," *Kidney international*, 1997, Vol. 52, Issue 6, pp. 1635–1644. [Online]. Available: https://doi.org/10.1038/ki.1997.496

[11] Pulmonary infiltrate. Available at: https://en.wikipedia.org/wiki/Pulmonary_infiltrate (accessed 2024, March)

[12] Collapsed Lung (Pneumothorax). Available at: https://www.pennmedicine.org/for-patients-and-visitors/patient-information/conditions-treated-a-to-z/collapsed-lung-pneumothorax (accessed 2024, March).

[13] 'Timika score' on x-rays may help identify complex TB cases. Available at: https://www.auntminnie.com/imaging-informatics/artificial-intelligence/article/15635618/timika-score-on-xrays-may-help-identify-complex-tb-cases (accessed 2024, March)

[14] A. Chakraborthy, A.J. Shivananjaiah, S. Ramaswamy, N. Chikkavenkatappa, "Chest X ray score (Timika score): an useful adjunct to predict treatment outcome in tuberculosis," *Advances in Respiratory Medicine*, 2018, Vol. 86, No. 5, pp. 205–210. [Online]. Available: https://doi.org/10.5603/ARM.2018.0032

[15] LungExpert. Available at: https://lungs.org.by (accessed 2024, March).

[16] TB DEPOT Data Dictionary. Available at: https://depot.tbportals.niaid.nih.gov/#/data-dictionary (accessed 2024, March)

[17] ImageNet. Available at: https://www.image-net.org (accessed 2024, March)

[18] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam., D. Parikh, D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," *International Journal of Computer Vision*, 2020, Vol. 128, No. 2, pp. 336–359. [Online]. Available: https://doi.org/10.1007/s11263-019-01228-7

[19] Top 7 text-to-image generative AI models. Available at: https://byby.dev/ai-text-to-image-models (accessed 2024, Apr)

[20] Rohit Kundu. Image Processing: Techniques, Types, & Applications [2023]. Available at: https://www.v7labs.com/blog/image-processing-guide (accessed 2024, Apr).

[21] M. Arsalan, M. Owais, T. Mahmood, J. Choi, K.R. Park, "Artificial Intelligence-Based Diagnosis of Cardiac and Related Diseases," *Journal of Clinical Medicine*, 2020, 9(3):871, pp 1–27. [Online]. Available: https://doi.org/10.3390/jcm9030871

[22] M.R. Zare, M. Mehtarizadeh, "An Ensemble of Deep Semantic Representation for Medical X-ray Image Classification," *2021 55th Annual Conference on Information Sciences and Systems (CISS)*, Baltimore, MD, USA, 2021, pp. 1–6. [Online]. Available: https://www.doi.org/10.1109/CISS50987.2021.9400268

[23] B. Chen, Z. Zhang, Y. Li, G. Lu, D. Zhang, "Multi-Label Chest X-Ray Image Classification via Semantic Similarity Graph Embedding," in *IEEE Transactions on Circuits and Systems for Video Technology*, April 2022, vol. 32, no. 4, pp. 2455–2468. [Online]. Available: https://www.doi.org/10.1109/TCSVT.2021.3079900

[24] V. Rostovtsev, "Intelligent health monitoring systems," *Otkrytye semanticheskie tekhnologii proektirovaniya intellektual'nykh system [Open semantic technologies for intelligent systems]*, 2023, pp. 237–240. [Online]. Available: https://libeldoc.bsuir.by/handle/123456789/51289

[25] A. Kayeshko, A. Efimova, "Decision support system for breast cancer screening," *Otkrytye semanticheskie tekhnologii proektirovaniya intellektual'nykh system [Open semantic technologies for intelligent systems]*, 2021, pp. 229–232. [Online]. Available: https://libeldoc.bsuir.by/handle/123456789/45424

ОБРАБОТКА РЕНТГЕНОВСКИХ ИЗОБРАЖЕНИЙ ГРУДНОЙ КЛЕТКИ НА ОСНОВЕ ТЕКСТОВЫХ АННОТАЦИЙ РАДИОЛОГОВ

Косарева А. А., Павленко Д. А.,Снежко Э. В.

Проанализировано более 11 000 рентгеновских снимков грудной клетки и соответствующих им текстовых аннотаций, а также проведены первые пилотные исследования по обработке изображений с учетом текстовых аннотаций специалистов-рентгенологов. Разработан конвейер обработки изображений для базы данных и нейронной сети. Проведено прогнозирование параметра «Общий процент аномального объема», для которого средняя абсолютная ошибка составила 11,073 при использовании нейросетевой модели InceptionResNet50V2.

Received 12.03.2024

Problems of Privacy and Heterogeneity for Federated Learning Applications in Medical Image Analysis

Aliaksei Himbitski
Belarusian State University
Minsk, Belarus
Email: alekseygimbickiy@gmail.com

Victor Zelenkovsky
Belarusian State University
Minsk, Belarus

Maksim Zhydovich
Belarusian State University
Minsk, Belarus

Vassili Kovalev
The United Institute of Informatics Problems of the National Academy of Sciences of Belarus biomedical image analysis laboratory
Minsk, Belarus

***Abstract*—Recently, machine learning has become one of the most promising directions in working with medical data. Deep neural network models are the most effective and accurate, but they require large volumes of information for training. This is a common problem in the case of medical data, especially images, as their creation involves significant costs.**

One solution to improve the quality of deep neural network models without increasing the training dataset is model aggregation. However, a problem arises with preserving the confidentiality of medical images. For example, if one model is trained on an image containing information about a specific patient, other models participating in the aggregation may also gain access to this information. As a result, information about a specific patient may be disclosed.

In an attempt to address the problem described above, this work aims to research and develop methods for aggregating machine learning models while preserving the privacy of medical images, particularly federated learning methods.

***Keywords*—Computer vision, machine learning, deep neural network, medical images analysis, image processing**

I. Introduction

In the modern world, deep neural networks are one of the most powerful tools for analyzing medical data, as they can extract complex relationships between different features. However, one of the main challenges in training large and very large neural networks is the computational limitations during training. This is due to the fact that training deep neural networks involves using backpropagation algorithms, which require a large number of iterations to achieve good training quality. And as the size of the neural network increases, the number of iterations required for training increases exponentially.

Thus, training large neural networks becomes highly challenging and demands significant computational resources. One way to address this problem is through model aggregation. Model aggregation in machine learning is the process of combining multiple models into a more powerful and efficient one.

II. Idea of federated learning

The idea of federated learning emerged and was first described in 2016 by researchers from Google in their paper titled "Communication-Efficient Learning of Deep Networks from Decentralized Data" [1]. This approach belongs to the domain of distributed machine learning, deployed across multiple clients, and enables training a unified global model on the server using several sources (clients), each of which is trained on its own dataset. More formally, let there be N clients C_1, C_2, ..., C_N participating in the construction of the global statistical model, each with its own dataset D_1, D_2, ..., D_N. The server coordinates the work of different clients and their training.

The process of federated learning can be divided into three key stages:

1) Initialization: At each step t, clients download the latest version of the model wt from the server.
2) Local training: Each client C_k performs iterative training based on its own local dataset D_k and a hyperparameter η. The client updates the weights of its local model after several training epochs, denoted as $w_t^k \leftarrow w_{t-1}^k$ (η, D_k), and sends them back to the server.
3) Model aggregation: The server performs the aggregation of weights received from the local models and updates the global model.

$$w_t^{global} \leftarrow Agregation(w_t^k, k \in 1, 2, ..., N) \qquad (1)$$

The goal of the entire process is to minimize the objective function, which can be expressed by the following formula:

$$min_w \sum_{k=1}^{N} p_k F_k(w), \tag{2}$$

where F_k is the local objective function for the k-th client, pk is a value reflecting the relative influence of each client, with $p_k > 0$ and $\sum_{k=1}^{N} p_k = 1$. In other words, at each step, each client updates the weights of its model, and then the server aggregates k sets of weights using a specific aggregation method (such as averaging aggregation, progressive Fourier aggregation, FedGKT, etc.). More detailed, the entire process of federated learning is shown in the figure 1.

Later, this strategy was referred to as centralized federated learning, and a decentralized federated learning strategy was also proposed. In the decentralized federated learning strategy, there is no need for a central server with which the model clients exchange data. Instead, each client individually communicates with some other clients and aggregates their updates. This strategy helps address the issue of a single point of failure, where the entire model does not break down due to isolated errors.

The main advantages of federated learning can be summarized as follows [2]:

- Scalability: The distributed nature of federated learning allows the system to easily adapt to changes in the number of participating devices.
- Model simplification: By enabling different collaborating devices to conduct multiple parallel training cycles with small amounts of data, federated learning simplifies the traditional centralized approach, where a single entity has to process a substantial volume of data each time.
- Faster convergence: By using simpler models, devices participating in federated learning can perform multiple iterations more quickly since they learn from the experiences of other devices, leading to the faster development of a reliable global model.

However, despite the aforementioned advantages, federated learning methods have limitations when it comes to the heterogeneity of data and local models. This article focuses on the analysis and solution of these specific challenges.

III. Federated Learning Algorithms

A. Federated Stochastic gradient descent (FedSGD)

In a typical machine learning system, an optimization algorithm like stochastic gradient descent (SGD) works with a large dataset uniformly distributed across servers in the cloud. The gradient is computed on a mini-batch, which is a random subset of the original data, for each step. However, in the case of federated learning, the data is distributed unevenly across millions of devices, and some devices may be unavailable at certain times.

To address these challenges, a modification of SGD called Federated SGD has been introduced. In this approach, the gradients are averaged by the server in proportion to the number of training samples on each node and used for performing the gradient descent step [3].

$$F_k(w) = \frac{1}{n_k} \sum_{i \in P_k} F_i(w) \tag{3}$$

$$g_k = \nabla F_k(w_t) \tag{4}$$

$$w_{t+1} \leftarrow w_t - \eta \sum_{k=1}^{K} \frac{n_k}{n} g_k \tag{5}$$

B. Federated averaging (FedAvg)

Federated averaging is a generalization of FedSGD that allows local nodes to perform more than one batch update on their local data before exchanging weights with other models. Now, instead of exchanging gradients, local models exchange weights. The rationale behind this generalization is that if all local models start from the same initialization, averaging gradients is strictly equivalent to averaging the weights themselves.

$$\forall k, w_{t+1}^k \leftarrow w_t - \eta g_k \tag{6}$$

$$w_{t+1} \leftarrow \sum_{k=1}^{K} \frac{n_k}{n} w_{t+1}^k \tag{7}$$

C. Federated learning with dynamic regularization (FedDyn)

In cases where the data from different client models is not homogeneous, minimizing the local loss functions of the clients may not decrease the global loss function. Therefore, it has been proposed to use regularization with the use of data statistics, such as data volume, transmission speed, and cost. This modification transforms the values of local loss functions into values of global loss functions.

D. FedProx

This method is an improvement over the previous approach and aims to mitigate the issue of local optimization inherent in stochastic gradient descent-based methods.

The problem at hand is as follows: performing numerous local iterative training steps in FedAvg can cause each client to prioritize achieving its own local objective rather than the global one, leading to suboptimal convergence or model divergence.

The solution proposed by this method involves adding a term $\frac{\mu}{2}||w_k^t - w_{global}^t||^2$ to the objective function to regulate the influence of local models and ensure convergence guarantees. When μ = 0, FedProx is equivalent to FedAvg, meaning that subsequent model aggregation and global updates are performed using the same principles as in FedAvg.

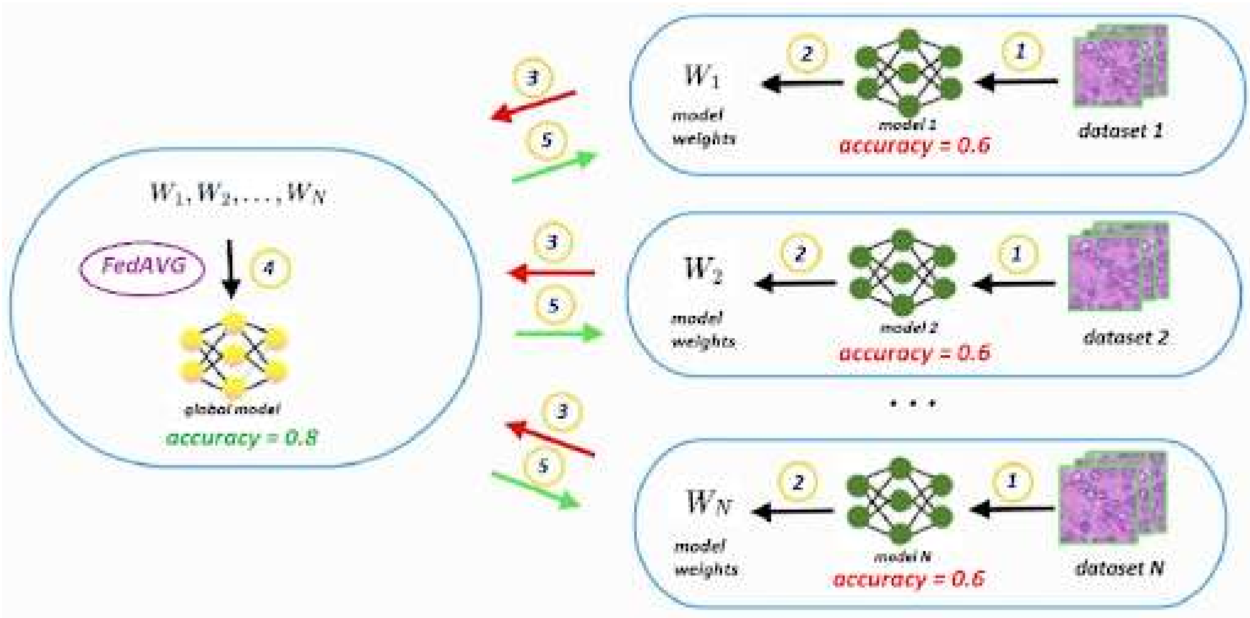


Figure 1. Diagram of the federated learning process.

E. FedNova

In the FedNova algorithm, the model aggregation stage of the FedAvg algorithm is modified to address the issue of model non-identicality. Before updating the global model, the algorithm applies normalization and scaling to the local updates from each client based on their local iteration number.

IV. The Problem of Heterogeneity in Federated Learning

Existing methods of federated learning have the following drawbacks that limit their use for medical imaging:

Data privacy concerns: During the aggregation phase, the server receives all model weights from the clients, which leads to a loss of data confidentiality. By having access to all weights and information about the model's hyperparameters, it becomes possible to reconstruct the images on which the local models were trained with some level of accuracy.

Difficulty in aggregating models with different architectures: In cases where the architectures of the local models differ significantly, the number and dimensions of weight matrices also differ greatly. This makes it impossible to apply basic federated learning methods.

Difficulty in aggregating heterogeneous data: Often, the data on clients may undergo different preprocessing. Moreover, the data may differ in class imbalance, which further complicates the use of basic federated learning methods.

To address these issues, a method is proposed that distinguishes itself from other approaches by having the server receive only a specific portion of the model weights during the aggregation phase. This feature allows for the preservation of the privacy of the local model and, consequently, the data on which it was trained.

Weight aggregation still occurs on the server, but now the server itself is a machine learning model (specifically, a neural network) trained on a similar task and data that is similar to the data used to train the local models.

1) Clients send a specific portion (not all) of the trainable weights to the server.
   $w_k^t, k \in [1, 2, ..., N]$
2) The server uses a transformation $F : R_k \to R_h$ to convert the received weights w_k^t from the R_k space (the weight space of the k-th client, where the weights transmitted by the clients may have different dimensions) to the hidden space R_h. This is done for the convenience of aggregating the weights, which are now in a unified space.
   $$w_k^t \leftarrow F(w_k^t) \tag{8}$$
3) Next, using the transformation $G : R_h- > R_h$, the server aggregates the weights obtained in the previous step as follows:
   $$w_{global} \leftarrow G(w_k^t) \tag{9}$$
4) Then, for all clients, the server applies the inverse transformation $F^{-1} : R_h- > R_k$ to convert the aggregated weights w_{global} back to the original weight space of the client R_k and sends them back to the client.

The described method has the same advantages as basic federated learning approaches, but it has an additional key advantage that other methods lacked: preserving the privacy of images.

Data leakage does not occur because clients only exchange a portion of the weights with the server, which is insufficient to reconstruct the original data on which the model was trained.

It is precisely this advantage that enables the use of this algorithm for training models used for medical images, ensuring the privacy and confidentiality of sensitive patient data.

Among the drawbacks of this approach, the following can be identified:

- The need for additional training of the server model: This requires extra time and additional data for training the server model.
- Lack of improvement in model quality when the number of weights transmitted by clients for aggregation is too small: If the amount of weight information shared by clients is insufficient, the overall model quality may not improve significantly.
- Potential degradation in model quality when the architectures of the models differ significantly: If the models used by the clients have vastly different architectures, the aggregation process may lead to a decrease in model quality instead of improvement.

V. Experiments and Results

For the analysis of the effectiveness of the developed method, a dataset of 12,000 histological images was used, divided into two classes: malignant tumors and benign tumors. The dataset consisted of 8,400 images for training and 3,600 images for testing. The training dataset was randomly divided into five parts: four clients and one server.

The training process involved a cycle of weight exchange between the clients and the server, followed by aggregation and sending back of the weights. This cycle was performed every 3 epochs of training, and a total of 10 cycles were conducted. The weights for aggregation were sent from each client. Therefore, the total number of training epochs for each client was 30.

The weights of the models for exchange (R_k space) were the weight matrices of the linear classification layer of the network, with a dimension of 1024x1024 for all clients.

The evaluation of the method was done by comparing the following values obtained with and without the application of this method during training (traditional training without weight aggregation for 30 epochs):

- The meaning of the loss function (cross-entropy) during training.
- The value of the loss function on the test dataset.
- The accuracy of predictions on the test dataset.

A. Models with the same architecture

For analyzing the effectiveness of the developed method in the case of homogeneity of local models, was used a simple neural network with the architecture shown in the figure 2:

```
        Layer (type)               Output Shape         Param #
================================================================
            Conv2d-1         [-1, 16, 222, 222]             448
         MaxPool2d-2           [-1, 16, 74, 74]               0
            Conv2d-3           [-1, 16, 72, 72]           2,320
            Conv2d-4           [-1, 32, 70, 70]           4,640
         MaxPool2d-5           [-1, 32, 23, 23]               0
           Flatten-6                [-1, 16928]               0
            Linear-7                 [-1, 1024]      17,335,296
            Linear-8                 [-1, 1024]       1,049,600
            Linear-9                    [-1, 2]           2,050
================================================================
Total params: 18,394,354
Trainable params: 18,394,354
Non-trainable params: 0
```

Figure 2. Simple neural network architecture.

As the global model (server), a pre-trained resnet18 network was used, fine-tuned on 1170 histological images for 30 epochs. For weight aggregation from clients, the last layer designed for classification was modified to the following:

For weight aggregation, a linear layer called "aggregate" is used. It takes weights from clients as input for aggregation and produces modified weights for each client as output.

In the figure 4 is a graph showing the evaluation of various quality metrics for client 0 (the graphs for other clients are similar).

Based on the analyzed data graphs, the following conclusions can be drawn about the performance of this method on simple models of the same architecture:

- For all clients, there is a decrease and stability in the values of the loss function during training when the method is applied.
- For 3/4 of the clients, there is higher stability in the accuracy of predictions on the test dataset when the method is applied.
- The average prediction accuracy did not change when the method was applied.

B. Models of various architectures

The following pretrained neural networks were used to analyze the effectiveness in the case of heterogeneity of local models:

- Client 1: SimpleModel (see above);
- Client 2: MobileNetV3 Large;
- Client 3: MobileNetV3 Small;
- Client 4: DenseNet121.

For each local model, the last layer of the neural network was replaced with the layer shown in the figure 5:

The transferred weights are the weights of the shared linear layer. The model for the server is similar to the model from the previous section.

```
(fc): Sequential(
  (fc): Linear(in_features=512, out_features=1048576, bias=True)
  (reshape): Reshape()
  (aggregate): Sequential(
    (0): Linear(in_features=1024, out_features=1024, bias=True)
    (1): ReLU()
  )
  (max_pool): MaxPool2d(kernel_size=(1024, 1), stride=(1024, 1), padding=0, dilation=1, ceil_mode=False)
  (squeeze): Squeeze()
  (classifier): Linear(in_features=1024, out_features=2, bias=True)
)
```

Figure 3. The last layer of the resnet18 network for weight aggregation.

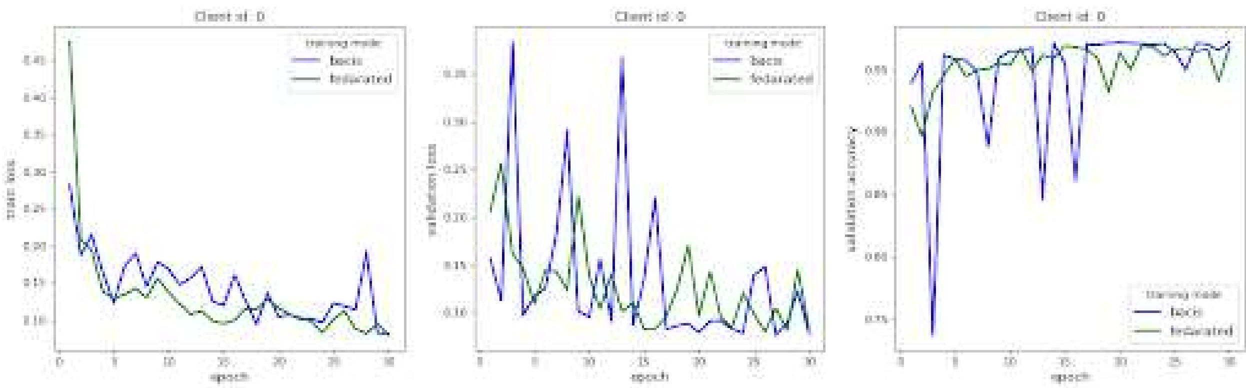


Figure 4. Graphs of analyzed metrics (SimpleModel).

```
(classifier): Sequential(
  (fc): Linear(in_features=1024, out_features=1024, bias=True)
  (relu1): ReLU()
  (shared): Linear(in_features=1024, out_features=1024, bias=True)
  (relu2): ReLU()
  (classifier): Linear(in_features=1024, out_features=2, bias=True)
```

Figure 5. The last layer of the local models.

In figures 6 and 7 are the plots showing the evaluation of various quality metrics for clients 0 and 2.

Based on the analyzed data graphs, the following conclusions can be drawn about the performance of this method on pre-trained models of various architectures:

- For all clients, there is a decrease and stability in the values of the loss function during training when the method is applied.
- There is higher stability in the prediction accuracy on the test dataset for simpler models and some pre-trained networks when the method is applied.
- On average, the prediction accuracy with the method applied has not changed.

C. Heterogeneous data

In real life, it is possible for data to be stored on different media. Additionally, the data can be heterogeneous.

To analyze the effectiveness in the case of data heterogeneity, consider a scenario where the original dataset is divided into two clients. The data of the first client contains 95% of class 0 objects and 5% of class 1 objects. Conversely, the second client has 5% of class 0 objects and 95% of class 1 objects. Architecture of client models is shown below

As the global model a pre-trained resnet18 network was used. For weight aggregation from clients, the last layer designed for classification was modified to the layer in the figure 9

During the training process of the second and third models, federated learning was used. The FedAVG algorithm was selected for aggregating information from client models.

The exchange of model weights between the client models and the server occurred every 3 epochs. Each of the client models was trained for a total of 30 epochs. The cross-entropy loss function was used.

The goal of the experiment was to study the dependence of model prediction accuracy on the partitioning of the dataset among clients.

Thus, we simulate a situation where the data from different clients is highly heterogeneous

In such a situation, it is not possible to achieve

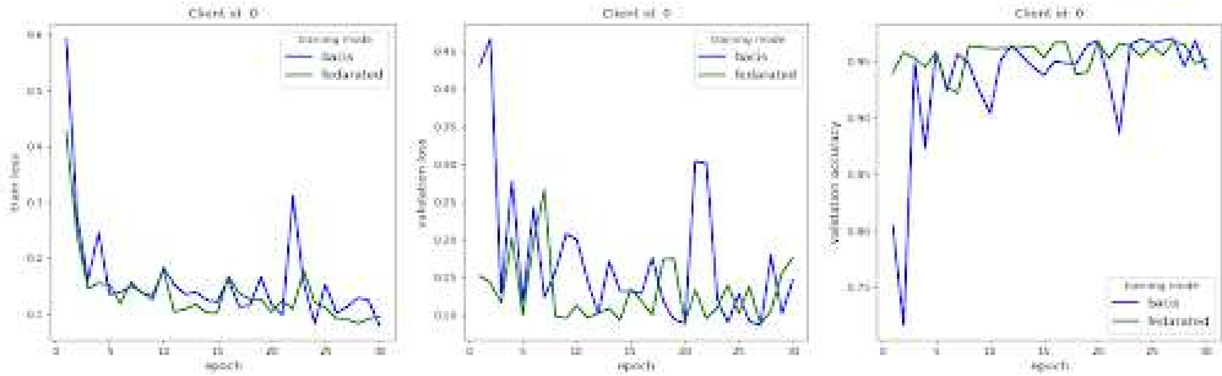


Figure 6. Graphs of analyzed metrics (SimpleModel).

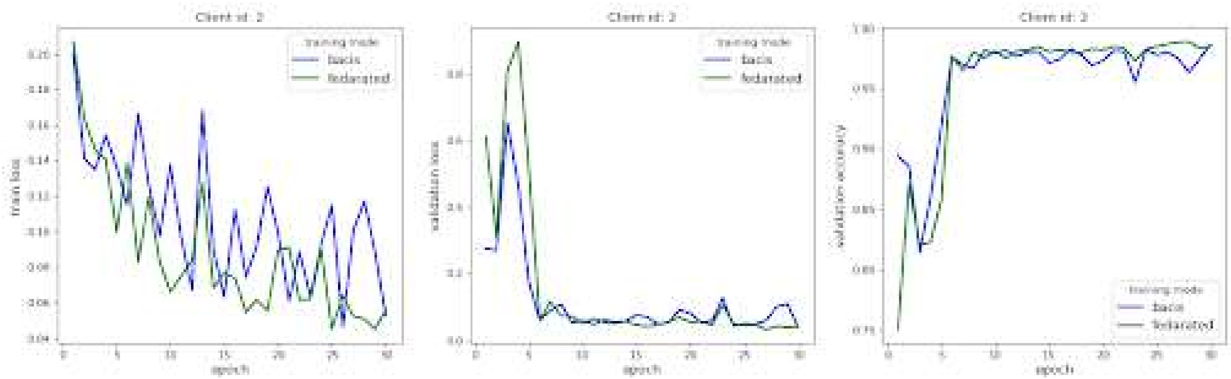


Figure 7. Graphs of analyzed metrics (MobileNetV3 Large).

```
        Layer (type)               Output Shape         Param #
================================================================
            Conv2d-1         [-1, 16, 222, 222]             448
         MaxPool2d-2           [-1, 16, 74, 74]               0
            Conv2d-3           [-1, 16, 72, 72]           2,320
            Conv2d-4           [-1, 32, 70, 70]           4,640
         MaxPool2d-5           [-1, 32, 23, 23]               0
           Flatten-6                [-1, 16928]               0
            Linear-7                 [-1, 1024]      17,335,296
            Linear-8                 [-1, 1024]       1,049,600
            Linear-9                    [-1, 2]           2,050
================================================================
Total params: 18,394,354
Trainable params: 18,394,354
Non-trainable params: 0
```

Figure 8. Client model architecture.

```
(fc): Sequential(
  (fc): Linear(in_features=512, out_features=1048576, bias=True)
  (reshape): Reshape()
  (aggregate): Sequential(
    (0): Linear(in_features=1024, out_features=1024, bias=True)
    (1): ReLU()
  )
  (max_pool): MaxPool2d(kernel_size=(1024, 1), stride=(1024, 1), padding=0, dilation=1, ceil_mode=False)
  (squeeze): Squeeze()
  (classifier): Linear(in_features=1024, out_features=2, bias=True)
)
```

Figure 9. The last layer of the resnet18 network for weight aggregation.

stable learning and good results when using sequential learning, where one client is trained first and then the other. However, by using federated learning methods, we were able to achieve the same metric values as in the case of balanced data. This is primarily because, during aggregation, the global model aims to acquire knowledge from all clients and average them.

Each of the client models learned to predict the classes present in its own dataset very well. However, when it comes to predicting classes that were rarely encountered before, the client models struggle on the test data. Nonetheless, the central model aggregated information from all the data, which can potentially improve its ability to predict such classes.

Thus, federated learning methods enable us to mitigate the impact of data heterogeneity when training models.

In the figure 10 is shown the graph of accuracy for heterogeneous data.

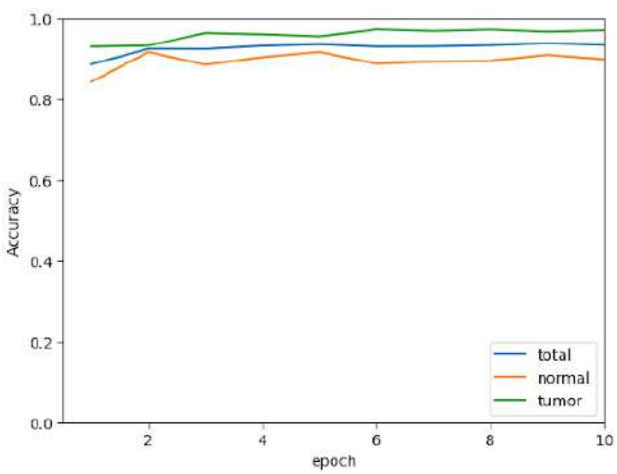


Figure 10. Graph of accuracy for heterogeneous data.

VI. Semantic technologies application

In addition to the above applications, you can use semantic technologies in federated learning. For example as follows.

- Distributed knowledge representation: Semantic technologies enable the representation of knowledge in a structured manner using formal languages. This can be useful in distributed model training, where each device can have its local knowledge representation and then combine these representations on a central server.
- Semantic data analysis: Semantic technologies can assist in the analysis and understanding of data collected from distributed devices. For example, they can be used to extract meaning and relationships between data, which can be beneficial for aggregation and merging of models on a central server.
- Unification of semantic understanding: Semantic technologies can help unify the understanding of data across different devices or servers. They can be used to create a shared model of knowledge or a semantic network, which can be utilized for harmonizing and collaboratively training models on different devices.

VII. Conclusion

In this article, the main problems arising in the task of biomedical image analysis were described, and a method to avoid them was proposed. Furthermore, experiments were conducted to demonstrate its practical applicability. Has been shown that federated learning helps preserve data confidentiality and also provides a significant improvement in quality when different clients have data from different classes. Various approaches to solving the problem of heterogeneous learning have been considered. The obtained conclusions and recommendations can be valuable for researchers in the field of biomedical informatics and medicine who aim to utilize advanced machine learning methods for image analysis in privacy-preserving conditions.

Acknowledgment

This research was supported by the United Institute of Informatics Problems of the National Academy of Sciences of Belarus (UIIP NASB).

References

[1] H. Brendan McMahan, E. Moore, D. Ramage, S.Hampson, B. Agüera y Arcas, "Communication-efficient learning of deep networks from decentralized data," Artificial Intelligence and Statistics, 1273–1282 (2017).

[2] P. Qi, D. Chiaro, A. Guzzo, M. Ianni, G. Fortino, F. Piccialli "Model aggregation techniques in federated learning: A comprehensive survey". Future Generation Computer System, 150, 272-293 (2023).

[3] Federated Learning for Mobile Keyboard Prediction, McMahan, Brendan, et al., 2017.

[4] EU. Regulation (eu) 2016/679 of the european parliament and of the council on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation). https://eur-lex.europa.eu/legal-content/EN/TXT., 2018.

[5] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communicationefficient learning of deep networks from decentralized data. In Aarti Singh and Xiaojin (Jerry) Zhu, editors, Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS 2017, 20-22 April 2017, Fort Lauderdale, FL, USA, volume 54 of Proceedings of Machine Learning Research, pp. 1273–1282. PMLR, 2017.

[6] Kairouz Peter, McMahan H Brendan, Avent Brendan, Bellet Aurélien, Bennis Mehdi, Bhagoji Arjun Nitin, Bonawitz Keith, Charles Zachary, Cormode Graham, Cummings Rachel, et al. Advances and open problems in federated learning. arXiv preprint arXiv:1912.04977, 2019.

[7] Hard Andrew, Rao Kanishka, Mathews Rajiv, Ramaswamy Swaroop, Beaufays Françoise, Augenstein Sean, Eichner Hubert, Kiddon Chloé, and Ramage Daniel. Federated learning for mobile keyboard prediction. arXiv preprint arXiv:1811.03604, 2018.

[8] Jin Yilun, Wei Xiguang, Liu Yang, and Yang Qiang. A survey towards federated semi-supervised learning. arXiv preprint arXiv:2002.11545, 2020.
[9] Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated learning: Challenges, methods, and future directions. IEEE Signal Process. Mag., 37(3):50–60, 2020.
[10] Zengpeng Li, Vishal Sharma, and Saraju P. Mohanty. Preserving data privacy via federated learning: Challenges and solutions. IEEE Consumer Electronics Magazine, 9(3):8–16, 2020.

ПРОБЛЕМЫ КОНФИДЕНЦИАЛЬНОСТИ И НЕОДНОРОДНОСТИ ПРИЛОЖЕНИЙ ФЕДЕРАТИВНОГО ОБУЧЕНИЯ ПРИ АНАЛИЗЕ МЕДИЦИНСКИХ ИЗОБРАЖЕНИЙ

Гимбицкий А. В., Зеленковский В. П., Жидович М. С., Ковалёв В. А.

В последнее время машинное обучение стало одним из самых многообещающих направлений в работе с медицинскими данными. Модели глубоких нейронных сетей являются наиболее эффективными и точными, но требуют больших объемов информации для обучения. Это общая проблема в случае медицинских данных, особенно изображений, так как их создание включает значительные затраты. Одним из решений для повышения качества моделей глубокого обучения без увеличения обучающего набора данных является агрегация моделей. Однако возникает проблема сохранения конфиденциальности медицинских изображений. Например, если одна модель обучается на изображении, содержащем информацию о конкретном пациенте, другие модели, участвующие в агрегации, также могут получить доступ к этой информации. В результате информация о конкретном пациенте может быть раскрыта.

В попытке решить описанную выше проблему, данная работа направлена на исследование и разработку методов агрегации моделей машинного обучения с сохранением конфиденциальности медицинских изображений, в особенности методов федеративного обучения.

Received 13.03.2024

Evaluation Metrics and Multi-level GAN Approach for Medical Images

Galina Kovbasa
Belarusian State University of Informatics and Radioelectronics
Minsk, Belarus
g.kovbasa@gmail.com

***Abstract*—This article examined methods for using GANs in medicine, their prospects, as well as problems with training generative adversarial networks associated with the increasing use of generated images for training other networks. The analysis of single-layer and multi-layer GANs concluded that although multi-layer GANs perform better statistically, they do not exactly match the distribution of the original dataset and, without medical supervision, such synthetic data should not be used when training new networks. Problems associated with the phenomenon of recursive learning, biased assessments of image realism, and non-optimized structures are considered. Approach is described in context of integrating generative adversarial network models into the OSTIS Technology based hybrid computer systems.**

***Keywords*—Multi-level GANs, recursive learning, synthetic data**

I. Introduction

Generative adversarial networks (GANs) are a remarkably popular technique for generating realistic synthetic data. Modern GANs can have different layers, backgrounds , complexity and be trained by semi-supervised and unsupervised learning. They gain their popularity because of the ability to implicitly modeling high-dimensional data distributions. [1] GANs are of particular interest in the processing, classification and evaluation of medical images, in the future making it possible to speed up and improve the analysis of the results of magnetic resonance imaging, computed tomography, X-rays and others. This can be solved by integrating a GAN neural networks into the OSTIS Technology. [2]

Integration with third-party technologies based on neural networks allows the development of universal hybrid systems. GANs are capable of not only processing images, but also creating new synthetic data on demand, which makes them valuable for creating datasets, anonymous educational materials, etc. GANs are actively changing during development and all these changes can be formalized in the OSTIS system using SC code. [3]

Thus, network artifacts of the processes of creation, training and further tuning, such as numbers and types of layers, weights, activation functions, can be stored in a general form and saved, supporting further replenishment and expansion of the general knowledge base. [4] The OSTIS intelligent system is also capable of saving several versions of the same model for later use, even restoration from a previous point. [5]

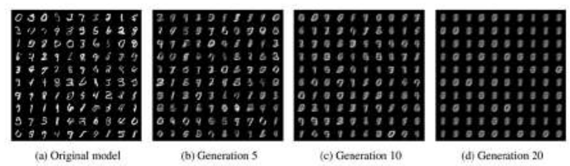


Figure 1. Model collapsing. Over generations, the generated data begins to look unimodal.

But one of the worst challenges facing GANs in medicine is that medical images are susceptible to various noise and artifacts common to different modalities and, what most important is, have a very little variety of datasets to be used. It should also be noted that much medical information is 3D structures, which can make it difficult to train a GAN on only 2D images.

A. The recursive learning problem

No generative adversarial network can reliably recreate the distribution of the original sample data. It may make mistakes or recreate the same data, that is, clone it. We must train the network so that it does not go beyond the distribution, but also does not repeat the picture. This solution allows us to avoid modern problems with GAN.

Generative adversarial networks are able to generate high-quality images on demand using the same distribution that is given, the score is also high. However, this does not mean that GAN can be used as a universal method or classifier, since it is impossible to assume that the distribution will be reliable or true. Likewise, the images produced by networks cannot be considered representative for training other networks.

However, more and more people are using GANs for reconstruction and generation, but this only leads to the fact that it can be used for a training set, since the generated materials are in the public domain. As a result, the networks degenerate, and the results of reconstruction and generation deteriorate. And as a result, the so-called mode collapse occurs (Fig. 1). [6]

This phenomenon (using generated images as training materials) is most dangerous for image generation, because in this case the networks mutually degrade. The problem of degradation of the second derivative of the loss of the original distribution leads to the fact that the learning network will take everything that the network gives without a critical attitude. It is particularly dangerous for GAN involved in the reconstruction of medical images. It is main case when a medical specialist plays a leading decisive role.

At the moment there is no complete solution to this problem. The search for ideal hyperparameters or best evaluation metric is still ongoing. Degradation can only be prevented by correct data preparation and correct interpretation of data. Even in the case of realistic results, a specialist's hand is still needed, otherwise it leads to the original problem. Partly, the problem of network degradation is related to the lack of training materials in the public domain, because all materials in one way or another relate to personal data, although they are as depersonalized as possible (but there is still a diagnosis verified by a specialist), which limits the number of networks available for training databases. In view of this fact, any decision of the GAN is only to some extent true.

B. Evaluation of GAN

1) Density estimation: Density estimation is a challenging unsupervised learning problem. Existing maximum likelihood approaches for density estimation are either limited or unable to generate high-quality samples. However, the lack of density estimation limits the application of sample data.

Density estimates are needed in a wide range of practical computer vision problems, especially when the likelihood of the generated samples is critical. This occurs, for example, when it is important to simultaneously explore and optimize the search space, when confidence estimates of a hypothesis are required, or when control over the level of generalization is important. Typical sampling quality metrics are inadequate because the generative model may simply remember the data or miss important modes.

Such methods are effective for determining whether a generative model has learned the correct statistics, but they are somewhat limited. Most techniques define the statistic to be zero if the generated GAN and the true samples belong to the same distribution. Difference between distributions is measured only on the basis of certain statistics, ignoring other. In particular, the manifold representation ignores the densities that the generator assigns to different parts of space, as well as whether the manifold is more abundant in regions around the true distribution.

Alternative density estimators, such as auto-regressive models, stream methods, or non-parametric methods such as kernel density estimation (KDE), are either too computationally intensive or require significant neural network engineering [7].

2) Log-likelihood: Log-likelihood is widely used to evaluate other families of generative models. On top of that, log-likelihood has been used before to demonstrate that a wide family of generative models assigns a greater likelihood to images outside of the training distribution. In [8] states that probability-controlled models that have much worse FID show better performance and overall distribution evaluation than state-of-the-art GANs. By evaluating GANs with low FIDs, we show that multi-level GANs are superior to single-level models in terms of average test log likelihood and generate subjectively better images on medical datasets. [9] indicates that AIS (Monte Carlo method that estimates an equation's integral by utilizing various intermediate distributions) is accurate enough to make reliable comparisons between models and can compete with other alternative density estimators. There is no guarantee that such approximation metrics will hold for real data, although [9] found that the behavior of AIS closely matches real and simulated data.

C. Multi-layer GAN

A multi-layer GAN, also can be called hierarchical or nested GAN, is a type of generative adversarial network (GAN), comprising with multiple networks of generators and discriminators organized in a hierarchical or nested structure. The main idea of multi-layer GANs is to improve the quality and variety of generated samples by using multiple layers of abstraction. By training generator and discriminator networks at different levels of abstraction, multi-layer GANs are able to learn more complex and diverse data distributions. In such a GAN, the generator network creates a sequence of samples that become increasingly refined and detailed, and the discriminator network evaluates the quality of these samples at each level of abstraction. Multi-layer GANs can be implemented using various architectures and techniques such as progressive growth, ladder networks, and recursive networks. These approaches differ in how they organize hierarchical or nested connections between networks of generators and discriminators, and in how they convey gradients and losses between layers.

II. Related work

A. Preparing of the data

In order to evaluate the realism of images based on medical data, two sets of datasets, The Automated Cardiac Diagnosis Challenge (ACDC) [10] and The Indian Diabetic Retinopathy Image Dataset (IDRID) [11], were selected to evaluate the quality of synthetic image reproduction using GAN. These datasets were chosen not only because of the high quality of the various types of images, but also because of the low number of uses in GAN research.

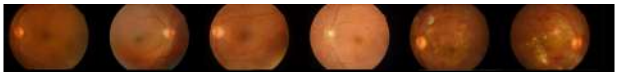


Figure 2. IDRiD images. From not affected to severe cases

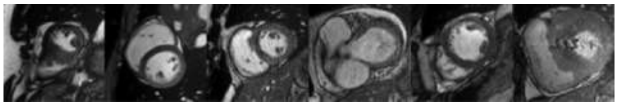


Figure 3. ACDC MRI image slices

ACDC dataset contains 150 studies of short-axis cardiac cine MRI of the University Hospital of Dijon patience, 1902 2D slices for training and 1078 2D slices for testing [10]. The 516 photos in the IDRID collection include both pathological and normal retinal fundus [11].

For the purpose of metrics analysis, we sampled a large number of images from each of our trained GANs to see how well it could learn the original data distribution. According to [12], although GANs are characterized by sensitivity to hyperparameters, a subset of GANs are characterized by their oversensitivity, as shown by the sharp change in FID score under different sets of hyperparameters.

Firstly, evaluating set of hyperparameters require some computational time to acxhive best performs for all the GANs described below.

Secondly, GANs are sensitive to the reference dataset: as it increases, image quality improves. To represent difference in GAN characteristics due to dataset size difference all estimation metrics will be evaluated on each dataset with different numbers of training samples.

Each GAN described below produced a batch of two thousand images without additional post processing.

B. *GAN approaches for realistic data generation*

1) Deep Convolutions in GAN: Deep convolutional GANs [13] were one of the first GANs to use convolutional layers and made significant contributions to balanced GAN learning. Although convolutional layers have been used in GAN architectures before, DCGAN offers an adapted architecture. Several rules have been proposed to create a stable convolutional architecture. Although convolutional networks have been used in GAN architectures before, DCGAN offers an adapted architecture. Several rules have been proposed to create a stable convolutional architecture. These rules are as follows [13]:

- Don't use merge layers. Instead, use straight line convolutions for the discriminator and fractional line convolutions for the generator.
- Use batchnorm in generator and discriminator layers.
- Do not use fully connected structures in hidden layers.
- In the output layer of the generator, use the Tanh function, in other layers - the Relu function.
- Use the LeakyReLU activation function on all discriminator layers.

Deep neural networks with a large number of parameters are very powerful machine learning systems. However, a serious problem for such networks is brute force. Large networks are also slow to use, making it difficult to combat overfitting by combining the predictions of many different large neural networks during testing. To solve this problem, the dropout technique is used [14]. Although DCGAN has a variety of advantages over non-convolutional models, in further chapter we show, that multi-level non-convolutional GANs outperforms it in terms of the quality of generative image for medical data.

2) Supervised vs Unsupervised networks: The variety of image structure is determined by the criteria of realism, which must be described, but at the end of training, such an approach is already meaningless, since the criteria need to be adjusted during the work process. Networks with this approach are called semi-supervised.

In recent years, deep generative models have dramatically pushed forward in generative modelling, achieving state-of-the-art semi-supervised learning results. Unsupervised networks also show good results, but they do not have the same advantages:

- The ability to predict H+1 classes (number of classifiers) during training time, forces class labels to be displayed. This extra class correlates with the outputs of generative models and can produce higher quality outputs. This may affect the results of the generative model. We can pass through this class to the corresponding outputs of the generative model, which allows us to improve performance and quality (fewer epochs). This can be compared to the feedback between a discriminator and a generative model.
- We can submit softmaxes of fake and original images.
- Better pictures and at the same time fewer epochs.
- Possibility of reducing the impact of type I error: fake images that are perceived as real [15].
  The common approaches that have been used in most multi-layer GANs is a progressive growth [16]

3) Progressive growth: A GAN training methodology that starts with low-resolution images and then gradually increases the resolution by adding layers to the network, as shown in the figure 4. This gradual nature of learning allows you to first detect the large-scale structure of the distribution of images (the small size of which, as noted by [16], is characterized by greater stability), and then switch attention to increasingly smaller details, refining and complementing the image, rather than studying all

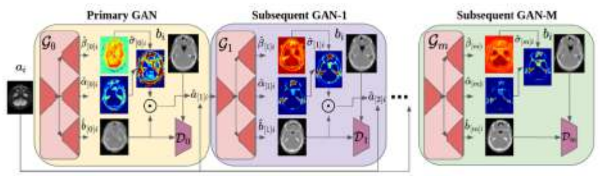


Figure 4. General architecture of Uncertainty-Guided Progressive GANs

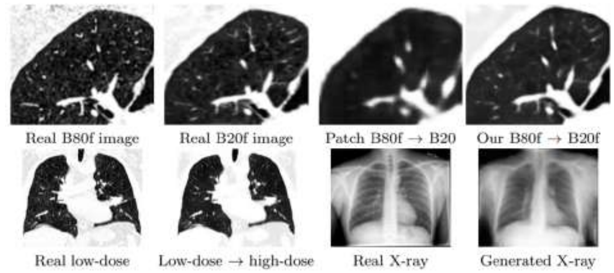


Figure 5. Results of Multi-Scale GANs (MSGANs)

scales simultaneously. Accordingly, the following advantages can be identified:

- More stable image generation in the early stages.This is due to less information about classes and fewer modes. The resolution enhancement approach simplifies the task for generating the final images. And this ultimately leads to image stabilization with reliable synthesis of results.
- Reduced training time.When using progressively growing GANs, most iterations are performed at a lower resolution, and comparable result quality is often achieved 2-6 times faster, depending on the final output resolution.
- Relatively low GPU requirement [17]. Because GANs at each scale are trained separately, the training process never exceeds a certain maximum size, so GPU requirements remain consistently low, making it easier to generate arbitrarily large images.

4) Multi-scale discriminator: The [17] used a multi-scale discriminator consisting of two parts of the same structure, which made it possible to solve problems that single-scale discrimination could not handle. The first discriminator, D1, looked at the smaller images to process the overall structure, and the second, D2, helped to reconstruct the finer details of the image. This approach allows you to train a network to find structures in an image more efficiently than single-level GANs. It is also worth noting that this way we can conclude that multi-layer GANs are inherently multi-layer discriminators. In [18], images are generated in several stages (Primary GAN, Subsequent GAN-1, ..., Subsequent GAN-N) (Figure 4).

The output of one phase serves as input to the subsequent GAN in the next phase, explicitly guided by the attention map derived from the uncertainty estimates. Using multi-level generators and discriminators, the authors achieved comprehensive correction of artifacts and noise, potentially replacing additional imaging procedures, which can reduce examination costs and time [18].

A similar approach was used in [19] and [20], where the image was also generated sequentially using enhancement modules (Fig. 6). At the same time, in [21] multi-level generators were used instead of a multilevel discriminator, since it was noted that the instability of the discriminator interferes with focusing on noise removal. This fact is important, since medical images are susceptible to various noises and artifacts characteristic of different modalities, so it makes sense to consider GAN network architectures where there are no discriminators at certain levels. The [21] architecture is discussed in more detail in a separate paragraph.

III. Material and Methods

A. Modern multi-layer GAN architectures

Unlike conventional GANs, multi-layer architectures produce higher-quality images. Since, presumably, the discriminator is the largest part responsible for the quality of image synthesis (a similar conclusion can be made based on [22]), in order to avoid errors of the second type. In the case of multi-level GANs, to improve the quality of generation, a load with additional discriminators is used. Also, a significant part of the result depends on the settings of the output layer loss function. In the case of multi-layer GANs, the result of the previous layer goes to the output layer function, which is scaled to the next layer. Thus, such an algorithm is not a multi-level synthesis, but a separation of correct characteristics from incorrect ones. The main advantage of the multi-level GAN architecture is that, unlike single-level ones, it is capable of such separation.

1) Uncertainty-Guided Progressive GANs (ProGANs): In Fig. 4 [18] shows a model consisting of cascaded GANs, where each generator is capable of estimating aleatory uncertainty as well as generating images. This solution removes the above-mentioned limitations of modern methods. Getting rid of them by modeling the underlying distribution of residuals across pixels as an independent but non-identically distributed zero-mean generalized Gaussian distribution (GGD), so $\widehat{b}_{ij} = b_{ij} + \epsilon_{ij}\epsilon_{ij} \sim GGD(\epsilon;0,\alpha_{ij},\beta_{ij}) \equiv \beta_{ij}(2\alpha_{ij}\Gamma(\beta_{ij}^{-1}))^{-1}exp(-\alpha_{ij}^{-1}|\epsilon|^{\beta_{ij}})$

2) Hierarchical GANs (HierGANs): HI-GAN [21] consists of four sub-networks, namely G_b, G_a, discriminator D_a and boost network G_c. Both G_b and G_a generators are DCNNs used for image desaturation. In addition, G_a is trained together with D_a to improve its ability to desaturate damaged images and preserve details. The advantage of G_a is its ability to solve the problem of

Figure 6. Results of Cascaded Refinement Networks (CRN)

Figure 7. General scheme for image generation using Multi-scale GANs

Figure 8. Examples of images with IS equal to 900.15

loss of high-frequency characteristics such as edges and texture, constantly competing with D_a in a repeated zero-sum game. In contrast, G_b learns on its own and does not need to compete with any network. G_b's strategy is to avoid the influence of discriminator instability and focus only on noise removal. Overall, $G_b$and Ga have different strategies and use different criteria to evaluate reconstruction performance, and neither is better than the other. For this reason, G_c is used to help G_b and Ga cooperate more efficiently and improve reconstruction efficiency. G_c takes the bleached images of G_b and Ga as input and generates a synthesized image whose PSNR is close to that of G_b and details are recovered more accurately than that of G_a.

3) Multi-scale GANs (MSGANs): As shown on Fig. 7 [20], different generator architectures were chosen for G_0 and $G_{1...n}$ because the tasks of generating low-resolution whole images and high-resolution patches differ in a number of requirements. LR GAN uses U-Net architecture, which is able to filter out many irrelevant details and generalize better due to its bottleneck. Its tendency to produce blurrier images is negligible in the context of low-resolution images. ResNet blocks were chosen to generate patches using HR GAN because they are known to produce clear results while maintaining the same resolution of the input image [20]. The risk of overkill in the absence of a bottleneck is reduced by stronger conditioning (on previous scales) and an overall higher number of patches compared to the number of images used. For discriminators $D_{0...n}$ the usual fully convolutional architecture [20] is chosen.

B. Evaluation metrics and network optimization

Let's consider popular metrics for assessing the quality of generative image models.

1) Inception Score (IS): Inception Score is a metric for automatically assessing the quality of generative image models. This metric has been shown to correlate well with human assessment of the realism of generated images from the CIFAR-10 dataset. Inception Score uses an Inception v3 network pre-trained on ImageNet and calculates statistics of the network's outputs when applied to the generated images.

$IS(G) = exp(\mathbb{E}_{x\sim p_g} D_{KL}(p(y|x)||p(y)))$ where $x \sim p_g$ means that x is an image sampled from p_g, $D_{KL}(p \parallel q)$ — KL-divergence between distributions p and q, $p(y|x)$ is the conditional class distribution, $p(y) = \int_x p(y \mid x) p_g(x)$ — marginal class distribution. exp is present in the expression to make it easier to compare values, so we will ignore it and use $ln(IS(G))$ without loss of generality [23].

In other words, IS can be interpreted as a measure of the dependence between the images generated by G and the marginal distribution of classes in y. The mutual information of two random variables is also related to their entropies:

Inception Score does show a reasonable correlation with the quality and variety of generated images, which explains its widespread use in practice. However, it is not entirely correct because it only evaluates Pg as an image generation model and not its similarity to Pr. Such gross violations as mixing in natural images from a completely different distribution completely deceive the Inception Score. As a result, this may push models to simply learn sharp and varied images (or even some unfavorable noise) instead of Pr. This also applies to Mode Score. Additionally, Inception Score is unable to detect overfitting because it cannot use the validation set [24].

Applying Inception Score to generative models trained on non-ImageNet datasets produces unreliable results. Most often, Inception Score is used for generative models trained not on ImageNet sets, but on CIFAR-10, since it is slightly smaller and more convenient for training than ImageNet.

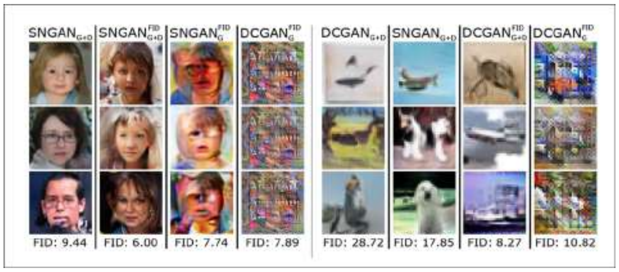


Figure 9. Example of the FID manipulation problem

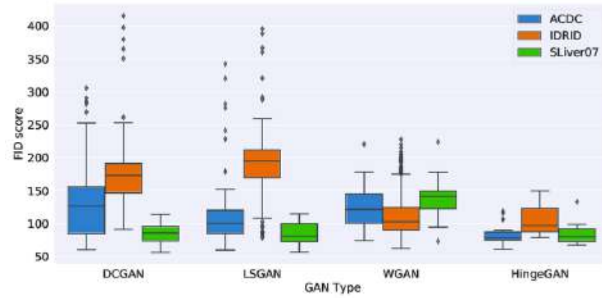


Figure 10. FID score for different GAN types

2) Frechet Inception Distance (FID): Frechet Inception Distance, or FID for short, is a metric for assessing the quality of generated images, specifically designed to evaluate the performance of generative adversarial networks. This metric was proposed as an improvement to the existing Inception Score, or IS. The Inception Score evaluates the quality of a collection of synthetic images based on how well the best Inception v3 image classification model classifies them as one of 1,000 known objects [22]. The scores combine both the confidence in the conditional class predictions for each synthetic image (quality) and the integral of the marginal probabilities of the predicted classes (diversity). The inception score does not reflect the comparison of synthetic images with real ones.

The FID score was designed to evaluate synthetic images based on the statistics of synthetic images datasets compared to the same statistics of real images from the goal distribution. As with inception estimation, FID estimation uses the inception v3 model. In particular, the model's encoding layer (the last pooling layer before output image classification) is used to capture specific features of the image, they will be generalized as a multivariate Gaussian, computing the mean and covariance of the images. These statistics are then calculated correspondingly for the collection of real and generated images.

However, it should be noted that this metric cannot fully evaluate images on an equal level with a person and it does not fully comply with human standards for image evaluation [22].

3) LL computation: As a high test LL corresponds to a low KL divergence between the generative distribution and the genuine data distribution, many generative models, aside from GANs, employ it as an additional evaluation variable. When decoder is not created to be reversible log-likelihood estimation in decoder-based models is typically intractable ([25], [26]). For GANs it is not obvious how to compute a good lower bound, unlike in the case of VAE-based models. Even when lower bounds are available, they have only been calculated in relatively few studies ([9], [27]). LL has received little attention and is never utilized specifically in GANs. [25]

Monte Carlo techniques such as AIS and nonparameteric density estimation methods such as KDE get around this by assuming a Gaussian observation model $p\theta(x|z)$ for the generator. In particular, [28] demonstrated that LL of GANs may be precisely approximated via annealed significance sampling. The AIS implementation in this study adheres to the [29] approaches. We applied a Metropolis-Hastings (MH) adjustment and 10 leapfrog steps with the HMC transition operator. The optimal 65 percent acceptance rate for HM [29] was used to tune the HMC. AIS algorithm used 16 independent chains, approximately 8000 intermediate distributions. It was found that AIS is precise enough to enable thorough comparisons between generative models for the majority of the models we looked at.

IV. Results

We used AIS to estimate log-likelihoods for all models under consideration, also was calculated FID and Inception metrics for all GANs specified at tables below. As the result, the log-likelihood assigned to the GAN does not correlate FID or Inception scores. As GAN tend to favor images with larger regions, with less local variance, there is little to no correlation between actual distribution of GAN generated images and common statistics approaches. All AIS results from represented GANs are listed in the table I. All FID and Inception results, if calculated from represented GANs, are listed in the table below. Examples of generated images from multi-layer GANs are shown in the figure 11.

V. Discussion

Training a GAN requires significant hyperparameter tuning and powerful computing resources. With the OSTIS system, it has become possible to automate the selection of the optimal neural network architecture and protects the core of the intelligent system from the formation of redundant approaches and confusing terms.

Neural network models as components of the ostis system are accelerated by reducing the number of parameters used. In the case of optimization and complexity reduction, it is sufficient to determine the maximum upper value of the number of neurons of each layer without the need to select these parameters during a

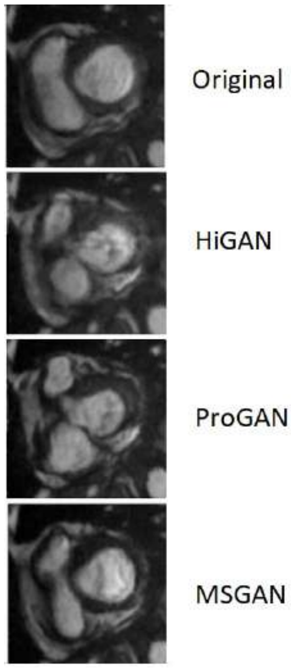


Figure 11. Examples of synthetic images from different nets

series of experiments. Deep models can often be slow and resource-intensive, which impacts overall system performance, leading to severe lags, especially in the absence of efficient and powerful computers and video adapters.

However, the trained model may not produce the desired result when considering the subsequent tasks for which the generated data is intended. Even if ProGAN or MSGAN achieve better results, they can produce unreliable images. There is no objective way to evaluate whether a medical image actually involves in gathering data responsible for the diagnosis with GAN approach without involvement of a medical expert. Due to the smaller scale of medical datasets not every GAN architecture is adapted to capture specific features of medical images.

The OSTIS technology promotes and supports integration of various neural network models. This system helps user achieve compatibility between different AI systems, effectively use knowledge to create code-based transcriptions and graphical interpretations of models. [2]

The OSTIS system also simplifies network development, expands functionality by using and reusing existing designs. This technology is capable to summarize results of a GAN training and testing and store different model versions.

VI. Conclusions

As a result of the literature review, it was concluded that the realism of the generated GAN images largely depends on the distribution of the original training data

Table I
AIS ACDC results

(Nats)	AIS Test	AIS Train
DCGAN	448.43±7.40	447.23±2.24
HiGAN	349.17±4.51	546.73±5.08
ProGAN	571.64±10.22	721.37±8.76
MSGAN	651.65±9.10	732.42±2.43

Table II
AIS ACDC results.
Evaluation for half of 2D slices

(Nats)	AIS Test	AIS Train
DCGAN	348.25 ± 6.32	444.13 ± 2.24
HiGAN	294.21 ± 7.52	446.73 ± 9.38
ProGAN	571.64 ± 10.22	652.50 ± 7.5
MSGAN	551.65 ± 9.74	732.42 ± 2.43

and how similar the output results are to the training ones. Viewing different parameters at different levels allows you to determine the most optimal ones for training multi-level GANs.

It was determined that multi-scale GANs are characterized by sensitivity to changes in hyperparameters, and some multi-scale GANs are hypersensitive to such changes. The main advantages for integrated computer ostis-systems that appear when using reduction as a way to reduce the dimensionality of neural networks are determined. The authors see the further development of the proposed approach in obtaining practical results for known deep architectures of models used to solve problems of computer vision and natural language processing.

The multi-layer GAN architecture coupled with a progressive learning method allows you to better perform with medical dataset, showing improved results in popular evaluation metrics such as FID and Inception. Having considered the problem of recursive learning, we can conclude that it is impossible to obtain a real data distribution using GSN. And when using generated

Table III
AIS IDRiD results

(Nats)	AIS Test	AIS Train
DCGAN	343.11±6.07	352.45±4.7
HiGAN	289.96±5.22	498.21±9.32
ProGAN	434.71±12.54	551.15±6.13
MSGAN	532.23±6.56	678.02±15.21

Table IV
FID and Inception results

Score	FID (IDRiD)	Inception (IDRiD)
DCGAN	77.34	-
HiGAN	66.54	-
ProGAN	27.21	721.10
MSGAN	32.8	832.82
	FID (ACDC)	Inception (ACDC)
DCGAN	59.34	-
HiGAN	46.54	-
ProGAN	47.21	755.16
MSGAN	37.8	712.45

images to train networks, it leads to complete degradation of the neural network. This problem can only be solved with the help of human control of neural networks. Another problem with CT scans data, which is obtained in 3D: common GANs may not produce a good quality image if trained solely on 2D images. This makes the study of GANs specifically built for medical data an interesting avenue of research and could lead to improvements in quality and ultimately clinical usability. Thus, we can conclude that it is impossible to obtain a real pseudo-real image of a medical nature without adjusting their structure, like in multi-level discriminators, and evaluate proper metrics to correctly estimate generated distribution to make reliable synthetic data.

References

[1] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath Generative adversarial networks: An overview. *IEEE signal processing magazine*, 2018, vol. 35, no 1, pp. 53-65.

[2] V. V. Golenkov, N. A. Gulyakina, and D. V. Shunkevich, *Otkrytaya tekhnologiya ontologicheskogo proektirovaniya, proizvodstva i ekspluatatsii semanticheski sovmestimykh gibridnykh intellektual'nykh komp'yuternykh sistem [Open technology of ontological design, production and operation of semantically compatible hybrid intelligent computer systems]*. Minsk, Bestprint, 2021, (In Russ.).

[3] (2024, Feb) Ostis applications. [Online]. Available: https://github.com/ostis-ai/sc-web

[4] (2024, Feb) OSTIS. [Online]. Available: https://github.com/ostis-ai

[5] M. Kovalev. Versioning model of neural network problem-solving methods in intelligent systems *Open Semantic Technologies for Intelligent Systems (OSTIS)*, 2023, Vol. 7, pp. 121–126.

[6] I. Shumailov, Z. Shumaylov, Y. Zhao, Y. Gal, N. Papernot and R. Anderson. The curse of recursion: Training on generated data makes models forget. *arXiv preprint arXiv*:2305.17493, 2023.

[7] M. E. Abbasnejad, Q. Shi, A. V. D. Hengel and L. Liu A generative adversarial density estimator. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10782-10791

[8] Ethan Fetaya, Jörn-Henrik Jacobsen, Will Grathwohl, and Richard Zemel.Understanding the limitations of conditional generative models.*arXiv preprint arXiv*:1906.01171, 2019.

[9] Y. Wu, Y. Burda, R. Salakhutdinov, and R. Grosse. *On the quantitative analysis of decoder-based generative models. arXiv preprint arXiv*:1611.04273, 2016.

[10] O. Bernard, A. Lalande, C. Zotti, F. Cervenansky, et al. Deep Learning Techniques for Automatic MRI Cardiac Multi-structures Segmentation and Diagnosis: Is the Problem Solved? *IEEE Transactions on Medical Imaging*. 2018. Vol. 37, no. 11, pp. 2514-2525.

[11] P. Porwal, S. Pachade, R. Kamble, M. Kokare, G. Deshmukh, V. Sahasrabuddhe, F. Meriaudeau. Indian Diabetic Retinopathy Image Dataset (IDRiD). *IEEE Dataport*. 2018.

[12] Y. Skandarani, P. M. Jodoin and A. Lalande Gans for medical image synthesis: An empirical study. *Journal of Imaging*, 2023, vol. 9, no 3, pp. 69.

[13] A. Öcal and L. Özbakır Supervised deep convolutional generative adversarial networks. *Neurocomputing*, 2021, no 449, pp. 389-398.

[14] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever and R. Salakhutdinov Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 2014, vol. 15, no 1, pp. 1929-1958.

[15] V. Wilmet, S. Verma, T. Redl, H. Sandaker and Z. Li A Comparison of Supervised and Unsupervised Deep Learning Methods for Anomaly Detection in Images. *arXiv preprint arXiv*:2107.09204, 2021.

[16] T. Karras, T. Aila, S. Laine and J. Lehtinen Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv*:1710.10196, 2017.

[17] J. He, X. Li, N. Liu and S. Zhan Conditional generative adversarial networks with multi-scale discriminators for prostate MRI segmentation. *Neural Processing Letters*, 2020, vol. 52, no 2, pp. 1251-1261.

[18] U. Upadhyay, Y. Chen, T. Hepp, S. Gatidis and Z. Akata (2021) Uncertainty-guided progressive GANs for medical image translation. In Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24 (pp. 614-624). *Springer International Publishing*.

[19] Q. Chen and V. Koltun Photographic image synthesis with cascaded refinement networks. In Proceedings of the IEEE international conference on computer vision, 2017, (pp. 1511-1520).

[20] H. Uzunova, J. Ehrhardt, F. Jacob, A. Frydrychowicz and H. Handels (2019). Multi-scale gans for memory-efficient generation of high resolution medical images. *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17*, 2019, Proceedings, Part VI 22, pp. 112-120. *Springer International Publishing*.

[21] D. M. Vo, D. M. Nguyen, T. P. Le and S. W. Lee HI-GAN: A hierarchical generative adversarial network for blind denoising of real photographs. *Information Sciences*, 2021, 570, 225-240.

[22] Y. Yu, W. Zhang and Y. Deng Frechet inception distance (fid) for evaluating gans. *China University of Mining Technology Beijing Graduate School*, 2021.

[23] S. Barratt and R. Sharma A note on the inception score. *arXiv preprint arXiv*:1801.01973, 2018

[24] Q. Xu, G. Huang, Y. Yuan, C. Guo, Y. Sun, F. Wu and K. Weinberger An empirical study on evaluation metrics of generative adversarial networks. *arXiv preprint arXiv*:1806.07755, 2018

[25] L. Dinh, D. Krueger, and Y. Bengio. Nice: non-linear independent components estimation. *arXiv preprint arXiv*:1410.8516, 2014.

[26] L. Dinh, J. Sohl-Dickstein, and S. Bengio. Density estimation using Real NVP. *arXiv preprint arXiv*:1605.08803, 2016.

[27] L. Theis, A. van den Oord, and M. Bethge. A note on the evaluation of generative models. *arXiv preprint arXiv*:1511.01844, 2015.

[28] R. B. Grosse, Z. Ghahramani, and R. P Adams. Sandwiching the marginal likelihood using bidirectional monte carlo. *arXiv preprint arXiv*:1511.02543, 2015.

[29] Neal, R. M. MCMC using Hamiltonian dynamics. *Handbook of markov chain monte carlo*, 2011, Vol. *2*, no 11, pp. 2

МЕТРИКИ ОЦЕНКИ И ПРИМЕНЕНИЕ МНОГОУРОВНЕВЫХ GAN ДЛЯ МЕДИЦИНСКИХ ИЗОБРАЖЕНИЙ

Ковбаса Г. А.

В данной статье были рассмотрены методы использования GAN в медицине, их перспективы, а также проблемы в обучении генеративно-состязательных сетей, связанные с увеличением использования сгенерированных изображений для обучения других сетей. Анализ одноуровневых и многоуровневых GAN пришел к выводу, что, хотя по статистике многоуровневые GAN работают лучше, они не совсем соответствуют распределению исходного набора данных, и без наблюдения медицинского специалиста такие синтетические данные не следует использовать при обучении новых сетей. Рассмотрены проблемы, связанные с явлением рекурсивного обучения, предвзятыми оценками реалистичности изображений и неоптимизированными структурами. Подход описан в контексте интеграции моделей генеративно-состязательных сетей в гибридные компьютерные системы на основе технологий OSTIS.

Received 13.03.2024

Generation of Description for Eye Fundus Desease

Elena Himbitskaya
Belarusian State University
Minsk, Belarus
Email: fpm.gimbicka@bsu.by

Vadim Ermakov
Belarusian State Medical University
Minsk, Belarus
Email: ermakouvv@gmail.com

Alexander Nedzved
Belarusian State University
United Institute of Informatics Problems NAS Belarus
Minsk, Belarus
Email: Anedzved@bsu.by

***Abstract*—This research proposes an algoritm for detecting and evaluating signs of optic nerve disc degeneration of various genesis. This algorithm is based on the analysis of a diagnostic protocol, which is built by extracting a semantic map from an ocular fundus image and analyzing the characteristics of object classes associated with the manifestations of opticoneuropathies.**

***Keywords*—automated model, ocular fundus, deep learning, convolutional neural network, segmentation**

I. Introduction

Optic nerve atrophy, accompanied by destruction of nerve fibres and progressing death of retinal ganglion cells (RGCs), is one of the most common causes of decreased visual function, up to complete blindness. Despite the variety of clinical manifestations, optic atrophy has a common feature — a change in the colour saturation of the optic nerve disc (OND). It has been discovered that this sign is most often detected before the appearance of functional disorders associated with the disease. In this case, pallor of the optic disc and replacement of its typical pink shades with grey tones turns out to be an indicator of a latently running disease [1]. The above is fully applicable to multiple sclerosis (MS) — a chronic progressive autoimmune-inflammatory and neurodegenerative disease characterized by the formation of multiple lesion foci mainly in the white matter of the central nevrous system (CNS) and progression of focal and diffuse brain atrophy [2]. It is shown that multiple sclerosis is accompanied by death of retinol ganglion cell axons and loss of visual functions. Early diagnosis of MS significantly increases the chances of long-term remission under the influence of medications that change the clinical course of the disease.

Various threshold methods measuring the degree of pallor of the RPE have been used to diagnose MS. However, when attempting to obtain generalized color characteristics, the objectivity of the results decreases, because this approach does not eliminate the dependencies on the settings and accuracy errors of the equipment used for photoregistration of the ocular fundus. The reliability of the result, in fact, remains within the limits of reliability of the results obtained by subjective, visual evaluation of the color of the optic nerve disc, which is currently used in practice [3].

Threshold method is quite suitable in cases of pronounced atrophy of the optic nerve disk, when the diagnosis is without a doubt. However, it is difficult to apply it in controversial cases when degenerative changes in the area of the neuroretinal rim of the optic nerve disc (pallor) have only initial or poorly expressed character. Meanwhile, detection of such changes is of special interest because, as it was mentioned above, degenerative changes can be detected much earlier than functional manifestations of pathologies (pallor of the neuroretinal rim) that caused these changes [4].

Fixation of the lesion zones, determination of their informatively significant characteristics [5] (which, apart from changes in color and shape of the OND, should include changes in the pattern of retinal vessels) can be performed with the help of the most advanced current methods of computer mathematics that use artificial intelligence — deep learning neural networks and mathematical models of machine learning [6].

The objective of the study is to develop an automated method for detecting and evaluating signs of optic disc degeneration of various genesis based on the analysis of a diagnostic protocol based on the spectrum of possible manifestations of optic neuropathies recorded on the surface of the optic disc [7]. The diagnostic protocol is compiled by extracting the semantic map of the ocular fundus image and analyzing the characteristics of class objects (manifestations of opticoneuropathies). The result of applying the voting algorithm to the diagnostic protocol of the image is then compared with the result of the developed neural network model that gives the assessment-diagnosis of multiple sclerosis.

II. Signs of different classes of neuropathies on ocular fundus images

Acquired atrophy can be due to two causes - compression of peripheral optic neurons (so-called primary form) or optic disc edema (so-called secondary form) [8].

A. Primary form

The primary or glaucomatous form of optic atrophy develops due to collapse of the lattice plate of the sclera which happens due to increased eye pressure. Glaucomatous optic neuropathy in the initial phase either does not cause pallor of the optic nerve disc at all, or the disc color change is very slight and can be spotted only in one segment. However, the development of the disease leads to the loss of ganglion cells, thinning of the optic disc and narrowing of the vessels of the eye fundus, as well as desolation and death of small vessels on the disc (Kestenbaum's symptom). A gradual thinning of the optic disc is happening. The pallor becomes more even, spreading over the entire surface of the disc. However, neither its shape nor size changes. A distinctive feature of glaucomatous neuropathy is the excavation zone increasing in size.

Although the growth of the excavation zone indicates the possible presence of pathology. But the size of excavation in healthy people is also variable: as a rule, a larger optic disc has a larger excavation zone, therefore, it is more important to assess the ratio of the optic disc size to the excavation zone. An increase in this ratio is most often evidence of abnormal contraction of glial cells. An important sign of pathology may be not only a change in the relative size of excavation, but also a difference in these parameters for two eyes.

B. Secondary form

The secondary form of atrophy is based on a pathological process in the optic disc itself, in which nerve fibers are replaced by neuroglia. In this case, there is a pronounced loss of pink shades of the optic disc neuroretinal rim, we can see disc swelling, its size increases, and its contours lose sharpness of borders.

Analysis of optic disc color should not be performed over the entire area, but only within the neuroretinal rim located in the area between the edge of the optic disc and the excavation zone. In this area ganglion cell axons are localized. In norm the color is pink and its change is a sign of retinal ganglion cells death. Blood vessels located in the optic disc area should also be excluded from the analysis of the color of the neuroretinal rim. Their selection as separate objects is necessary not only because they will influence the generalized color characteristics of the neuroretinal rim, but also because the condition of blood vessels is a diagnostic sign in itself. A decrease in the number of small blood vessels on the disc surface (Kestenbaum's sign), as well as weakening (thinning) of blood vessels around the disc may indicate the presence of pathology.

III. ResNet50 classifier for pathology detection

To realize the qualitative performance of the ResNet [9] classifier, the input parameters are tensors with dimensionality 16 x 256 x 256 x 256 x 3, where 16 is the batch size, 256 x 256 are the spatial dimensions of the images compatible with the neural network architecture, and 3 corresponds to the number of image channels. The model is trained and tested on a dataset of images obtained by semantic segmentation of the original images and containing classes of disk and excavation area and a region around the disk equal to its radius. The model contains 49 convolutional layers and defines 2 classes of pathology. Model's architecture is shown in Fig. 1.

Binary classification is performed using the Sigmoid function. After a fully connected layer, the values are brought to the range [0, 1], which means the probability of disease. A threshold = 0.5 is used for decision making and subsequent comparison with the true image label. The model was trained for 20 epochs with a gradual decrease in the learning rate parameter using torch.optimize.lr_scheduler. On the test data, the model concludes the diagnosis with 94% accuracy. Table I shows the error matrix for the test data on 203 patients.

Thus Recall of the model is about 0.95 and Precision is about 0.93.

Table I
Results as error matrix

Class	***Result***
TP (correctly predicted patient)	96
FN (incorrectly predicted healthy)	5
FP (incorrectly predicted sick)	7
TN (correctly predicted healthy)	101

IV. Semantic formation of disease description

According to our research to identify informative features for further analysis (as it is shown in Fig. 2) we should segment:

- neuroretinal rim;
- the excavation zone;
- blood vessels.

To form a semantic map of the image of informative objects (the optic nerve disk, excavation zone and vascular network) was chosen a convolutional neural network of Unet [10] architecture with a resnet18 backbone pretrained on the ImageNet dataset.

The model contains 23 convolutional layers and consists of convolutional (encoder) and up-convolutional (decoder) parts. To reduce each 64-component vector to the required number of classes, 1×1 convolutions are applied on the last layer. The input image size is determined by the need for even values of height and width for adequate application of subsampling operation (2×2 max pooling).

The network is trained by stochastic gradient descent based on the input images and their corresponding segmentation maps (masks). Applied function, soft-max

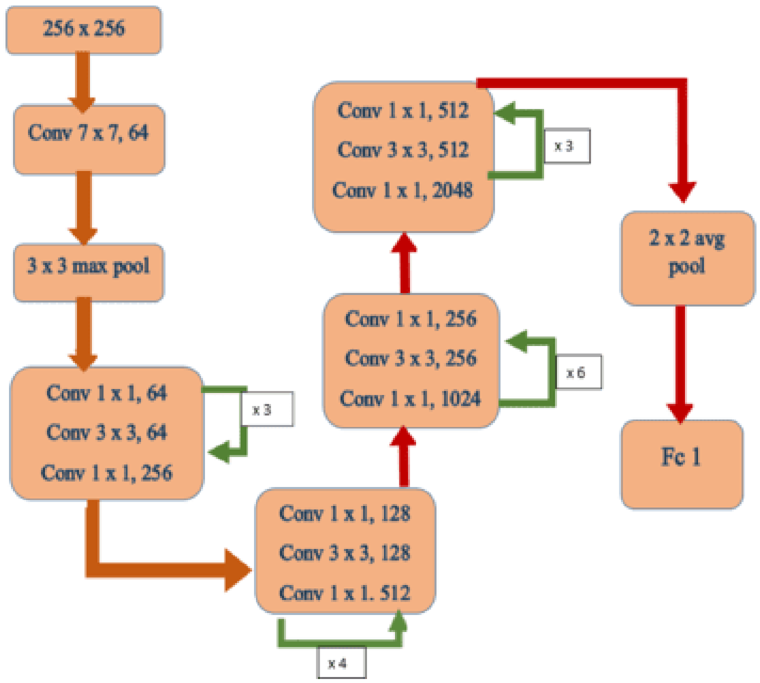


Figure 1. Classifier architecture.

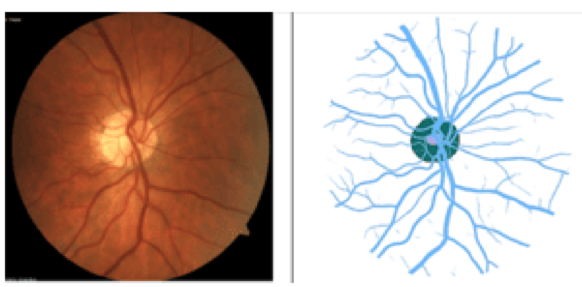


Figure 2. Semantic cart of an image.

brings the model prediction to the mask view. The loss function is a binary cross-entropy + jaccard functions. The accuracy is calculated by the BinaryIOU() [11] function, which finds the ratio of the correctly predicted mask to the union of the predicted and true masks.

A. Segmentation of blood vessels

The training of the model extracting the vascular network of the image was carried out in 2 stages. In the first stage, the network was trained on an additional set of 300 labeled data from publicly available datasets such as DRIVE [12], CHASE DB1 [13] and HRF [14]. In the second stage, the network was trained on target images.

Initially, the analyzed three-channel (RGB) image was compressed to a size of 996 x 996. After that, it was split into 9 slices with a resolution of 352 x 352 so that each slice captures a part of the neighboring slices (10 pixels). This is to eliminate distortion at the boundary between two tiles. Vessel segmentation by the neural network was performed for each tile. After the tiles were merged into a single image with the boundary 5 pixels cropped on each of them. They were then merged into rows. To smooth the transition between two tiles, their 5-pixel boundaries are overlaid and the resulting brightness is calculated using alpha blending to obtain a smooth transition. This process is shown in Fig. 3.

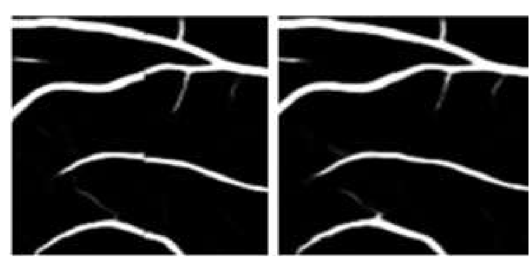


Figure 3. left — result of segmentation model on neighboring tiles, right — tiles merged with alpha blending

The same way the obtained 3 rows of tiles are combined into a vessel mask of the whole image. After that the obtained mask is stretched to the size of the original image.

The results of segmentation model for vascular network on training set and validation set are shown in Fig. 4 and their losses on training and validation set are shown in Fig. 5.

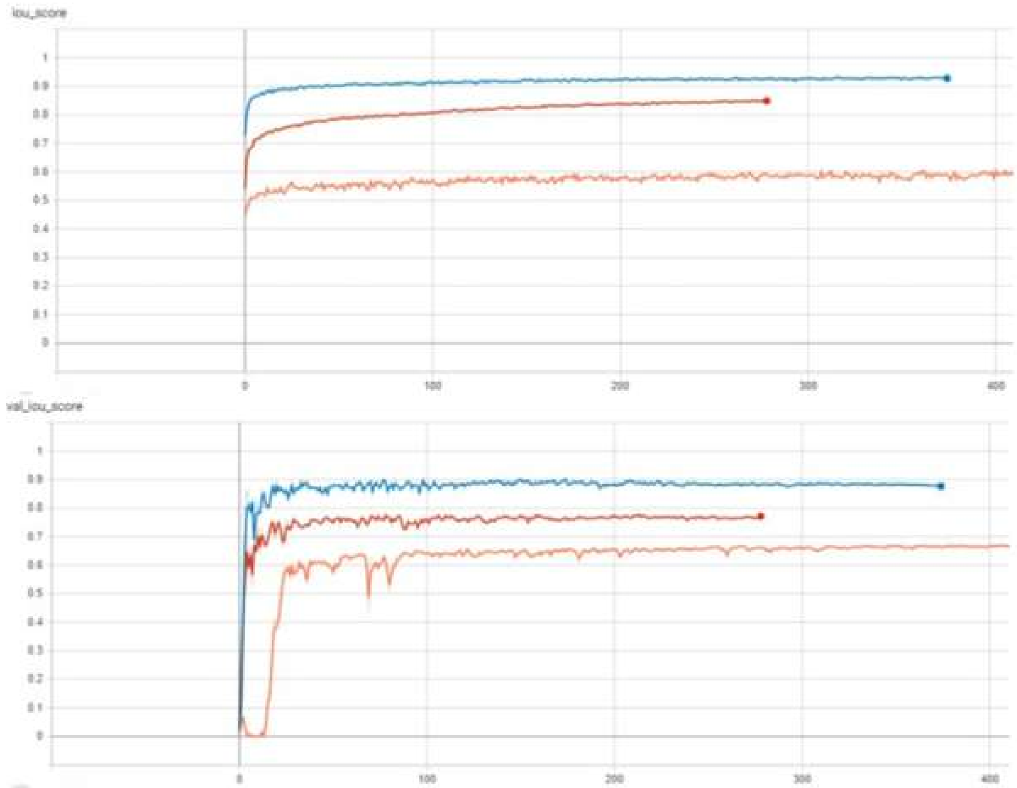


Figure 4. IoU score. first — training, second — validation. Blue — disk segmentation, red — excavation zone, orange — blood vessels

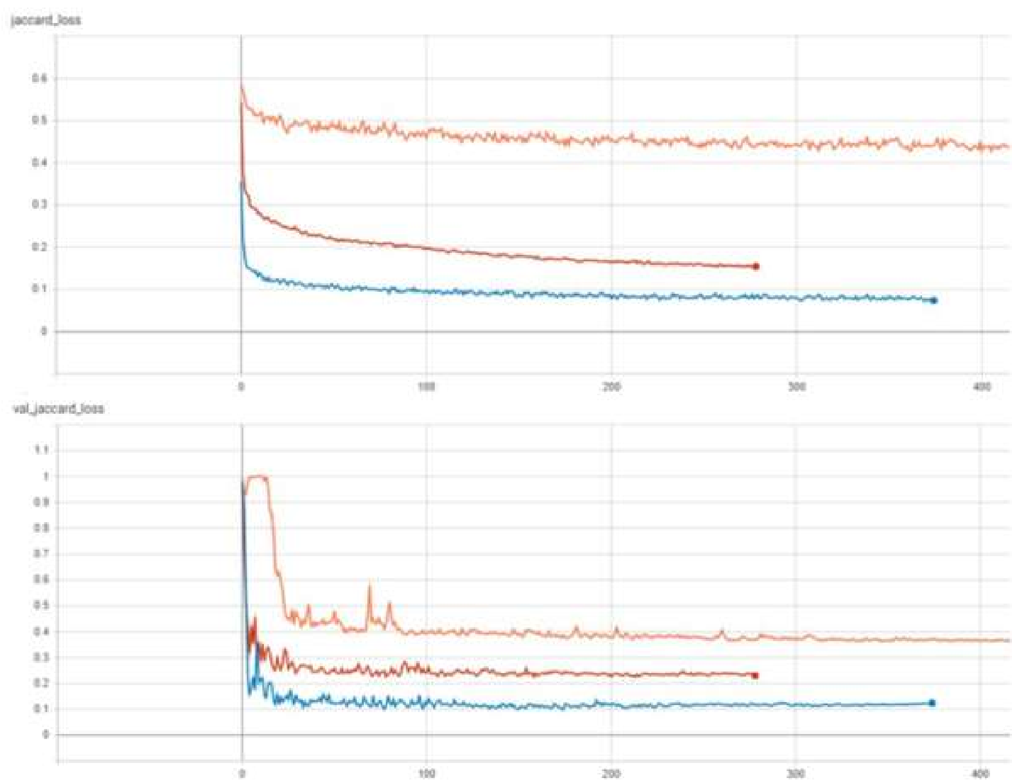


Figure 5. Jaccard loss. first — training, second — validation. Blue — disk segmentation, red — excavation zone, orange — blood vessels

B. Segmentation of optic nerve disk and excavation zone

Initially, the optic disc itself is segmented on the ocular fundus images compressed to a size of 352 x 352 using the same binary segmentation neural network (Fig. 6). The approximate radius of the disk is calculated, then the area of the original image containing the disk and having dimensions equal to three diameters of the disk is cut out. This image is transformed and used for segmentation of the excavation zone (Fig. 7).

The results of disk and excavation zone segmentation model on training set and validation set are shown in Fig. 4 and their losses on training and validation set are shown in Fig. 5.

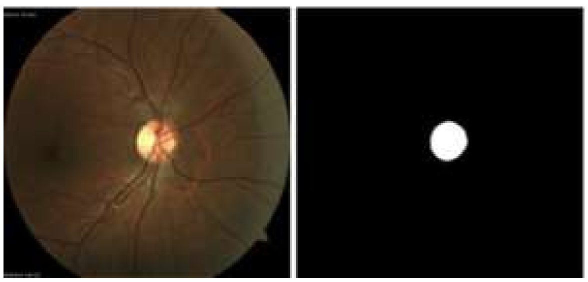


Figure 6. Image and mask of a disk

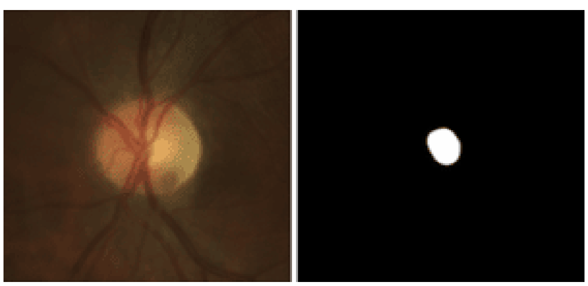


Figure 7. Image and mask of excavation zone

C. Signs of pathology manifestation

From the semantic segmentation map of these objects we need to extract such signs as:

- loss of clarity of the optic disc boundaries;
- enlargement of the cup in relation to the disk;
- Dying off of small vessels on the disk;
- branching characteristics of vessels.

Also from the provided image, should be extracted information about color signs of neuroretinal rim (disc area without excavation zone):

- graying of the neuroretinal rim (parameter H in HVS color model);
- graying of the neuroretinal rim (parameter S in HVS color model).

We should keep in mind that there is no clear relationship between these signs and the presence of early stage pathology, while at later stages of the disease the manifestation of primary optic atrophy increases, but perhaps the whole collection of signs will allow us to make some conclusions. Fig. 8 shows a scheme of the proposed automated method using the extracted features.

Now let's enlarge upon informative features and methods of their extraction. We will also consider methods of numerical expression of these signs.

1) Loss of clarity of the optic disc borders: Let's assume that this sign is manifested by the fact that the contour of the optic disc is clearer for images without pathology. To estimate the boundary clarity, we use the estimate of the intensity gradient on the contour and the variance of this gradient (Fig. 9), which is expected to be lower on the images with pathology. Specifically, we:

- extract the optic disc boundary obtained using the semantic map of the whole image as the image to be evaluated;
- apply the gradient operator, in our case the Sobel operator, to the boundary;
- calculate the total intensity gradient using the Euclidean norm of gradients along the horizontal and vertical axes.

2) The enlargement of the cup (excavation zone) relative to the disk: To estimate the enlargement of the cup (excavation zone) relative to the disk using information from the semantic map (Fig. 10), we calculate the ratio of the area of the excavation zone to the area of the optic nerve disk.

3) Dying off of small vessels on the disk: For this feature on the mask of the vascular network obtained from the semantic map, we will find and count the number of small vessels.

In this method, we will use the built-in functions of the cv2 library that remove "noise" in the image.

Let's assume that we call "small" vessels those that are less than 5 pixels wide. Then, taking them as noise, we use a kernel of 5 x 5 pixels to traverse the vascular network mask obtained from the semantic map and paint all small vessels. The obtained mask with only large vessels will be subtracted from the mask of all vessels and we will get only small vessels (Fig. 11).

4) Vascular network area: Calculating the area occupied by vessels can provide insight into the density and distribution of vessels. The method of measurement:

- counting all pixels that make up the vascular system. This can be done by simply counting the number of mask pixels that have a value corresponding to the vessel class;
- convert this number of pixels to area, given the scale of the image.

5) Vessel branching:

- we apply a skeletonization operation to the selected vessels to obtain their skeletal representation. In this

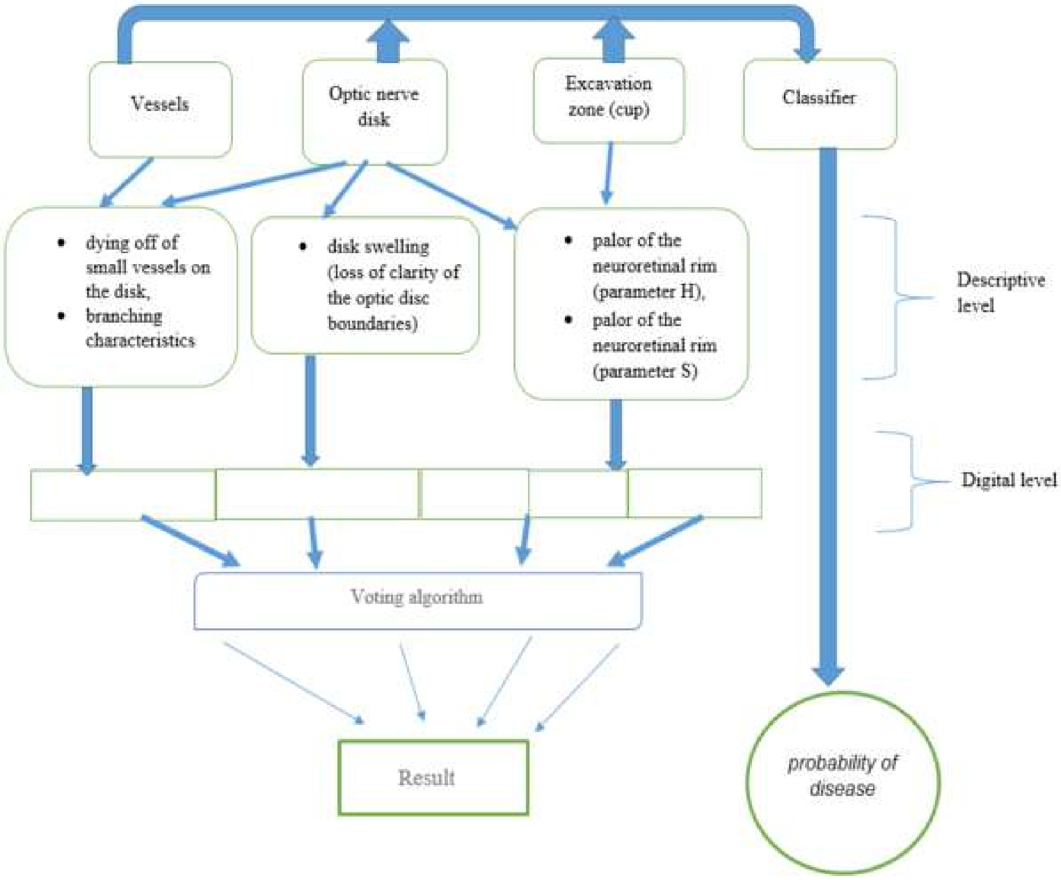


Figure 8. Pipeline of the automated method of semantic description generation for eyes diseases

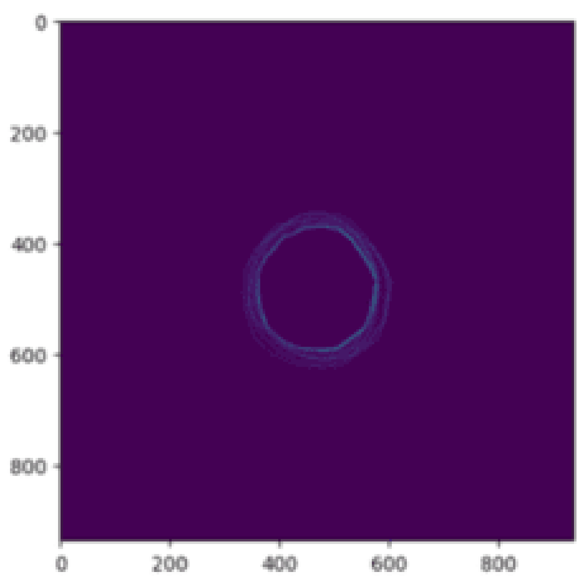


Figure 9. Mean gradient is 0.48. Gradient variance is 51

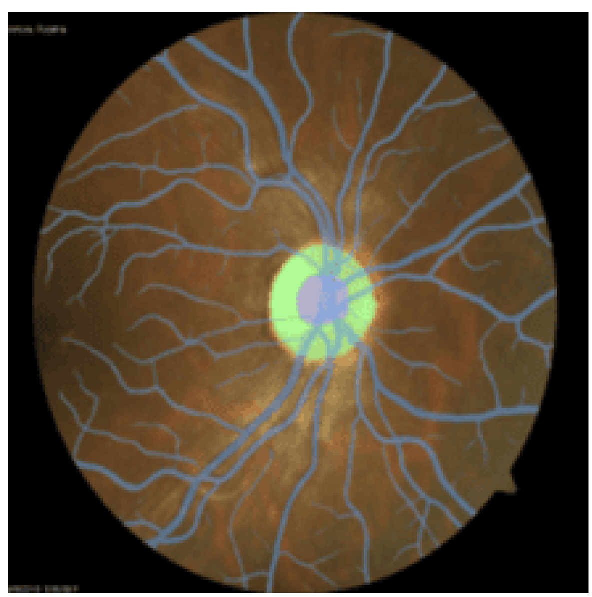


Figure 10. All classes highlighted

way, each vessel will be compressed to a width of 1 pixel;

- then for each pixel on the vessel skeleton, we perform the following steps:

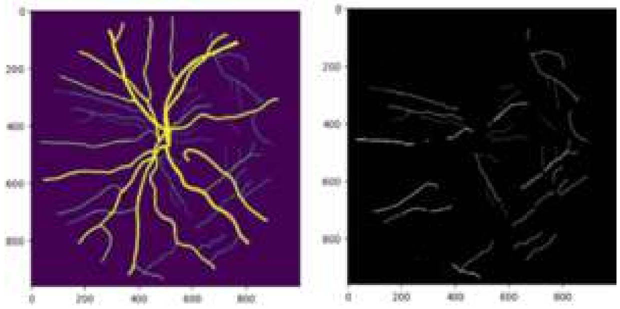


Figure 11. Small vessels extraction

  - define the 9-neighbor neighborhood of the pixel;
  - count the number of neighboring "active" pixels in the neighborhood;
  - if the number of neighboring "active" pixels is more than 3, it means that there is a branching of vessels in this place;
  - mark this pixel;
- since this algorithm can mark one branching several times, we will consider the number of connected marked components as the number of branching.
  Fig. 12 illustrates skeletonized vessel mask and it's branching points

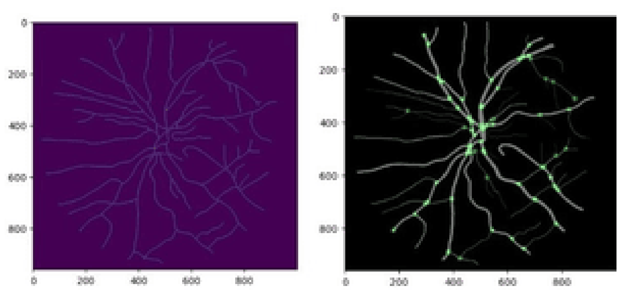


Figure 12. Number of branching points: 46

V. Conclusion

Systems capable of making semantic description for eyes diseases based on feature analysis represent a powerful tool for detecting and classifying pathologies. Such systems can be useful for monitoring and detecting early stages of different diseases.

However, we understand that the method requires additional research and validation on more diverse clinical data. It is also important to keep in mind that the final decision and diagnosis should always be based on a combination of data, expertise and clinical context.

Acknowledgment

The authors would like to thank the group of Laboratory of Information and Computer Technologies of Belarusian State Medical University for their valuable comments and help with data collecting, in particular, Grigory Karapetyan and Kosik Ivan.

References

[1] Ambika Selvakumar. Understanding Optic Disc Pallor - Shades of White. Mode of access: https://www.eophtha.com/posts/understanding-optic-disc-pallor-shades-of-white. – Date of access: accessed 24, Jan.

[2] Massimo Filippi, Amit Bar-Or, Fredrik Piehl, Alessandra Solari, Sandra Vukusic, Maria A. Rocca Multiple sclerosis. *Nature Reviews Disease Primers*, 2018, volume 4, Article number: 43.

[3] Shahzaib Iqbal, Tariq M. Khan, Khuram Naveed, Syed S. Naqvi, Syed Junaid Nawaz Recent trends and advances in fundus image analysis: A review. *Computers in Biology and Medicine*, 2022, Volume 151, Part A.

[4] Starovoitov V. V., Golub Y. I., Lukashevich M. M. Digital fundus image quality assessment. *System analysis and applied information science*, 2021, Volume 4, pp. 25–38.

[5] Bonilha VL. Age and disease-related structural changes in the retinal pigment epithelium. *Clin Ophthalmol*, 2008 Jun, vol. 2, no. 2, pp. 413–424. [Online]. Available: doi: 10.2147/opth.s2151

[6] Rohit Thanki A deep neural network and machine learning approach for retinal fundus image classification. *Healthcare Analytics*, 2023, vol. 3, 100140. [Online]. Available: doi: https://doi.org/10.1016/j.health.2023.100140

[7] Sarkar P, Mehtani A, Gandhi HC, Dubey V, Tembhurde PM, Gupta MK. Atypical optic neuritis: An overview. *Indian J Ophthalmol*, 2021 Jan, vol. 69 (1). [Online]. Available: doi: 10.4103/ijo.IJO_451_20

[8] T. S. Teleusova, M. M. Lepesova, A. L. Glaznye proyavleniya rasseyannogo skleroza [Ocular manifestations of multiple sclerosis]. *Vestnik AGIUV [Bulletin of the ASMRI]*, 2014, №3, pp 46-50.

[9] Liang, Jiazhi Image classification based on RESNET. *Journal of Physics: Conference Series*, 2020, 1634, 012110. [Online]. Available: doi: 10.1088/1742-6596/1634/1/012110

[10] Suri, Jasjit and Bhagawati, Mrinalini and Agarwal, Sushant and Paul, Sudip and Pandey, Amit and Gupta, Suneet and Saba, Luca and Paraskevas, Kosmas and Khanna, Narendra and Laird, John and Johri, Amer and Kalra, Manudeep and Fouda, Mostafa and Fatemi, Mostafa and Naidu, Desineni UNet Deep Learning Architecture for Segmentation of Vascular and Non-Vascular Images: A Microscopic Look at UNet Components Buffered With Pruning, Explainable Artificial Intelligence, and Bias. *IEEE Access*, 2022.[Online]. Available: doi: 10.1109/ACCESS.2022.3232561

[11] Adrian Rosebrock, Deep Learning for Computer Vision, PyImageSearch, 2017, 330 p.

[12] Paperswithcode. Available at: https://paperswithcode.com/dataset/drive

[13] Paperswithcode. Available at: https://paperswithcode.com/dataset/chase-db1

[14] Paperswithcode. Available at: https://paperswithcode.com/dataset/hrf

ГЕНЕРАЦИЯ ОПИСАНИЯ ДЛЯ ЗАБОЛЕВАНИЯ ГЛАЗНОГО ДНА

Гимбицкая Е. В., Ермаков В. В., Недзведь А. М.

В данной работе предлагается алгоритм для детекции и оценки клинических проявлений рассеянного склероза на снимках фундус-камерой глазного дна. Так как симптомы заболевания проявляются на различных объектах снимка, с помощью создания семантической карты исходного изображения, для каждого отдельного объекта оцениваются проявления симптомов. В результате работы данного алгоритма составляется диагностический протокол снимка. Далее полученный протокол можно сравнить с результатом модели-классификатора, работающей только с областью диска и оценивающей вероятность заболевания.

Received 19.03.2024

Intelligent Analysis in Text Authorship Identification

Ilya Trukhanovich and Anton Paramonov
Belarusian State University of Informatics and Radioelectronics
Minsk, Republik of Belarus
ilya.trukhanovich@gmail.com, a.paramonov@bsuir.by

***Abstract*—The problem of text authorship identification is considered. A generalized review of the current state of the problem is provided. A solution based on the modification of ensemble machine learning methods is proposed. The possibilities of applying sophisticated computational methods such as natural language processing, machine learning algorithms and stylometric analysis to identify individual writers based on their distinctive linguistic models are investigated. A hypothesis is put forward about the need for multidimensional text analysis and a new approach is proposed for identifying the authorship of the text in the form of a hybrid intelligent system.**

***Keywords*—authorship identification, semantics, natural language processing, machine learning, quantum technologies**

I. Introduction

Assigning authorship to texts that are disputed or anonymous requires the identification of text authors, which is a basic problem in linguistics. Conventional approaches depend on human analysis and linguistic knowledge, but with the development of intelligent analysis methods, there is an increasing interest in using computational tools to improve and automate this process. Researchers are looking into novel ways to reliably identify writers based on their writing traits by utilizing machine learning models, natural language processing (NLP) methods, and stylometric aspects.

The use of intelligent analytic techniques has attracted a lot of attention in the field of text authorship identification because of its potential to improve the precision and effectiveness of author identification based on writing styles. This paper explores the potential of using mining methods to identify text authorship, with the goal of highlighting the prospects, challenges, and possible uses of these modern techniques. A hypothesis is discussed about how intelligent analysis can transform text authorship identification and open the door for more trustworthy and robust attribution techniques in linguistics through a thorough investigation of the body of research and case studies.

II. Current authorship identification usage for plagiarism detection

A. General

In academics, publishing, and other businesses, plagiarism detection is an essential procedure for guaranteeing the uniqueness and integrity of written content. It is comparing a given text with a sizable database of previously published content using specialized tools and procedures to find any instances of plagiarism or unoriginal content. Plagiarism detection software can identify possible instances of purposeful or inadvertent plagiarism by comparing the content to other sources.

Detecting plagiarism is essential for sustaining academic integrity, credibility in research and publications, and a writing culture that values originality and ethics. It supports integrity, authenticity, and respect for intellectual property rights by assisting publishers, educators, researchers, and content producers in spotting and dealing with cases of plagiarism.

Plagiarism detection uses various techniques, such as citation analysis, machine learning (ML), natural language processing, text comparison algorithms, etc. and using these techniques the software analyzes the document's text structure, linguistic structure, citation style and other texts characteristics which are more accurate to its originality may be considered [1].

Programs for text scanning can be used to detect plagiarism. Using these tools, users can upload documents or enter text for research purposes. A report that displays any matching information is generated and stored in users' databases. By reading these reports, users can identify potential piracy cases and take necessary actions to correct the problem.

Properly crediting sources, properly paraphrasing, using quotations when quoting directly, committing to originality in writing and citing any borrowed ideas well is also important to help people develop good writing habits that encourage originality, stealing inadvertently prevents it.

B. Plagiarism detection approaches

Based on how they identify plagiarism, methods and systems for doing so can be categorized into many groups [2].

At the moment, there are various researches in this direction, which include analyzing current achievements [3].

These are a few of the principal approaches and strategies:

- Matching methods are useful for plagiarism detection because they can identify comparable strings in documents.
- Syntax-based methods match words based on their grammatical structures to find similarities in documents. It

works well for finding exact duplicates, but may not work well for altered texts with the same concept.

- The external method of plagiarism detection uses reference cards from which pieces can be extracted verbatim. Screens suspicious documents for adjacent or similar excerpts in the reference database and sends the results to the person in charge for review.
- Citation pattern analysis with citation patterns can be a strong marker of semantic similarity between documents when compared using this method. Comparability is determined by the size of the citations in the sample and by the similarity of their structure and/or scope.
- Style measurement analyzes an author's specific writing style to identify possible patterns of copying. This works well for spotting annotated and masked plagiarism, but may not work for highly annotated or translated pieces.
- Intrinsic plagiarism detection systems detect plagiarism just look at the text that has to be assessed; they don't compare it to other documents. Their objective is to identify any alterations in an author's distinctive writing style that might be signs of possible copying.

At the same time, authorship identification methods need to be modified in order to advance the field, increase efficiency, handle complicated features, adjust to changing textual environments, increase accuracy, and withstand adversarial tactics. These changes open the door to stronger and more reliable authorship attribution techniques in the fields of text analysis.

Further development in solving the problem of authorship identification seems to be the use of an integrated approach, taking into account the capabilities of various promising areas, for example, in the form of hybrid text analysis systems.

III. Modifications of machine learning methods for authorship identification and plagiarism detection

A. *Ensemble methods*

Advanced strategies known as ensemble machine learning integrate several models to increase prediction performance and accuracy. By combining the advantages and diversity of multiple models, these techniques produce better outcomes than any one model could on its own. For complicated situations, the main objective of ensemble approaches is to improve predictions while lowering bias and variance.

There are several types of ensemble methods.

Bagging is the process of training several models — typically of the same kind — on several training dataset subsets. By randomly selecting replacement samples from the original dataset, these subsets are formed, allowing for the appearance of some samples in a subset more than once and others not at all. For regression issues, the final prediction is determined by averaging the forecasts, and for classification problems, it is determined by majority vote. The goal of bagging is to prevent overfitting and lower variation. The Random Forest method is a well-known example of bagging applied to decision trees.

By using boosting techniques, a series of models is trained so that each model tries to fix the mistakes caused by the models before it. The models are added one after the other, with greater weight given to data samples that earlier models mispredicted, causing later models to concentrate more on challenging cases. A weighted total of all the models' forecasts makes up the final forecast. By eliminating bias, boosting creates a powerful prediction model from a number of weaker ones. Gradient Boosting Machines (GBM), XGBoost, and AdaBoost are a few examples of boosting algorithms.

Stacking is the process of training several models on the same dataset, then combining their predictions using a different model—often referred to as a blender or meta-model. The meta-model uses the predictions of the basic models, which were trained on the entire dataset, as input characteristics to determine the final prediction. By integrating the predictions of several models in a useful way, stacking can greatly increase forecast accuracy while maximizing the benefits of each underlying model.

Ensemble methods have several advantages:

- Enhanced accuracy. When dealing with intricate issues involving noisy data, ensemble approaches frequently yield better results than single models.
- Decreased overfitting. By averaging the predictions of several models or concentrating on hard-to-predict samples, strategies like bagging and boosting can lower the chance of overfitting.
- Flexibility. Ensemble approaches are applicable to a wide range of machine learning tasks, such as regression and classification.

Ensemble methods have several disadvantages:

- Enhanced complexity. Compared to single models, ensemble models are more difficult to comprehend, apply, and justify.
- Computing cost. Ensemble approaches are more expensive to run since training numerous models takes more time and computer resources.
- Overfitting. Some ensemble techniques, particularly boosting, have the potential to cause overfitting on the training set if not used properly.

In conclusion, by combining different models, clustering methods are effective machine learning tools that can significantly improve the prediction performance. They have several advantages, such as reduced overloading and increased accuracy, but also disadvantages, such as high computational cost and complexity

B. *Quantum-inspired methods*

Machine learning with quantum inspiration is a recent development that combines classical machine learning techniques with quantum mechanical concepts. Utilizing mathematical algorithms and insights from quantum physics, this multidisciplinary approach seeks to improve the performance and capabilities of machine learning models without the need for actual quantum computing hardware

Quantum-inspired machine learning is characterized by the use of quantum concepts in classical computational frameworks, which separates it from advanced quantum machine learning (QML) It shows current developments and practical and advanced applications in research areas, including dequantized algorithms and tensor network simulations — Strives to provide a comprehensive review of induced machine learning, clarifies its implications, and highlights the potential for further learning

Machine learning inspired by quantum includes new techniques based on digital simulations of certain quantum objects. These approaches provide new ways to train machine learning models, potentially improving productivity and efficiency.

Quantum state discrimination techniques using classical algorithms such as k-nearest neighbors are quantum inspired machine learning mainly used in classification problems This method has shown the ability to reduce complexity and increase accuracy in classification problems

In addition to focusing on real-world applications, the discipline also delves into the concepts underlying quantum-inspired machine learning. This includes studying the theoretical implications of fusing quantum mechanics and machine learning and analyzing how quantum materials can improve machine learning models

Quantum-inspired machine learning is expected to further enrich the discipline as it evolves to incorporate more ideas from traditional machine learning, quantum computation, and quantum mechanics.

Quantum-enhanced machine learning offers new ways to improve machine learning models and solve challenging problems. It's an interesting hybrid of quantum physics and machine learning. Driven by theoretical research and real-world applications, the future of this interdisciplinary field is bright.

IV. Proposed example of classification model for authorship identification

A. General

We proposed method at the nexus of classical and quantum machine learning is an ensemble machine learning classifier that integrates a quantum component and distinguishes and translates features for different kinds of classifiers when summarizing the voting results in the final solution.

Authorship attribution problem is a tuple (A, K, Q), where A is the set of candidate authors, K is the set of reference (known authorship) texts, and Q is the set of unknown authorship texts. For each candidate author $a \in A$, we are given $K_a \subset K$, a set of texts written by a. Each text in Q should be assigned to exactly one $a \in A$. From a text categorization point of view, K is the training corpus and Q is the test corpus.

The approach is showed on Figure 1.

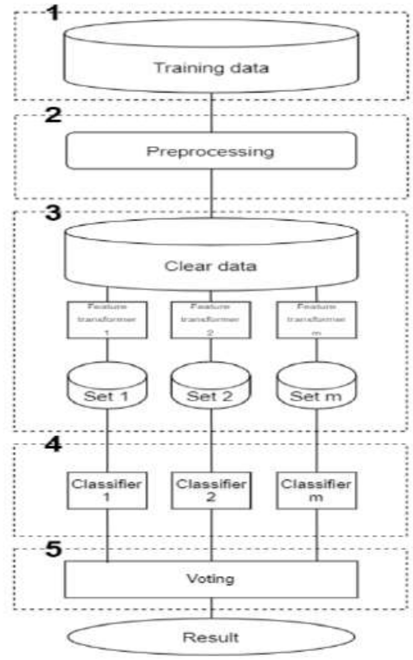


Figure 1. Proposed model architecture.

B. Description

This model has five components. Figure 1 shows them. Also this model has three main attributes.

The first attribute is feature translation and differentiation. This ensemble classifier uses a technique to translate and distinguish characteristics for various classifier types inside the ensemble. By customizing features to each classifier's unique capabilities, the ensemble model's overall prediction power is improved.

So, after preprocessing of training data we take several types of text features (lexical, statistical and so on), create feature sets and submit them to appropriate classifiers. Then we collect the results of the classifiers and pass them to the voting component. This component make decision about text ownership.

The second one is integration of quantum-inspired components. Using quantum principles or methods to process data or reach judgments as a group could constitute the quantum component of this ensemble classifier. By utilizing quantum phenomena this integration of quantum elements seeks to increase classification efficiency and accuracy.

The third one is voting results summarization. Following the predictions generated by each classifier in the ensemble, these results are combined using a voting method. The weighted majority voting algorithm is a technique for compiling expert forecasts. Expert forecasts based on weighted voting are aggregated. In order to get a final conclusion, it combines several forecasts, dynamically adjusting the expert weights according to the accuracy of each prediction.

C. Results

The model was tested in different contexts including general datasets for authorship identification and custom sets of students works.

Table I and Table II show the averaged results.

Table I
Averaged results for general datasets

Number	Classifier	Accuracy
1	Decision tree	0.38
2	K-Nearest Neighbors	0.53
3	Random forest	0.62
4	Gradient boosting	0.51
5	Proposed model	0.68

Table II
Averaged results for students works

Number	Classifier	Accuracy
1	Decision tree	0.41
2	K-Nearest Neighbors	0.46
3	Random forest	0.56
4	Gradient boosting	0.52
5	Proposed model	0.65

V. Multidimensional text analysis which includes semantics

A. General

Using statistical and computational methods, multidimensional text analysis, which includes semantics, examines the connections and patterns between different linguistic elements in a document. This method seeks to capture not just the syntactic and structural aspects of words and phrases, but also their meaning and context by integrating semantic analysis.

Researchers can find co-occurrence patterns of linguistic elements such as sentiment, topic, genre, and authorship style in multidimensional text analysis. This makes it possible to comprehend the text's features more thoroughly and provides subtle insights into the text's meaning and content.

Through the utilization of methodologies such as factor analysis, cluster analysis, and topic modeling, scholars are able to examine texts from several angles, unveiling latent patterns and structures that would not be discernible through conventional text analysis approaches. This approach will be particularly useful in identifying authorship in cases where the semantic component of the text must be taken into account.

B. Perspective of semantic analysis in authorship identification

Automatic NLP actively uses semantic analysis and subject domain ontology. The technique of comprehending natural language literature by the extraction of meaningful data from unstructured sources, such as sentiments, emotions, and context, is known as semantic analysis. It entails dissecting sentences' grammatical construction, including how words, phrases, and clauses are arranged, in order to ascertain the connections between independent concepts within a given context. These approaches are currently under active development and offer a wide range of potential opportunities [4].

A subject domain ontology is a collection of ideas and classifications inside a certain domain that illustrates their characteristics and connections. It is a type of knowledge representation that aids in the arrangement and structuring of data related to a certain area. A common vocabulary and conceptual framework are provided by ontologies for a variety of applications, including NLP.

When subject domain ontology and semantic analysis are combined, NLP systems perform better because the text is understood more accurately. Domain-specific ontologies enable NLP systems to analyze and understand text more effectively within a given context, improving the accuracy and applicability of the findings [5].

For example, in sentiment analysis, topic domain ontology can offer a systematic representation of the domain-specific concepts and relationships, while semantic analysis can assist in comprehending the meaning of words and their context. This may result in the classification and interpretation of sentiment more accurately.

Subject domain ontology and semantic analysis can be crucial parts of NLP systems. They offer an organized depiction of concepts and relationships unique to a given subject and aid in comprehending the content and context of text. NLP systems can analyze and interpret texts more accurately and relevantly by combining these strategies [6].

The applied use of semantic analysis should be considered from several approaches: software and hardware.

VI. Potential semantic analysis approaches: software

Text meaning is taken into consideration throughout the semantic text classification process. It comprehends context, sentiment, and word relationships in addition to just matching keywords. Semantic approach can use different methods.

Bidirectional encoder representations from transformers (BERT), a type of deep learning model, has been found to outperform conventional machine learning techniques in a variety of text classification tasks, such as question answering and sentiment analysis. By taking into account the complete context in which words appear, these models are able to capture intricate semantic patterns in text.

With multi-level semantic features text is classified by combining the extraction of semantic information at several levels, including keywords, local context, and global context. To collect keyword semantic information, for instance, enhanced TF-IDF based on category correlation coefficients can be employed, and neural network models with attention mechanisms, such as TextCNN and BiLSTM, can capture local and global semantic information, respectively.

Text meaning and class descriptions are compared via semantic matching. When there are few labeled examples but thorough class descriptions, this can be especially helpful in few-shot text categorization.

Text classification logic components involve classifying text using logical reasoning and rules.

In rule-based classification text is categorized using a set of pre-established logical rules. For example, the text can be categorized based on the presence or absence of specific keywords or phrases.

Logic programming and machine learning can be combined in inductive logic programming. It can be applied to build classification models that include logical rules representing domain knowledge.

Probabilistic logic models address text classification uncertainty by fusing probability theory and logic. They can make assumptions about a text's propensity to belong in a particular class based on specific feature presence.

More reliable text classification systems can result from the integration of logic and semantic components. For instance, a system may employ logical reasoning to improve categorization based on domain-specific rules after using deep learning to comprehend the semantics of text. Text categorization is known to create issues due to high dimensionality and low density of text representation, which this integration can aid with.

In essence, the application of sophisticated NLP techniques to comprehend the subtle meaning of text and the application of logical reasoning to apply classification rules are the foundations of the semantic and logic components of text classification. This combination makes it possible to create complex models that correctly classify text in a wide range of intricate situations.

For this purpose, a proper review is necessary so that it is possible to understand what features are currently available.

VII. Potential semantic analysis approaches: hardware

A. General

Apart from hardware meant for general-purpose computing and software solutions, there are also specialist hardware solutions made for text processing. The special needs of text processing jobs, like quick data processing, quick response times, and effective resource management, are catered for in these hardware solutions. These are a few instances of specialized text processing hardware.

Integrated circuits created specifically for a given purpose or task are known as Application-Specific Integrated Circuits (ASIC). ASICs may be made to work very well and with minimal power consumption on a variety of text processing applications, including sentiment analysis, keyword extraction, and regular expression matching. Real-time text analysis and NLP are two examples of text processing activities where ASICs can be especially helpful. These tasks call for high-speed data processing and low reaction times.

Field-Programmable Gate Arrays are hardware components that can be programmed to carry out particular functions. Text processing jobs requiring high throughput and low latency, like NLP and real-time text analysis, are especially well suited for FPGAs. Certain text processing activities, such regular expression matching and keyword extraction, can be efficiently and power-efficiently carried out by FPGAs through programming.

Hardware components called Tensor Processing Units (TPUs) are specially made for deep learning and machine learning applications. Matrix multiplications and other operations frequently found in machine learning models, such neural networks, are optimized for TPUs. With its ability to drastically reduce training and inference times for machine learning models, TPUs are especially helpful for large-scale text processing applications like sentiment analysis and text classification [7].

Clusters of computers specifically built for high-performance computing workloads are known as high-performance computing (HPC) clusters. Multiple networked computers, or nodes, each with its own processing capacity and memory make up an HPC cluster. Large-scale text processing jobs requiring high-performance computing resources can be handled by HPC clusters [8].

B. Remote services

Semantic text analysis cloud services offer a strong and adaptable foundation for handling and evaluating massive amounts of text data. These services include a variety of techniques and features, including as sentiment analysis, entity recognition, and NLP, for deriving insights and meaning from text. One example of a cloud service for semantic text analysis is Google Cloud Natural Language. This service recognizes the language of a given document and extracts important phrases using machine learning. Along with emphasizing entity extraction, sentiment analysis, syntax analysis, and categorization, Google's in-depth learning modules fuel the API.

Another cloud-based tool for semantic text analysis is Amazon Comprehend. With the help of machine learning, this service extracts insightful information from language in emails, social media feeds, support requests, documents, and more. By extracting text, important phrases, subjects, sentiment, and more from documents like insurance claims, it also streamlines document processing operations.

VIII. Quantum technologies in semantics

Unlike classical NLP, Quantum NLP represents linguistic aspects utilizing the quantum theory mathematical framework. By using this method, lexical meanings can be encoded as quantum states that can be processed by quantum circuits in simulators or specialized hardware.

The theoretical basis of QNLP is the DisCoCat (categorical distributional compositional) model, which uses string diagrams to convert grammatical structures into quantum processes.

Heavy preprocessing and syntax-dependent architectures are two drawbacks of syntactic analysis in traditional NLP that are addressed by the creation of quantum self-attention neural networks (QSANN). By including a self-attention mechanism, QSANN enhances the scalability and efficacy of quantum neural networks for bigger datasets, and enables their implementation on near-term quantum devices [9].

Sentiment analysis has benefited greatly from the application of QNLP, which has produced flawless test set accuracy in a number of simulations. This demonstrates how quantum computing may lead to major improvements in the comprehension and analysis of textual human emotions.

However, neither traditional nor quantum architecture currently provides a full-fledged semantic platform for solving various tasks in the field of text processing.

The considered promising developments are still partial solutions and require significant modernization.

IX. Conclusion and prospects

Summarizing all the described above, we can formulate a generalized scheme for the proposed approach of intelligent text classifier, including the role of authorship identification when attributing anonymous texts to one or another author (class).

In the scheme, after aggregating the raw data, the texts are sent to three groups for further processing.

In the first one, text processing is performed on familiar hardware using widely used proven classification algorithms such as decision trees, nearest neighbor method and so on.

In the second one text features are encoded in quantum space, then they are transferred to hardware supporting this form of representation (quantum processors or classical devices simulators).

In the third one text is presented in the some semantic model form, and processed by specialized hardware designed for semantic analysis [10]. The semantic model can be described in the form of some abstract agent-oriented model of information processing, for example, using the CS-code [11].

The final decision about text belonging to a class (author) is made taking into account each of the above results.

The approach is showed on Figure 2.

Involvement of semantic and quantum spaces in text classification for authorship identification requires further discussion and testing, as it increases the cost of resources used.

Nevertheless, using the described components in addition to existing established methods, it is possible to increase the accuracy of the final result by analyzing not only lexical and syntactic elements, but also other components of the text. These components that may include hidden borrowings.

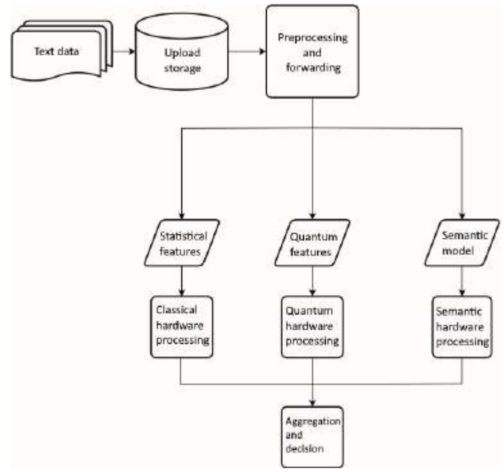


Figure 2. Potential analysing architecture.

References

[1] A. Paramonov, I. Trukhanovich, U. Kuntsevich, "Dynamic features selection in authorship identification problem,“ Otkrytye semanticheskie tekhnologii proektirovaniya intellektual'nykh system [Open semantic technologies for intelligent systems], pp. 309–312, 2021.

[2] M. Gavrilova. Analyzing approaches to plagiarism detection. Molodezhniy nauchno-tekhnicheskiy vestnik [Electronic periodical youth scientific and technical bulletin]. 2015, no. 1, pp. 27–35. (In Russian).

[3] E. Moskalenko, Y. Slesarev, Methods of checking the uniqueness of the text documents. Available at: https://web.snauka.ru/issues/2016/06/69137 (accessed 2023, December).

[4] V. Ivashenko, "Semantic Space Integration of Logical Knowledge Representation and Knowledge Processing Models,“ Otkrytye semanticheskie tekhnologii proektirovaniya intellektual'nykh system [Open semantic technologies for intelligent systems], pp. 95–114, 2023.

[5] D. Shunkevich, "Ontology-based Design of Knowledge Processing Machines,“ Otkrytye semanticheskie tekhnologii proektirovaniya intellektual'nykh system [Open semantic technologies for intelligent systems], pp. 73–94, 2017.

[6] V. Golenkov, N. Gulyakina, "The Main Directions, Problems and Prospects of the Development of the Next-Generation Intelligent Computer Systems and the Corresponding Technology,“ Otkrytye semanticheskie tekhnologii proektirovaniya intellektual'nykh system [Open semantic technologies for intelligent systems], pp. 15–26, 2023.

[7] Tensor Processing Unit : Architecture, Working & Its Applications. Available at: https://www.elprocus.com/tensor-processing-unit/ (accessed 2024, February).

[8] High-performance computing solutions. Available at: https://www.ibm.com/high-performance-computing (accessed 2024, February).

[9] G. Li, X. Zhao and X. Wang, "Quantum Self-Attention Neural Networks for Text Classification," 2022. [Online]. Available: https://arxiv.org/abs/2205.05625

[10] M. Tatur, A. Paramonov, "Open semantic technology as the foundation for new generation intelligent systems,“ Otkrytye semanticheskie tekhnologii proektirovaniya intellektual'nykh system [Open semantic technologies for intelligent systems], pp. 61–66, 2023.

[11] Vladimir Golenkov, *Technology of Complex Life Cycle Support of Semantically Compatible Next-Generation Intelligent Computer Systems*, V. Golenkov, Ed. Minsk: Bestprint [Bestprint], 2023.

ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ В ИДЕНТИФИКАЦИИ АВТОРСТВА ТЕКСТА

Труханович И. А., Парамонов А. И.

Рассматривается проблема идентификации авторов текстов. Делается обобщенный обзор текущего состояния проблемы. Предлагается решение на основе модификации ансамблевых методов машинного обучения. Исследуются возможности применения сложных вычислительных методов, таких как обработка естественного языка, алгоритмы машинного обучения и стилометрический анализ, для идентификации отдельных писателей на основе их отличительных лингвистических моделей. Выдвигается гипотеза о необходимости многомерного анализа текста и предлагается новый подход для идентификации авторства текста в виде гибридной интеллектуальной системы.

Received 29.03.2024

Belarusian Language Oriented Intelligent Voice Assistants

Yauheniya Zianouka, Volga Dydo,
Maxim Lutich, Vitalij Chachlou, Juras Hetsevich
United Institute of Informatics Problems
National Academy of Sciences of Belarus Minsk, Belarus
{evgeniakacan, olgadydo1,
maksim.lyutich, vitalikhokhlov, yuras.hetsevich}@gmail.com

Vadim Zahariev, Veronika Krischenovich
Belarusian State University of
Informatics and Radioelectronics
Minsk, Belarus
{zahariev, krish}@bsuir.by

***Abstract*—The article represents Belarusian intelligent voice assistants platform in open access and free use. It depicts an architecture of question-answering systems, their versions and the principles of work. Also, Belarusian modern speech synthesis and recognition systems of new generation are described in detail, which are the core of AI-assistants. The platform employs a structured approach, including an input interface, data processing, and information search, to ensure relevant and accurate answers for users. The OSTIS technology is shown as a means of improving AI-assistants' responses.**

***Keywords*—artificial intelligence, question-answering system, natural language processing, voice assistant, text-to-speech system, speech recognition system, large language model, chatbot, OSTIS technology**

I. Introduction

Artificial intelligence (AI) is playing an increasingly significant role in the development of modern society. With the advent of new technologies and opportunities, the use of AI is becoming an integral part of our lives, influencing all areas of activity, from business and science to education and medicine. For example, question-answering systems use Natural Language Processing (NLP) techniques, speech synthesis and recognition systems, machine learning, dialogue systems and other algorithms to understand questions, search for relevant information and generate answers based on available data [1]. They solve the difficult task of understanding natural language, which is one of the key components of artificial intelligence.

Such question-answering systems as *Chat GPT-4, Midjorney, Google Gemini, GPT Yandex, Burd, Copilot, Mistral* are already well known. They process requests in different languages [2], with the exception of the Belarusian language. For Belarusian speakers, the Speech Synthesis and Recognition Laboratory of the UIIP of the National Academy of Sciences of Belarus has developed an interactive Voice AI-assistant platform [3] which contains a set of Belarusian-speaking female and male question-answering assistants. The concept of the platform is based on the provision of an effective and easy-to-use mechanism for performing general information and solving user problems in the Belarusian language [4]. Assistants are represented in three versions (Web-version, iOS and Android platforms for mobile applications, and chatbots on the Telegram social network). Each system is built using speech recognition and synthesis technologies, machine translation, and dialogue systems. They allow users to ask questions verbally or in text form and receive an audio/printed response quickly, with high quality and accuracy.

II. The architecture of intelligent voice assistants

The AI-assistants of the intelligent question-answering platform have the following structure (Fig. 1) [5]: 1. Input interface. The user can ask an assistant a question verbally or in text format. Voice requests are more difficult to process than text messages. Therefore, at this stage, *the Belarusian-language speech recognition system (BSRM)* [6] is used to convert the audio signal into text form for searching the most relevant response. The system processes voiced requests efficiently, which reduces the likelihood of an answer that does not match the request.

2. Processing of input data. The system performs a re-interpretation of the question to understand its meaning and context. This step includes a set of natural language processing tools that highlight the main semantic units in the question. If a chart with this user has been created before, a dialogue summary is loaded from the database for more accurate and related responses.

3. Information search, ranking of answers, and their extraction. The voice assistant is looking for suitable information that may contain the answer to the question in online resources. The GPT language model (namely the ChatGPT-3,5 model) is used in Belarusian-speaking question-answering systems. The model is pre-trained on huge sets of text data; therefore, GPT can generate text that makes sense, uses the correct grammar and sentence structure. If there are several possible answers, the system can use a ranking algorithm to select the most relevant one. ChatGPT-3.5 supports the Belarusian

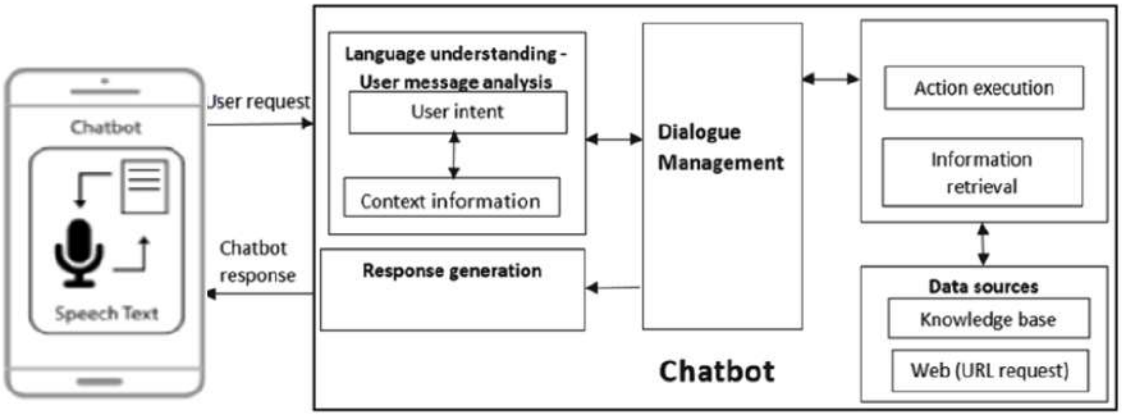


Figure 1. An architecture of intelligent voice assistant

language, but the quality of responses is not very good. To improve it, an additional machine translation unit has been developed, in which all queries are automatically translated from Belarusian to English using the Google Translate system. Then the most accurate answer given by ChatGPT goes back to the machine translation block, where it is already converted in Belarusian. Also, at this stage, the system saves answers to the language model in the database to create a summary of the dialogue with the user.

4. Forming a response. The voice assistant generates a response to the user in the form of a text or voice message, depending on the user's desire. An updated model of the Belarusian-language online speech synthesizer is used to output a voice message.

III. Belarusian speech recognition system using deep machine learning

To process the spoken question, a high-quality Belarusian speech recognition system (BSRS) [6] is used (fig. 2). It is based on an end-to-end architecture using deep learning and hosted on the Hugging Face platform, which allows users to create and share machine learning models and datasets.

To develop the BSRS, a large corpus of well-read texts in the Belarusian language was collected. The total duration of the audio recordings is 987 hours,voiced by 6160 speakers. The high variability of the collected data both from the point of view to speakers (gender, age, speech tempo, other features) and to recording conditions (various microphones, background noise, etc.) shows its quality. This is the first example of such big datasets for the Belarusian language.

To build a speech recognition model, a deep neural network architecture **wav2vec2** was chosen [7]. Its advantage is pretraining on a corpus of non-annotated data to study the ways of qualitative selecting features from the input recording. The obtained features are used for further subtasks, for example, to teach the model to convert speech into text. We used *"facebook/wav2vec2-base"* as the pre-studied model. Its further training was conducted on the Belarusian speech data set collected on *the Common Voice platform*. The training, validation and test samples were left unchanged; that is, the limit on the number of vocalizations of the same sentence was not removed, and the data set size was 345 909 audio recordings.

The speech recognition systems consist of 2 main components:

1) *The acoustic model* is a speech recognition system unit that builds a sequence of phonemes (or letters) that are pronounced with the greatest probability. It is based on the features selected from the input audio signal.
2) *A language model*, which is needed to translate a set of phonemes or letters obtained from an input audio recording into a set of the most likely words – the final transcription.

For the acoustic model, all audio recordings were converted to the following format: sampling rate 16 kHz, 1 channel (mono). We reduced text transcriptions to lowercase; removed all characters except letters of the alphabet and numbers; and replaced each sequence of characters-spaces (space, tab, etc.) by 1 space. We used *CTC* as a loss function to train the model for speech recognition tasks. The parameters were optimized using *the AdamW algorithm*, a corrected version of the popular Adam optimization algorithm. *the Gradient Checkpointing* method was also applied to optimize memory consumption. The selection of the best model was carried out using the*WER (Word Error Rate Metric)* on a validation sample [8].

The software implementation of acoustic model training was fulfilled on a popular framework for training NLP and ASR models *HuggingFace*. It is a shell over another framework — *PyTorch* [9]. The training was conducted on a server with 3 NVIDIA GeForce RTX

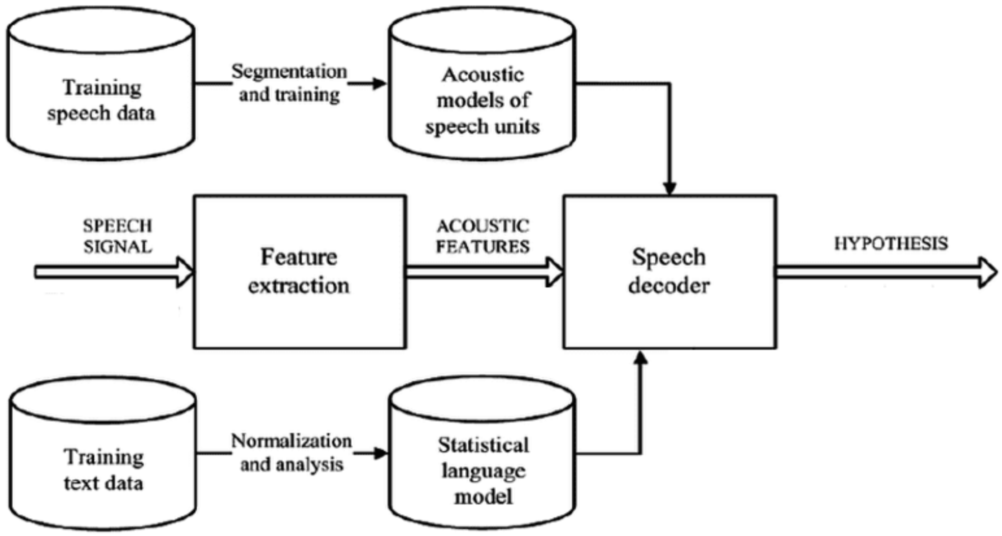


Figure 2. The structure of Belarusian Speech Recognition System

2080 Ti graphics cards. The size of the batch for training and evaluating the quality of the model was 48 (16 elements in the batch on each of the 3 video cards). The average time per epoch was about 8 hours. In this regard, and with acceptable metric values, training was discontinued after 5 epochs. Quality evaluation and compliance with intermediate parameters (checkpointing) were carried out several times during each epoch. To improve the predictions of the acoustic model, a 5-gram language model was constructed using modified *Kneser-Ney smoothing*. The popular *KenLM library*, created by the authors, was used for building such a model. The language model was trained on the text corpus described above. The total number of sentences needed to build a language model was 314 676. Adding a language model allowed for a reduction of WER from 0.187 to 0.124 in the test sample. The final result, WER 0.124 (or 12.4 per cent) is quite good for recognition models. For example, the current best value of test WER for the German Common Voice dataset is 5.7 per cent. Now the model is able to recognize arbitrary Belarusian speech at a fairly good level [10].

IV. A new generation Belarusian text-to-speech synthesizer

A new generation Belarusian-text-to-speech synthesizer is based on *The VITS (Variational Inference with adversarial learning for Text-to-Speech)* [11]. This is a one-stage non-autoregressive text-to-speech model capable of generating more natural sound than existing two-stage models such as *Tacotron 2, Transformer TTS, or even Glow-TTS*. Using a variational framework, VITS models a latent space of speech features, reflecting the inherent variability and uncertainty in speech generation. Competitive learning environment at VITS enhances the synthesis process. Collaborative learning involves training the discriminator network to distinguish between real and synthesized speech, while the generator network aims to produce speech that successfully fools the discriminator.

This adversarial interaction helps to improve the overall quality and naturalness of the synthesized speech samples. VITS serves as a stand-alone text-to-speech solution as it does not require a separate vocoder. The general architecture of VITS is depicted in fig. 3. It consists of *a Posterior encoder, a Prior encoder, a Decoder and a Stochastic Duration Predictor*. The Posterior Encoder and Decoder Discriminator modules are used only during training, not for speech output. For the Posterior Encoder, *16 residual WaveNet blocks* are used, consisting of layers of extended scrolls with an activation block and a communication pass. The back-end encoder takes *xlin* linearly scaled logarithmic spectrograms as input and produces latent variables Z with 192 channels. The idea behind the Posterior Encoder is to translate the audio data from the mel-spectrogram space to the normal distribution space. That is why the model uses a linear layer above the Posterior Encoder to obtain the average variance of the normal posterior distribution. Prior Encoder consists of *Text Encoder, Projection Layer, Normalizing Flow, and uses Monotonic Alignment Search (MAS)*. Like Posterior Encoder, Prior Encoder aims to map textual data from phoneme space to normally distributed space [12].

Collecting data for model training is an essential step for the development of a Belarusian text-to-speech system. Given the limited availability of specialized datasets for the Belarusian language, the *CommonVoice dataset* was chosen as the main data source. It is a large collection of voice recordings collected from voluntary participants who read sentences in different languages.

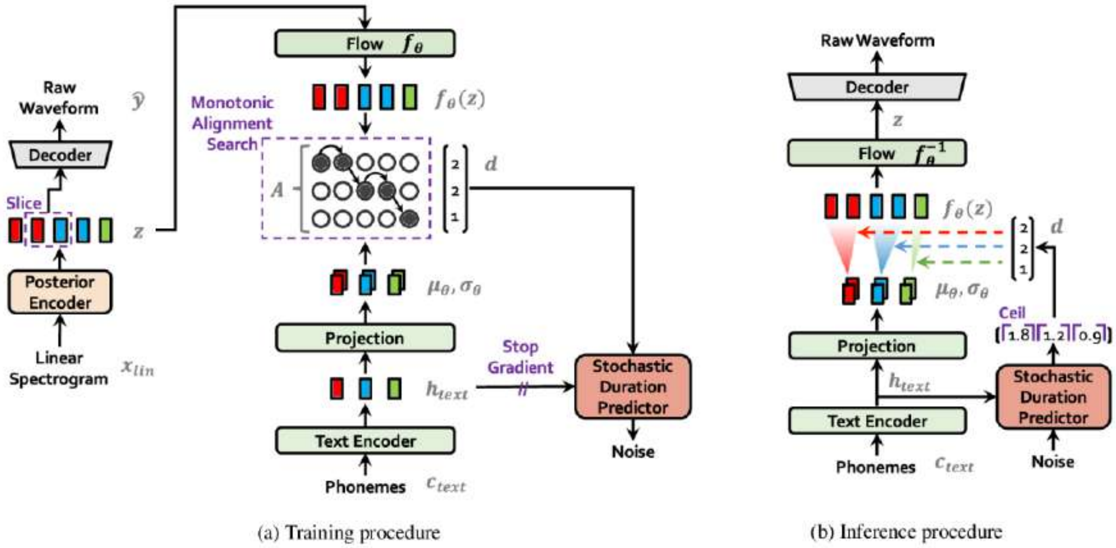


Figure 3. The structure of VITS

This dataset is available for free use and distributed under an open license, which makes it a valuable source for the development of linguacoustic resources in the Belarusian language. The disadvantage of the dataset from CommonVoice is that it is intended for speech recognition tasks and not specifically for synthesis. In this regard, the recordings from the dataset can be unprofessional and contain various noises or flaws that negatively affect the quality of speech synthesis.

To fix this, pre-filtering and selection of recordings were performed to ensure that the selected data were suitable for the speech synthesis task and of good quality. An audio material (about 20 000 words) met the selection criteria of the largest amount of available audio and a relatively low level of noise and artefacts. The analysis of the data was carried out with the aim of obtaining statistical information and understanding the peculiarities of the Belarusian language in the context of speech synthesis. It included an assessment of the distribution of phonetic units, the length of phrases, and other characteristics that may affect the quality and naturalness of synthesized speech.

The VITS model was trained using the *Coqui TTS library*, a popular open-source toolkit for TTS. Coqui TTS provides a complete set of tools and utilities for training and deploying TTS models [13]. The VITS model used the *Weights and Biases recorder* during the training process. It is a platform for tracking and visualizing machine learning experiments. It allows researchers and developers to log in and track training progress, metrics, and model performance in real-time.

By leveraging the capabilities of the Coqui TTS library and the integration of the Weights and Biases register, the VITS model was trained using robust TTS development toolkit. The use of Coqui TTS and the Wandb logger facilitated efficient experimentation, model optimization, and performance monitoring throughout the learning process. A server with an *Nvidia RTX4090 graphics card* was used for this. Parameters were optimized using the textitAdamW algorithm, a corrected version of the popular Adam optimization algorithm. The batch size for model training and evaluation was 74, and the model training time was 72 hours.

With the help of neural networks, the training and learning of the acoustic database were carried out. Therefore, it is created automatically and has quite accurate results. However, this synthesizer is large in size, which can reduce the speed of text output. The lack of processing numbers, figures, dates, and abbreviations are also considered a big drawback, which is the object of development for new methods and algorithms to correct this bug.

V. Versions of intelligent voice assistants

Voice assistants are available on the official website of the platform, the interface of which is presented in Belarusian, English, Russian and Chinese [14]. The user can choose a personal virtual interlocutor *(AIAlesBot, AIAlesiaBot, AIAlenaBot, AIBorisBot, AIKirylBot, AsistentBot)* and chat on the Internet by selecting the Web version or a smartphone by installing the application on Android or iOS operating systems. Question-answering systems are also available in the form of chatbots in Telegram Messenger. Launching an official website, the user chooses a convenient version, and then he is redirected to the version he has chosen, and enters a request. In addition, everyone can chat with chatbots (AIVasil, AIVasilina), which are narrowly focused on various fields of activity (general assistant, architect, business analyst, financial consultant, recruiter, project manager, legal adviser, marketer, engineer, programmer, teacher, and writer). The main feature of all voice assistants is that they process text or audio responses only in Belarusian, whereas text requests can be sent in English. Voice queries are also recognized only in Belarusian.

Currently, mobile applications for iOS and Android versions are being developed. The process of their creation consists of several stages: approving the design concept of the application and its functional features; work on the back-end; direct implementation of the project. The iOS mobile application is realized in *Swift* using the *UIKit framework*. This is a high-quality combination of technical solutions and a design concept. Technical tools include *the Massage Kit library* for creating a functional chat, as well as the integration of methods with the native *AVFoundation library* for flawless work with audio files. *The UICollectionView class* was selected to build the interface.

The writing language of the Android version is *Kotlin*. The following technologies were also used: Android Architecture, MVVM Architecture (using ViewModels), WorkManager, Kotlin coroutines, Java Threads, OkHttp, SQLite database (Room technology), Canvas.

The applications work as follows. The logo or the start video is loaded on the first page, and then the user can go to the settings or contacts screen. To select a certain assistant, the user must click on the assistant card, and then the chat window opens. Then a chatbot greets the user, after which the second one enters a text or voice request. After entering a request, he cannot set another request until he receives a response from the bot. The system automatically voices the message in the manner of the selected speaker.

The verification of question-answering systems takes place daily. According to statistics, chatbots are good at answering simple questions like *“How many colours are in a rainbow?", and "How much is 1+1?")*, as well as difficult ones (for example, *“Tell me about the very first film”, “I need an interesting story about Shakespeare"*). Text queries and responses in English and Belarusian are quite high-quality. If the question was asked in Russian or some other languages, the bots will answer in English. On average, the user can receive a response from the chatbot within 10-30 seconds, which is a good result of the server side.

VI. Application prospects of OSTIS technology for voice assistants

Semantic technologies, particularly the OSTIS technology, offer numerous advantages when applied in the development of voice assistants tailored for specific languages, like Belarusian. Firstly, they provide a flexible framework for constructing dialogue rules, enabling efficient handling of user queries in natural language. This flexibility ensures adaptability to diverse conversational contexts, resulting in more intuitive and user-friendly interactions [16].

Secondly, semantic technologies, including OSTIS, offer mechanisms to limit the size and context of responses, addressing the issue of information overload and reducing irrelevant or excessive outputs. This targeted response approach enhances the user experience by delivering concise and relevant information.

Moreover, the application of such technologies contributes to reducing hallucinations, a common challenge faced in language generation models like GPT and other LLM’s. By leveraging semantic understanding and contextual awareness, OSTIS-based voice assistants can generate responses that are more coherent and accurate, minimizing nonsensical or misleading outputs.

The integration of OSTIS technology also facilitates the development of intelligent voice assistants capable of capturing and analyzing user profiles and dialogue histories. This feature enables personalized interactions, where the assistant can tailor responses based on individual preferences and past interactions, thereby enhancing user satisfaction and engagement.

Furthermore, semantic technologies allow for seamless integration of domain-specific knowledge with general knowledge bases within voice assistant systems. This integration involves efficient handling of domain-specific queries and contextually relevant responses, ensuring the assistant’s effectiveness across diverse topics and applications.

VII. Conclusion

The article depicts Belarusian question-answering systems, which are available on the AI-assistant platform. The goal of these assistants is to provide easy access to information in the Belarusian language. With the use of artificial intelligence, question-answering systems provide quick and accurate responses on various topics, including scientific discussions and entertainment suggestions. To do this, assistants invoke such technologies as text-to-speech and recognition systems, effective algorithms of search, and machine translation.

The relevance of the assistants is due to the lack of competitive chat-bots that support the Belarusian language, whereas there are a huge number of question-answering systems for other languages. Such developments make them more accessible to local users, offering a convenient option for searching for information and communication in a modern global Internet network and using computer technologies in their native language. Future plans for the platform include new functions such as creating custom bots, gathering user feedback, saving message history, supporting developers, and extracting information from the internet to enhance effectiveness.

Applying semantic technologies, specifically OSTIS, in the development of voice assistants brings numerous benefits, including enhanced dialogue management, reduced hallucinations, and personalized interactions. This integration not only improves response relevance and coherence, but also underscores the significance of semantic technologies for user satisfaction and engagement.

References

[1] Bansal,H.,Khan,R. A review paper on human computer interaction.Int.J.Adv.Res.Comput. Sci. Softw. Eng. 8, 53 (2018). Avaliable at: https://doi.org/10.23956/ijarcsse.v8i4.630 (accessed 11.05.2023).

[2] Adamopoulou, Eleni Moussiades, Lefteris An Overview of Chatbot Technology. In IFIP International Conference on Artificial Intelligence Applications and Innovations. Avaliable at: https://www.researchgate.net/publication/341730184 (accessed 17.02.2024).

[3] AI Voice Assistant. Avaliable at: https://asistent.io/ (accessed 09.01.2024).

[4] Hiecevic,Ju, Zianouka, Ja., Dydo, V. Liucic, M., Pavucina, M. Bielaruskamounaja halasavaja pytalna-adkaznaja sistema [The Belarusian-language voice question-answering system]. Science and innovation. Minsk. 2023. №7 (245). P. 13-16. (In Belarusian)

[5] Hiecevič, J.S., Zianoŭka J.S., Trafimaŭ A.S., Bakunovič A.A., Latyševič D.I., Drahun A.JA., Sliesarava M.M., Tukaj M.S., Komplieks srodkaŭ realizacyi zadač štučnaha inteliektu dlia bielaruskaj movy [A complex of means to implement tasks of artificial intelligence for the Belarusian language]. Piervaja vystavka-forum IT-akadiemhrada Iskusstviennyj intielliekt v Bielarusi. UIIP NASB. Minsk. 2022. P. 64-73. (In Belarusian)

[6] Belarusian Speech Recognition. Avaliable at: https://corpus.by/BelarusianSpeechRecognition/?lang=en (accessed 31.09.2023).

[7] Dhankar, Abhishek Study of deep learning and CMU sphinx in automatic speech recognition. 2017 International Conference on Advances in Computing, Communications and Informatics,2017, pp. 2296-2301.

[8] Graves Alex, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. Proceedings of the 23rd international conference on Machine learning, 2006, pp. 369-376.

[9] Heafield Kenneth, Ivan Pouzyrevsky, Jonathan H. Clark, and Philipp Koehn Scalable modified Kneser-Ney language model estimation. Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, Volume 2: Short Papers, 2013, pp. 690-696.

[10] Trafimau, A. S. Autamatycnaje pierautvarennie bielaruskaha mauliennia u tekst [Automatic conversion of Belarusian speech into text]. XXI International Scientific and Technical Conference "Development of Informatization and State System of Scientific and Technical Information RINTI-2022", UIIP NASB, Minsk. 2022.P. 241-245. (In Belarusian)

[11] Jaehyeon Kim, Jungil Kong, Juhee Son Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech. Avaliable at: https://arxiv.org/abs/2106.06103 (accessed 03.02.2023).

[12] Jungil Kong, Jaehyeon Kim, Jaekyoung Bae HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis. Avaliable at: https://arxiv.org/abs/2010.05646 (accessed 13.02.2023).

[13] Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, Tie-Yan Liu FastSpeech: Fast, Robust and Controllable Text to Speech. Avaliable at: https://arxiv.org/abs/1905.09263 (accessed 17.03.2023).

[14] Dydo, V. V., Liucic, M. S., Pavucina, M. A., Drahun, A. Ja., Chachlou, V. A., Trafimau, A. S., Zianouka, Ja. S., Hiecevic Ju. S. BIELARUSKAMOUNY HALASAVY AI-ASISTENT [BELARUSIAN-SPEAKING VOICE AI ASSISTANT]. XXII International Scientific and Technical Conference "Development of Informatization and State System of Scientific and Technical Information RINTI-2022", UIIP NASB, Minsk. 2023. P. 190-194. (In Belarusian)

[15] Golenkov V., Gulyakina N. A., Shunkevich D. V. Open technology of ontological design, production and operation of semantically compatible hybrid intelligent computer systems. Bestprint [Bestprint]. Minsk. 2021.

[16] Zahariev, V., Shunkevich, D., Nikiforov, S., Azarov, E. Intelligent Voice Assistant Based on Open Semantic Technology. Open Semantic Technologies for Intelligent System. OSTIS 2020. Communications in Computer and Information Science, vol 1282. Springer, Cham. 2020, pp. 121-145.

ИНТЕЛЛЕКТУАЛЬНЫЕ ГОЛОСОВЫЕ АССИСТЕНТЫ, ОРИЕНТИРОВАННЫЕ НА БЕЛОРУССКИЙ ЯЗЫК

Зеновко Е., Дыдо О., Шуст М., Хохлов В., Гецевич Ю., Захарьев В., Крищенович В.

В статье представлены интеллектуальные голосовые ассистенты на белорусском языке, размещенные на Интернет-платформе Voice AI-assistant. Главной целью разработки ассистентов является обеспечение эффективного и простого в использовании механизма предоставления общей информации и решения вопросов пользователей на белорусском языке. Вопросно-ответная платформа "Голосовой ИИ-ассистент"позволяет пользователю задать вопрос на белорусском языке текстовым или голосовым сообщением и получить на него звуковой или напечатанный ответ. За счет использования искусственного интеллекта она дает возможность получать быстрые, качественные и точные ответы по различным темам.

Ассистенты представлены в трех версиях (Web-версия, iOS- и Android-платформы для мобильных приложений и чат-боты в социальной сети Telegram). Каждая система построена с использованием технологий распознавания и синтеза речи, машинного перевода и диалоговых систем. Описанные в работе белорусскоязычные системы синтеза и распознавания речи свидетельствуют о высоком уровне развития речевых технологий на белорусском языке.

Актуальность данных ассистентов обусловлена отсутствием конкурентоспособных чат-ботов, поддерживающих белорусский язык, тогда как для других языков существует огромное количество голосовых ассистентов. Разработка устройств на белорусском языке делает их более доступными для белорусскоязычных пользователей, предлагая удобный вариант поиска информации и коммуникации в современной глобальной интернет-сети и использования компьютерных технологий на их родном языке.

Received 14.03.2024

An Approach to Automate the Entire Process from the Generation of Test Questions to the Verification of User Answers in Intelligent Tutoring Systems

Wenzu Li
Belarussian State University Informatics and Radioelectronics
Minsk, Belarus
lwzzggml@gmail.com

***Abstract*—This article is dedicated to the issues of automatic generation of test questions and automatic verification of user answers in the new generation of intelligent systems based on semantic representation of information. An approach to implement automatic generation of test questions of various types using knowledge bases and automatic verification of user answers according to the semantic representation structure of the knowledge is detailed in this article.**

***Keywords*—test question generation, user answer verification, OSTIS Technology, intelligent systems based on semantic representation of information, semantic structure, knowledge base**

I. Introduction

With the rapid advancement of artificial intelligence technology in areas like natural language processing and image processing, researchers have started integrating artificial intelligence into the field of education. One of the most representative products of the combination of artificial intelligence and education is **Intelligent Tutoring Systems** (ITS). The significance of ITS lies in bringing revolutionary changes to the field of education. By combining artificial intelligence technology with educational theory, ITS can personalize guidance for students, providing tailored educational content and feedback based on each student's learning style, progress, and needs. This personalized teaching approach helps stimulate students' interest in learning, improve learning efficiency, and facilitate knowledge absorption and mastery. In conclusion, the emergence of ITS injects new vitality into education, offering students and teachers more effective learning and teaching tools, thus driving the progress and development of education [1].

In ITS, the automatic generation of test questions and the automatic verification of user answers play a crucial role. By generating test questions automatically, the system can provide personalized tests based on students' learning content and level, aiding them in comprehensive review and reinforcement of knowledge points. On the other hand, the automatic verification of user answers allows the system to promptly check students' answers and provide immediate feedback or guidance on correctness. This helps students correct errors in a timely manner during the learning process, enhancing their understanding and retention of knowledge points. In conclusion, the automatic generation of test questions and the automatic verification of user answers enhance personalization and immediacy in ITS, leading to improved learning outcomes and an increase in teaching quality [2], [3], [4], [5].

In recent years, with the continuous development of artificial intelligence technology, a variety of approaches to automatically generate test questions and automatically verify users' answers have been proposed. However, these approaches still have some limitations, mainly in the following areas:

- current approaches to test question generation only allow for the generation of test questions of simple types;
- some of the existing approaches to verifying user answers (e.g., keyword matching and probabilistic statistics) do not take into account the semantic similarity between answers;
- semantic-based approaches to verifying user answers only allow verification of answers with simple semantic structures [5], [6], [7].

Therefore an approach to develop a subsystem for automatic generation of test questions of various types using knowledge bases and automatic verification of user's answers according to the semantic representation structure of information in intelligent tutoring systems of the new generation based on semantic representation of information is proposed in this article [2], [3], [8]. Using the developed subsystem not only allows to generate test questions of various types using the knowledge bases and to verify the correctness and completeness of the user's answers based on semantics, but also to automate the entire process from the generation of test questions to the grading of the test papers. The ITS for discrete mathematics will be used as a demonstration system for the developed subsystem.

The remainder of this article is organized as follows. Section II presents a review of several existing approaches to generating test questions and verifying user answers. Section III describes our proposed approach to automatically generate test questions and automatically verify user answers. Section IV evaluates the effectiveness of the subsystem developed using our proposed approach. Section V concludes this article.

II. Related works

A. Automatic generation of test questions

Approach to automatic generation of test questions mainly focuses on automatic generation of test questions using electronic documents, text corpora and knowledge bases. Knowledge bases store highly structured knowledge that has been filtered, and with the development of the semantic networks, automatic generation of test questions from knowledge bases has become the most important research direction in this field. The basic principles of automatic generation of objective questions using knowledge bases are described in detail in references [7], [9].

Objective questions usually have a unique standard answer. In this article, objective questions include: multiple-choice questions, fill in the blank questions and judgment questions. Objective questions differ from subjective questions, which have more than one potential correct answer. Subjective questions in this article include: definition explanation questions, proof questions and problem solving task.

The primary issues with current approaches to generating test questions are as follows:

- using electronic documents to automatically generate test questions necessitates a substantial amount of sentence templates;
- compiling a text corpus demands significant human resources to gather and process diverse knowledge;
- existing approaches only allow to generate simple objective questions [10], [11].

B. Automatic verification of user answers

Automatic verification of user answers is divided into verification of answers to objective questions and verification of answers to subjective questions. The basic principle of verification of answers to objective questions is to determine whether the string of standard answers matches the string of user answers. The basic principle of verification of answers to subjective questions is to calculate the similarity between standard answers and user answers, and then to implement automatic verification of user answers based on the calculated similarity and the evaluation strategy of the corresponding test questions [12], [13]. The more similar the standard answer and the user answer are, the higher the similarity between them. The verification of user answers to subjective questions is divided into the following categories according to the approaches used to calculate the similarity between the answers:

- based on natural language;
- based on semantic graph.

The basic principle of approaches to calculate the similarity between answers to subjective questions based on natural language is to convert natural language text into vectors using a series of tools for modelling natural language text (for example, Jaccard similarity, vector space models (TF-IDF, Doc2Vec), deep learning models (Transformer, BERT), and etc.), and then to calculate the similarity between the vectors. Since test questions of various types and their answers are described in the form of semantic graphs in the knowledge base of ITS, this article focuses on approaches to compute the similarity between answers using semantic graphs [14]. A semantic graph is a network that represents semantic relationships between concepts. In the reference [4] we have described in detail the approaches to calculate the similarity between answers to subjective questions based on natural language and compared their advantages and disadvantages.

The basic principle of calculating the similarity between answers to subjective questions (i.e., sentence or short text) based on semantic graphs is that the answers are first converted into semantic graph representations using natural language processing tools (such as syntactic dependency trees and natural language interfaces) and then the similarity between them is calculated (i.e., similarity between answers). The main advantage of this type of approach is computing the similarity between answers based on semantics. One of the most representative approaches is SPICE (Semantic Propositional Image Caption Evaluation) [15].

These approaches primarily encounter the following issues:

- the keyword phrase-based approach does not take into account the order between words in a sentence;
- the VSM-based approach leads to the generation of high-dimensional sparse matrices, which increases the complexity of the algorithm;
- semantic graph-based approaches supporting only the description of simple semantic structures;
- these approaches cannot determine whether the sentences are logically equivalent to each other;
- these approaches are dependent on the corresponding natural language.

In ITS information is described in the form of semantic graphs and stored in the knowledge base. Therefore for the above reasons an approach to automatically generate test questions of various types using knowledge bases and to automatically verify user answers according to the similarity between the semantic graphs of the answers in intelligent tutoring systems based on semantic representations of information is proposed in this article.

III. Proposed approach

The main task of this article is to develop a subsystem for automatic generation of test questions and automatic verification of user answers in intelligent tutoring systems of the new generation based on semantic representation of information. To achieve this task OSTIS Technology is proposed to be used [2], [3], [8].

A. Description of the used technology

OSTIS technology is a complex open semantic technology for the design and development of intelligent systems. Intelligent systems developed using the OSTIS Technology are called ostis-systems. Each ostis-system for different application fields includes a platform for interpretation semantic models of ostis-systems, as well as a semantic model of ostis-systems using SC-code (sc-model of ostis-systems). At the same time, the sc-model of the ostis-systems includes the sc-model of the knowledge base, the sc-model of the problem solver and the sc-model of the interface (in particular, the user-oriented intelligent interface). As a basis for knowledge representation within the OSTIS Technology, a unified coding language for information of any kind based on semantic networks is used, named SC-code. Texts of the SC-code (sc-texts) are unified semantic networks with a basic set-theoretic interpretation. Within the framework of the technology, several universal variants of visualization of the SC-code constructs are also proposed, such as SCg-code (graphic version), SCn-code (non-linear hypertextual version), SCs-code (linear string version). The methods and rules for the detailed design of intelligent systems using OSTIS technology are described in the reference [2].

OSTIS Technology offers the following possibilities:

- unified knowledge description language SC-code;
- models describing knowledge of various types, models and tools for the development of knowledge bases;
- integration of various problem-solving models;
- multi-agent approach to developing problem solvers.

Next we will describe in detail the process of developing a subsystem for the automatic generation of test questions and the automatic verification of user answers in intelligent tutoring systems developed using OSTIS technology.

B. Architecture of the subsystem

The task of this article is to develop a subsystem using the proposed approach for generating test questions and verifying user answers. Fig. 1 shows the organisation of the developed subsystem in the ostis-system.

As can be seen in Fig. 1 in the ostis-system information is passed between components in the form of semantic fragments constructed using SC-code. Natural language interfaces are used to implement the interaction between the intelligent system and the user. An approach to developing natural language interface using OSTIS Technology is described in the reference [16]. In order to facilitate the explanation of the working principle of the subsystem, the illustrations and knowledge base fragments we choose in this article are all in English, but it needs to be emphasized that the subsystem developed does not depend on natural language.

The subsystem is divided into two functional components according to the functions to be realised:

- component for automatic generation of test questions;
- component for automatic verification of user answers.

Fig. 2 shows the complete working process of the component for automatic generation of test questions.

The work corresponding to stages 1 and 5 in Fig. 2 is done using a natural language interface.

Fig. 3 shows the complete working process of the component for automatic verification of user answers.

The work corresponding to stages 1 and 5 in Fig. 3 is done using a natural language interface.

C. Development of formal semantic model of test questions

The formal semantic model of test questions is aimed at defining ways to describe test questions of various types in the form of semantic graphs in the knowledge bases of ostis-systems. The formal semantic model of test questions is defined as shown below:

$$M_{TQ} = \{M_{MCQ}, M_{FBQ}, M_{JQ}, M_{DIQ}, M_{PSQ}\} \quad (1)$$

The parameters are defined as shown below:

- M_{MCQ} — semantic model of multiple-choice question;
- M_{FBQ} — semantic model of fill in the blank question;
- M_{JQ} — semantic model of judgment question;
- M_{DIQ} — semantic model of definition explanation question;
- M_{PSQ} — semantic model of proof question and problem solving task.

The semantic model of multiple-choice question is defined as shown below:

$$M_{MCQ} = \{S_{MBI}, S_{MAI}, R_{MBI}, R_{MAI}\} \quad (2)$$

The parameters are defined as shown below:

- S_{MBI} — a set of concepts that specifies basic information about multiple-choice question, including

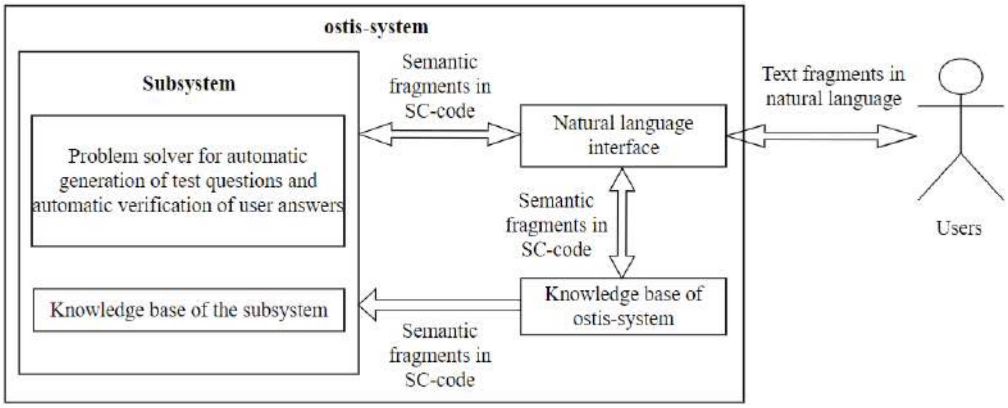


Figure 1. Organisation of the developed subsystem in the ostis-system.

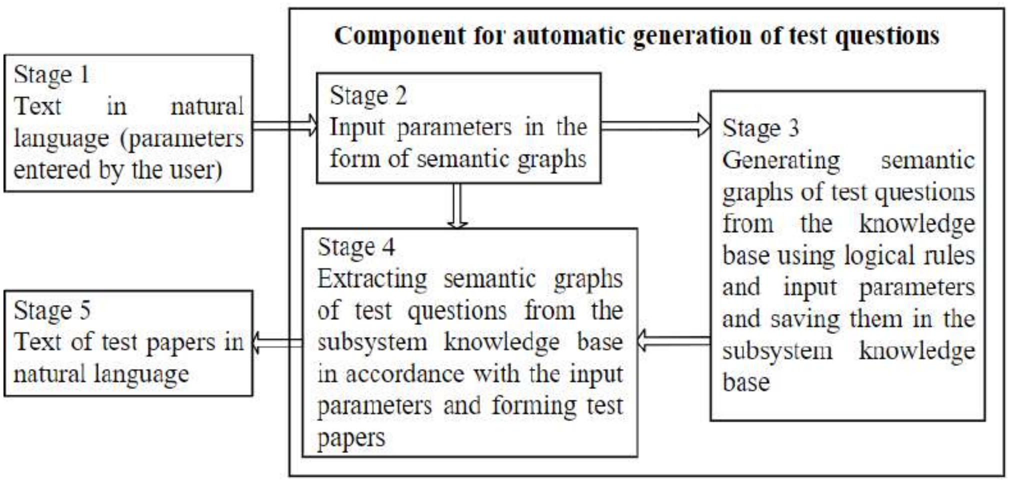


Figure 2. Working process of the component for automatic generation of test questions.

describing its type, key elements, options, and so on;

- S_{MAI} — a set of concepts that specifies the contextual information about the answer to multiple-choice question, including describing the standard answer and the user answer;
- R_{MBI} — a set of relations that specifies concepts from S_{MBI}, including "key sc-element' ", "standard answer*", etc.;
- R_{MAI} — a set of relations that specifies concepts from S_{MAI}.

Similarly semantic models of test questions of other types can be defined in the way described above.

D. Automatic generation of test questions

The basic principle of automatic generation of test questions of various types (including objective questions and subjective questions) in the ostis-systems is to first extract the corresponding semantic fragments from the knowledge base using a series of test question generation strategies summarized based on the knowledge representation approach and the knowledge description structure in the framework of OSTIS Technology, then add some test question description information to the extracted semantic fragments, and finally store the semantic fragments describing the complete test questions in the corresponding section of the subsystem [4]. The subsystem allows a series of test questions to be extracted from the subsystem and formed into test papers according to the user's requirements when test papers need to be generated. Test papers consisting of semantic graphs of test questions are converted to natural language descriptions using a nature language interface. In the following, the basic principles of automatic generation of objective questions in the ostis-systems will be introduced using test question generation strategy based on class as examples.

The "inclusion*" relation is one of the most frequently used relations in the knowledge base of the ostis-systems, which is satisfied between many classes (including subclasses), so that the inclusion relation between classes can be used to generate objective questions. The set

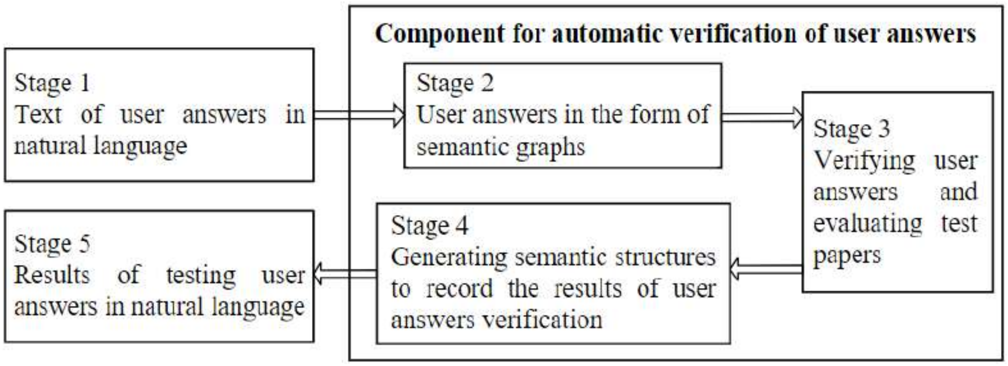


Figure 3. Working process of the component for automatic verification of user answers.

theory expression form of inclusion relation between classes is as follows: $S_i \subseteq C(i \geq 1)$, ($S$-subclass, i-subclass number, C-parent class) [5], [8]. The following shows a semantic fragment in the knowledge base that satisfies the inclusion relation in SCn-code:

binary tree
$\Leftarrow$ *inclusion**:
directed tree
$\Rightarrow$ *inclusion**:
- *binary sorting tree*
- *brother tree*
- *decision tree*

Consider the example of a multiple-choice question generated using this semantic fragment according to the strategy of inclusion relations, which has the natural language form shown below:

<<Binary tree does not include ( )?>>

A. directed tree B. brother tree

C. decision tree D. binary sorting tree

Fig. 4 shows the semantic graph of this generated multiple-choice question that was constructed based on the semantic model of the test questions.

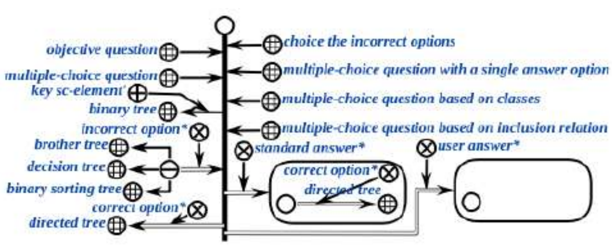


Figure 4. An example of semantic graph of multiple-choice question.

It should be emphasized that the semantic graph of this multiple-choice question was converted into the corresponding natural language description form using a natural language interface.

Fig. 5 shows an example of a logic rule for generating this multiple-choice question constructed based on a strategy of inclusion relation.

Logic rules of test questions in the ostis-systems are constructed using SC-code strictly according to the test question generation strategies. Each logic rule of test questions consists of two main parts: 1. search template; 2. generate template. The search template is used to find all the semantic fragments that satisfy the conditions in the knowledge base. The generation template uses the searched semantic fragments to generate the semantic graph of the test question.

Similarly objective questions of other types can be generated in a similar way using this strategy.

Other strategies used to generate objective questions include:

- Test question generation strategy based on elements;
- Test question generation strategy based on identifiers;
- Test question generation strategy based on axioms;
- Test question generation strategy based on relation attributes;
- Test question generation strategy based on image examples.

In reference [4] we describe in detail the approach to generating objective questions of various types.

The process of generating subjective questions is shown below:

- searching the knowledge base for semantic fragments describing subjective questions using logic rules;
- storing the found semantic fragments in the knowledge base of the subsystem;
- using manual approaches or automatic approaches (such as natural language interfaces) to describe the definition, proof process or solution process of the corresponding test question according to the knowledge representation rules in SCg-code or SCL-code (a special sub-language of the SC language intended for formalizing logical formulas), (this part

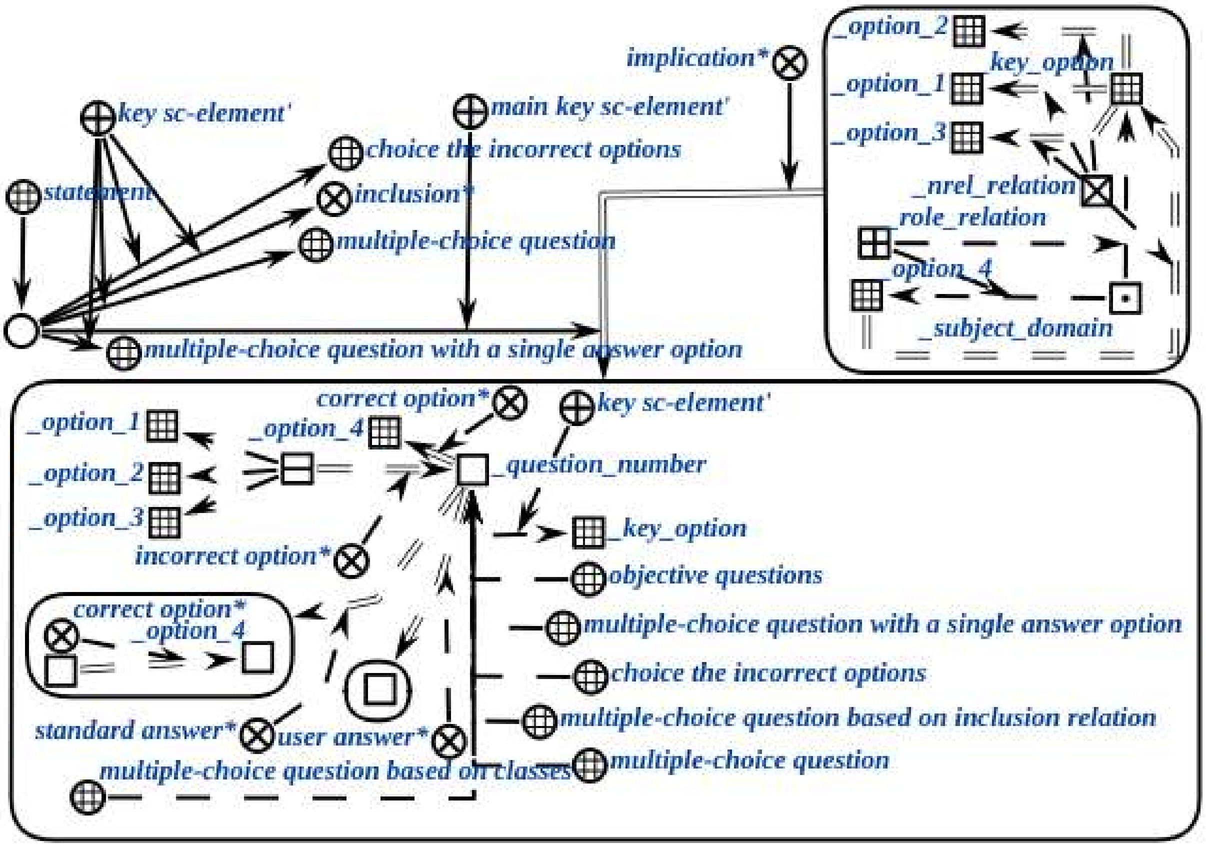


Figure 5. An example of a logic rule for generating multiple-choice question.

of the work can also be done before generating test questions) [8].

The suggested approach to generating test questions offers the following key benefits:

- within the framework of OSTIS Technology, knowledge is described in a unified semantic network language, so that the component developed using the proposed approach to generating test questions can be used in different ostis-systems;
- the developed semantic model of test questions does not rely on natural language, which greatly simplifies the use and processing of semantic graphs of test questions in the knowledge base;
- by utilizing the suggested approach, both objective and subjective questions of high quality can be automatically generated.

E. Automatic verification of user answers

Test questions in the ostis-systems are stored in the knowledge base as semantic graphs. Therefore, the most crucial step in verifying user answers is to calculate the similarity between the semantic graphs of the answers. Once the similarity is determined and combined with the evaluation strategy for the respective test questions, the correctness and completeness of user answers can be verified.

Since the knowledge types and knowledge structures used to describe different types of test questions are not the same, answer verification is further divided into: 1. verification of answers to objective questions; 2. verification of answers to subjective questions.

F. Verification of answers to objective question

Semantic graphs of answers to objective questions are described using factual knowledge according to the same knowledge structures. As a result, the similarity between the semantic graphs of answers to objective questions of different types can be calculated using the same approach.

Factual knowledge refers to knowledge that does not contain variable types, and this type of knowledge expresses facts. When the user answers to objective questions in natural language are converted into semantic graphs, they are already integrated with the knowledge already in the knowledge base. So the similarity between answers is calculated based on the semantic description structures [15]. The process of calculating the similarity between the semantic graphs of the answers to the objective questions is shown below:

- decomposing the semantic graphs of the answers into sub-structures according to the structure of the knowledge description;
- using formulas (3), (4), and (5) to calculate the precision, recall and similarity between semantic graphs.

$$P_{sc}(u, s) = \frac{|T_{sc}(u) \otimes T_{sc}(s)|}{|T_{sc}(u)|} \quad (3)$$

$$R_{sc}(u, s) = \frac{|T_{sc}(u) \otimes T_{sc}(s)|}{|T_{sc}(s)|} \quad (4)$$

$$F_{sc}(u, s) = \frac{2 \cdot P_{sc}(u, s) \cdot R_{sc}(u, s)}{P_{sc}(u, s) + R_{sc}(u, s)} \quad (5)$$

The parameters are defined as shown below:

- $T_{sc}(u)$ — all substructures after the decomposition of the user answers u;
- $T_{sc}(s)$ — all substructures after the decomposition of the standard answers s;
- $\otimes$ — binary matching operator, which represents the number of matching substructures in the set of two substructures.

Once the similarity between the answers is obtained, the correctness and completeness of the user answers can be verified by combining it with the corresponding evaluation strategy. Fig. 6 shows an example of verification of user answer to multiple-choice question in SCg-code.

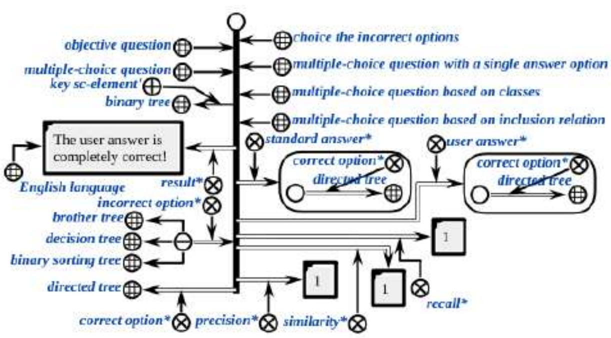


Figure 6. An example of verification of user answer to multiple-choice question.

G. Verification of answers to subjective questions

The approach to calculating the similarity between the semantic graphs of answers to subjective questions, according to the knowledge description structure of the different types of subjective questions, has been divided into: 1. the approach to calculating the similarity between answers to definition explanation questions; 2. the approach to calculating the similarity between answers to proof questions and problem-solving task.

Calculating the similarity between answers to definition explanation questions

The answers to the definition explanation questions are described based on logical formulas (SCL-code). Logic formulas are powerful tools for formal knowledge representation in the framework of OSTIS Technology, which are expanded based on the first-order predicate logic formulas [3], [8]. In the process of calculating the similarity between the semantic graphs of answers to this type of test question, the following tasks need to be solved:

- establishing the mapping relationship of potential equivalent variable sc-node pairs between the semantic graphs of the answers;
- calculating the similarity between semantic graphs;
- if the similarity between semantic graphs is not equal to 1, they also need to be converted to the prenex normal form (PNF) representation separately, and then the similarity between them is calculated again [17].

The semantic graphs for definition explanation questions are built using logical formulas, incorporating variable sc-nodes (bound variables) within the graphs. To compute the similarity between these semantic graphs, a mapping relationship of potential equivalent variable sc-node pairs between them must be established.

In the ostis-systems, the sc-construction composed of sc-tuple, relation sc-node, role relation sc-node and sc-connector is used to describe logical connectives (such as negation ($\neg$) and implication ($\rightarrow$), etc.) and quantifiers (universal quantifier ($\forall$) and existential quantifier ($\exists$)), atomic logic formula (various sc-constructions) or multiple atomic logic formulas that satisfy conjunctive relation are contained in the sc-structure and connected with the corresponding sc-tuple, and these sc-elements together constitute the semantic graph of answers to the definition explanation questions [4]. Its structure is a tree.

If the standard answer and the user answer are exactly equal, it means that the atomic logic formulas with the same semantics between the answers have the same position in the semantic graph. Thus a mapping relationship between variables sc-nodes can be established by determining the position in the semantic graph of each sc-construction containing the variable sc-nodes and the semantic connotation it expresses [18]. Fig. 7 shows some sc-constructions used to describe the information in the knowledge base.

The process of establishing the mapping relationship of the potential equivalent variable sc-node pairs between answers is shown below:

- each sc-tuple and sc-structure in the semantic graph is numbered separately according to the depth-first search strategy (DFS), (for indirectly determining the position of variables sc-nodes in the semantic graph);

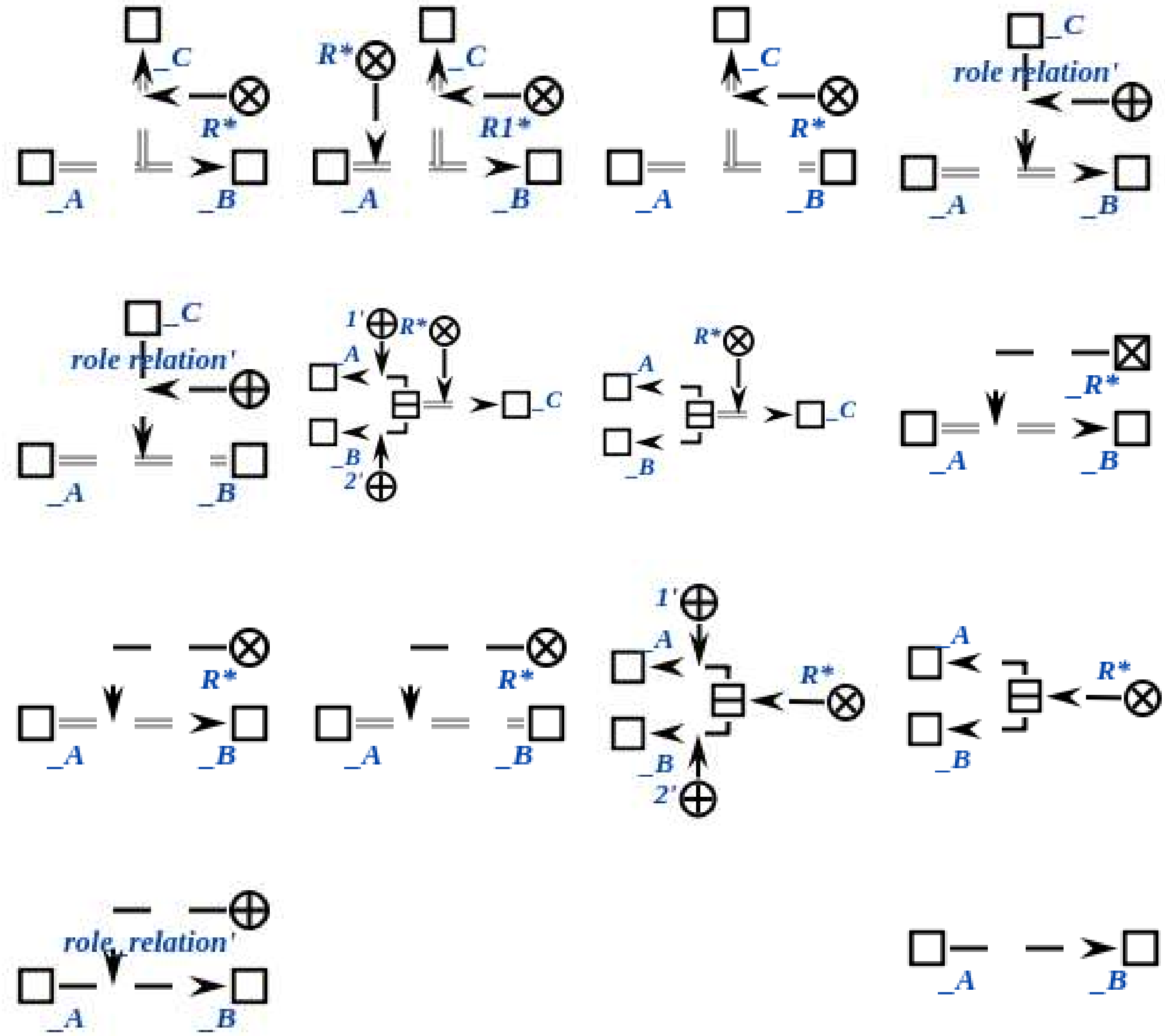


Figure 7. Some examples of sc-constructions.

- according to the matching relationship of each sc-element between each sc-construction pair with the same number in the semantic graph of the standard answer and the semantic graph of the user answer, the mapping relationships of potential equivalent variable sc-nodes pairs between the semantic graphs are established.

When the mapping relationship between the potential equivalent variable sc-node pairs between the semantic graphs is established, the similarity between answers can be calculated, and the detailed calculation process is shown below:

- decomposing the semantic graphs of the answers into substructures according to the structure of the knowledge description;
- establishing the mapping relationship of potential equivalent variable sc-node pairs between the semantic graphs;
- using formulas (3), (4) and (5) to calculate the precision, recall and similarity between semantic graphs.

Fig. 8 shows an example of calculating the similarity between semantic graphs of answers to a definition explanation question in SCg-code.

In Fig. 8, the definition of the inclusion relation is described $(\forall A \forall B((A \subseteq B) \leftrightarrow (\forall a(a \in A \to a \in B)))$.

If the similarity between semantic graphs is not equal 1, it is also necessary to determine whether their logical formulas are logically equivalent. Fig. 9 and 10 show examples of logical equivalence between semantic graphs,

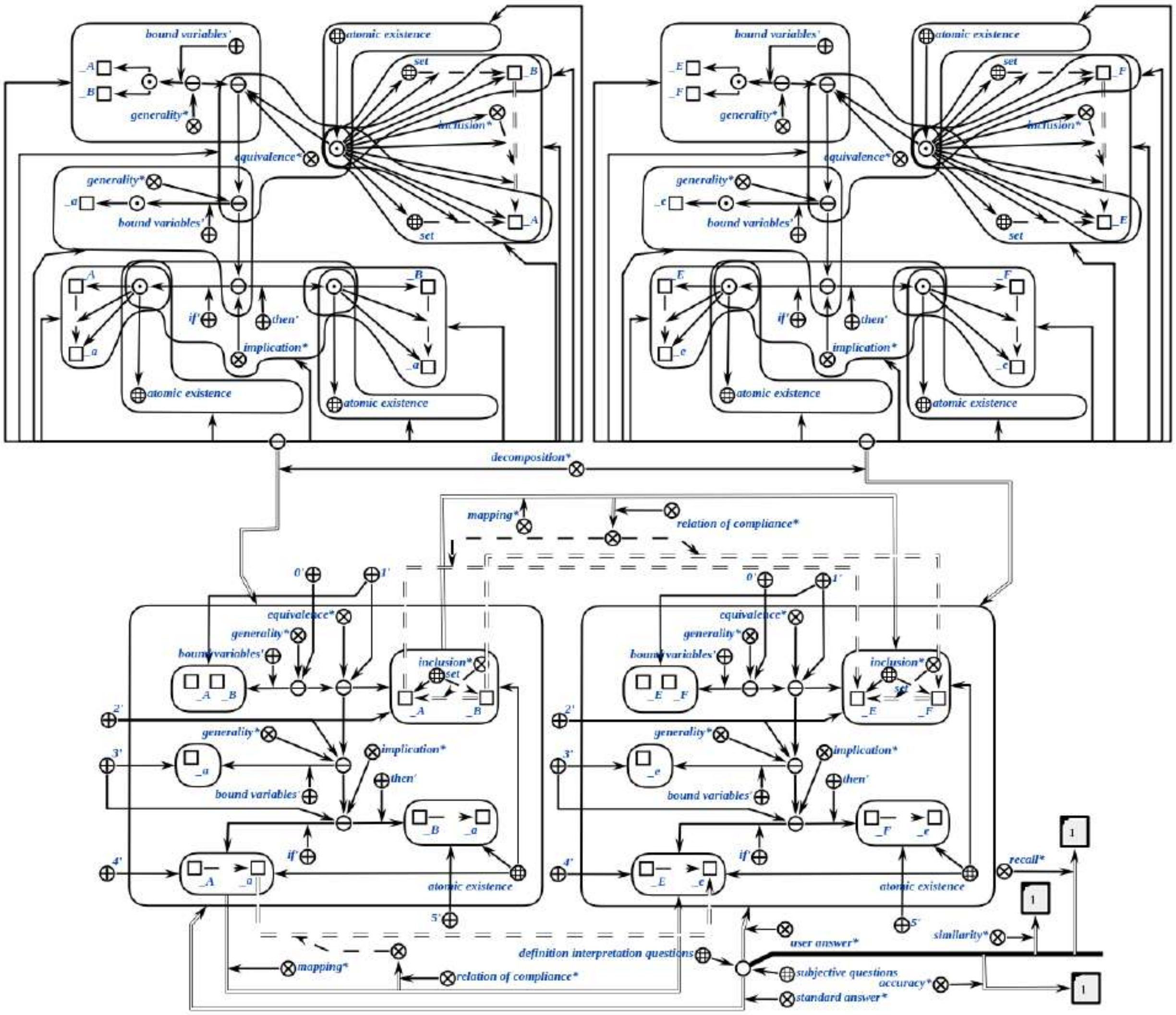


Figure 8. An example of calculating the similarity between semantic graphs of answers to a definition explanation question.

respectively.

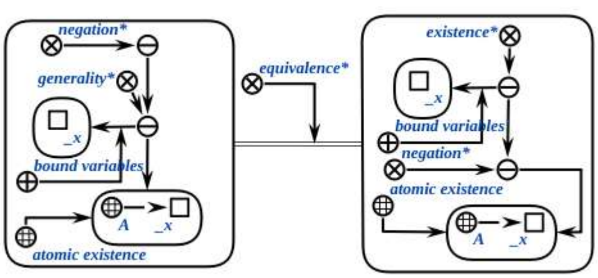


Figure 9. An example of semantic graphs satisfying logical equivalence ($\neg\forall x A(x) \Leftrightarrow \exists x \neg A(x)$).

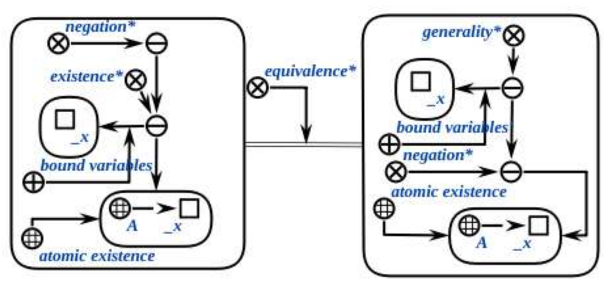


Figure 10. An example of semantic graphs satisfying logical equivalence ($\neg\exists x A(x) \Leftrightarrow \forall x \neg A(x)$).

Therefore, based on the approach to convert predicate logic formulas into PNF and characteristics of logic formulas in ostis-systems, an approach to convert logic formulas into unique (deterministic) PNF according to strict restriction rules is proposed in this article [17], [19]. The strict restrictions mainly include the following:

- renaming rule is preferred when converting logical

formulas to PNF;
- existential quantifier is moved to the front of the logical formula in preference;
- the logical formula can usually be expressed in the following form: $(Q_1x_1Q_2x_2...Q_nx_n(A \leftrightarrow B))$, where $Q_i(i = 1, ...n)$ is a quantifier. A is used to describe the definition of a concept at a holistic level, and it does not contain any quantifiers. B is used to explain the semantic connotation of a definition at the detail level, and it is usually a logical formula containing quantifiers [6], [19]. Therefore, in order to simplify the knowledge processing, it is only necessary to convert the logical formula B to PNF;

The process of converting the semantic graph constructed based on logic formula into PNF descriptions is shown below:

- if there are multiple sc-structures connected by the same conjunctive connective, the sc-constructions contained in them are merged into the same sc-structure;
- eliminating all the implication connectives;
- moving all negative connectives to the front of the corresponding sc-structure;
- using renaming rules so that all bound variables in the semantic graphs are not the same;
- moving all quantifiers to the front of the logical formula;
- merging again the sc-structures in the semantic graphs that can be merged.

Fig. 11 shows an example of converting a semantic graph into PNF representation in SCg-code $(\forall A \forall B((A \subseteq B) \leftrightarrow \forall a(a \in A \rightarrow a \in B)) \Leftrightarrow \forall A \forall B((A \subseteq B) \leftrightarrow \forall a(\neg(a \in A) \vee (a \in B))))$.

It should be emphasized that if the calculated similarity between the semantic graphs of PNF representation is not 1 ($F_{sc} < 1$), the similarity between the semantic graphs calculated for the first time is used as the final answer similarity.

Calculating the similarity between answers to proof questions and problem-solving task

Both proof questions and problem-solving task follow a common task-solving process:

1) the set (Ω) of conditions consisting of some known conditions;
2) deriving an intermediate conclusion using some of the known conditions in Ω and adding it to Ω. Each element in Ω can be regarded as a solving step;
3) repeat step 2) until the final result is obtained [20], [21].

This task-solving process is abstracted as a directed graph, whose structure is in most cases an inverted tree, and is called a reasoning tree (i. e. the reasoning tree of the standard answer). The automatic verification process of user answers to this type of test questions is the same as the traditional manual answer verification process, i.e.,

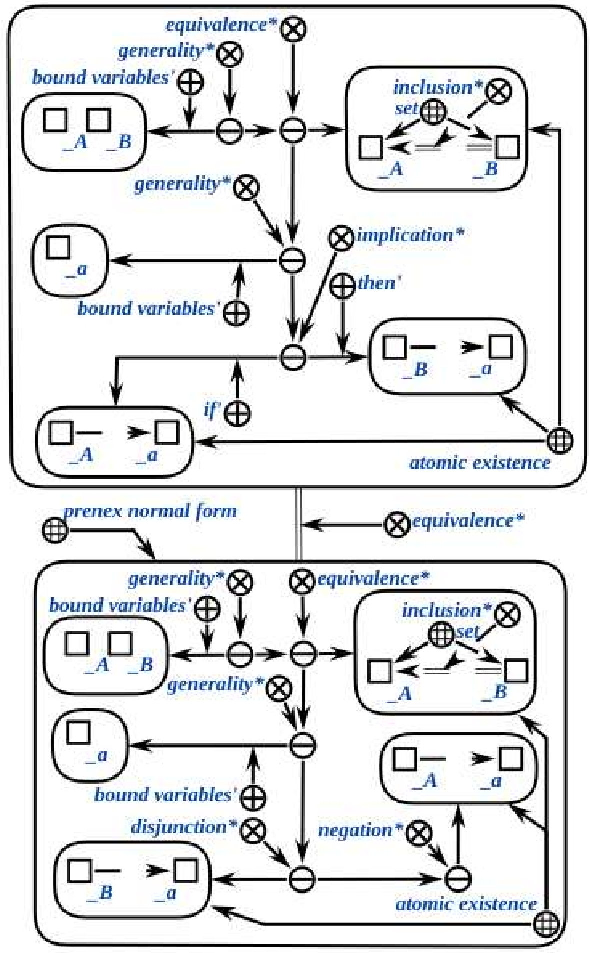


Figure 11. An example of converting a semantic graph into PNF representation.

verifying whether the current solving step of the user answer is a valid conclusion of the partial solving step preceding that step. This means whether the solving step in the user answer corresponding to the parent node in the reasoning tree always is located after the solving steps in the user answer corresponding to the child nodes [4].

The semantic graphs of user answers to proof questions and problem-solving task in the ostis-systems are linear structures consisting of some semantic sub-graphs for describing the solving steps and some semantic fragments for describing the logical order and transformation processes between the semantic sub-graphs. The semantic graph of standard answers to this type of test questions is an reasoning tree consisting of a number of search templates (which can be abstracted as the nodes in the tree). Each search template is constructed using SCL-code in strict accordance with the standard solution steps corresponding to the test question. The search template is used to search in the knowledge base for all semantic fragments corresponding to it [2], [4].

Since the user answers in natural language are converted into semantic graphs they are already integrated with the knowledge already available in the knowledge base. Therefore, when calculating the similarity between the semantic graphs, it is not necessary to consider the differences of the concepts at the natural language level. For example, Segment AB and Segment BA are represented by the same sc-node, they are just two identifiers of the sc-node [4], [16]. An approach to calculate the similarity between the semantic graphs of answers to proof questions and problem-solving task according to the reasoning tree of standard answer (semantic graph of standard answer) is proposed in this article, and the specific calculation process is shown below:

1) numbering each semantic sub-graph in the semantic graph of user answer (the numbering order started from 1);
2) each node in the reasoning tree (search template) is traversed in turn according to the DFS strategy. At the same time, the corresponding semantic sub-graph that is included in the semantic graph of the user answer are searched in the knowledge base using the search template currently being traversed. If such a semantic sub-graph exists, then determine whether the searched semantic sub-graph number is smaller than the semantic sub-graph number corresponding to the search template of the current search template parent node (except for the root node of the reasoning tree), and if so, the searched semantic sub-graph is considered correct;
3) repeat step 2) until all search templates in the reasoning tree have been traversed and the number of correct semantic sub-graphs is counted at the same time;
4) using formulas (3), (4) and (5) to calculate the precision, recall and similarity between answers.

Since this article focuses on presenting the fundamentals of the entire process of automatic generation of test questions to verification of users' answers in a holistic manner, it briefly describes the fundamentals of answer verification for subjective questions. In reference [4] we describe in detail the process of verifying user answers to subjective questions in the ostis-systems.

Once the similarity between the answers to the subjective questions is obtained, the correctness and completeness of the user answers can be verified combined with the evaluation strategy for the subjective questions. The proposed approach to automatic verification of user answers has the following advantages:

- verifying the correctness and completeness of user answers based on semantics;
- the logical equivalence between answers can be determined;
- the similarity between any two semantic graphs in the knowledge base can be calculated;
- the developed component using the proposed approach can be easily transplanted to other ostis-systems.

H. Development of the subsystem knowledge base

The knowledge base of subsystem is mainly used to store automatically generated test questions of various types. Therefore, in order to improve the efficiency of accessing the knowledge base of the subsystem and the efficiency of extracting the test questions, an approach to construct the knowledge base of the subsystem according to the type of test questions and the generation strategy of the test questions is proposed in this article.

The basis of the knowledge base of any ostis-system (more precisely, the sc-model of the knowledge base) is a hierarchical system of subject domains and their corresponding ontologies [2], [3], [8]. Let's consider the hierarchy of the knowledge base of subsystem in SCn-code:

Section. Subject domain of test questions
⇐ *section decomposition**:
  {• *Section. Subject domain of subjective questions*
    ⇐ *section decomposition**:
      {• *Section. Subject domain of definition explanation question*
       • *Section. Subject domain of proof question*
       • *Section. Subject domain of problem-solving task*
      }
   • *Section. Subject domain of objective questions*
    ⇐ *section decomposition**:
      {• *Section. Subject domain of multiple-choice question*
       • *Section. Subject domain of fill in the blank question*
       • *Section. Subject domain of judgment question*
      }
  }

It should be emphasised here that objective questions can be divided into more specific types (e.g., multiple-choice question with multiple answer options and multiple-choice question with a single answer option, etc.) according to their characteristics and the corresponding test question generation strategy. In references [4] and [5] we describe in detail the categorisation of objective questions and the process of construction of their subject domains.

I. Development of problem solver

One of the most important components of every intelligent system is the problem solver, which provides the ability to solve a variety of problems. The problem solver of any ostis-system (more precisely, the sc-model of the ostis-system problem solver) is a hierarchical system of knowledge processing agents in semantic memory (sc-agents) that interact only by specifying the actions they perform in the specified memory [2], [3].

Therefore, a problem solver for automatic generation of test questions and automatic verification of user answers has been developed based on the proposed approach, and its hierarchy is shown below in SCn-code:

Problem solver for the automatic generation of test questions and automatic verification of user answers
$\Leftarrow$ *decomposition of an abstract sc-agent**:
{
- *Sc-agent for automatic generation of test questions*
  $\Leftarrow$ *decomposition of an abstract sc-agent**:
  {
  - *Sc-agent for quick generation of test questions and test papers*
  - *Sc-agent for generating single type of test questions*
  - *Sc-agent for generating a single test paper*

  }
- *Sc-agent for automatic verification of user answers*
  $\Leftarrow$ *decomposition of an abstract sc-agent**:
  {
  - *Sc-agent for automatic scoring of test papers*
  - *Sc-agent for calculating similarity between answers to objective questions*
  - *Sc-agent for calculating the similarity between answers to definition explanation questions*
  - *Sc-agent for converting a logical formula into PNF*
  - *Sc-agent for calculating the similarity between the answers to proof questions and problem-solving task*

  }

}

The function of the sc-agent for quick generation of test questions and test papers is to automate the entire process from test question generation to test paper generation by initiating the corresponding sc-agents (sc-agent for generating single type of test questions and sc-agent for generating a single test paper).

The function of the sc-agent for automatic scoring of test papers is to implement automatic verification of user answers to test questions and automatic scoring of test papers by initiating sc-agents for calculating the similarity between user answers and sc-agents for converting a logical formula into PNF.

IV. Evaluation of the effectiveness of the developed subsystems

The effectiveness of the developed subsystem will be evaluated from the following aspects:

- availability of the generated test questions;
- closeness between automatic scoring and manual scoring of user answers to subjective questions;
- reduced time costs due to the use of subsystem.

In order to assess the availability of automatically generated test questions using the knowledge base, 200 automatically generated test questions were taken from the discrete mathematics tutoring system and the euclidean geometry tutoring system, respectively. The percentage of test questions that could be used directly was also manually counted. The evaluation results of the availability of test questions showed that 94 $\%$ of the automatically generated test questions in the discrete mathematics tutoring system and 92 $\%$ in the euclidean geometry tutoring system could be used directly.

In order to evaluate the closeness between the results of automatic scoring and manual scoring of user answers to subjective questions, the following work was done. Firstly, 40 second year university students were randomly selected and their knowledge was tested using an automatically generated test paper containing 4 subjective questions (the maximum score for each subjective question is 10 grades). The answers of 40 students were then checked using the subsystem and manual methods respectively and the error between the automatic and manual assessment results of these students' answers was calculated (Table I).

Table I
Table. Results of scoring error statistics for user answers to subjective questions

Error range (Φ)	Definition explanation question 1	Definition explanation question 2	Proof question 1	Proof question 2	Total	Proportion
$\Phi \leq 1$	35	31	26	28	120	75 %
(1-1.5]	2	4	8	8	22	13.75 %
(1.5-2]	2	3	5	3	13	8.125 %
$\Phi > 2$	1	2	1	1	5	3.125 %

The formula for calculating the error Φ is shown below (6):

$$\Phi = |x - y| \tag{6}$$

The parameters are defined as shown below:

- x — is the manual scoring of user answers to the test questions;

- y — is the automatic scoring of user answers to the test questions;

From the Table I, it can be seen that the automatic scoring and manual scoring of user answers to subjective questions in the tutoring system for discrete mathematics generally remained consistent, and that when the maximum score for a subjective question was 10, the sample size with an error $\Phi \leq 1.5$ between scores was over 88 %.

In order to evaluate the reduced time cost due to the use of the subsystem, a random sample of a class of second-year university students (30 students in total) was selected and the average duration of the process of testing the knowledge level of these students using the subsystem was calculated, as well as the duration of the process of testing the knowledge level of these students using the traditional examination method, respectively. Since it is possible to test the knowledge level of 30 students at the same time using the subsystem, the average duration of the testing process was counted. The evaluation result of reduced time cost showed that the use of the developed subsystem in the process of testing the knowledge level of 30 students can save 89 % time cost.

The duration of the development of the knowledge base should normally also be taken into account when assessing the reduced time cost due to the use of the subsystem. In reference [22], the duration of development a knowledge base for discrete mathematics using the reusable component approach was estimated to be 513 h. Therefore, the relationship between the average duration of generating each test paper using the developed subsystem and the number of test papers generated is evaluated in this article, taking into account the duration of the development of the knowledge base. The experimental result shows that when the number of test papers generated using the developed subsystem is more than 20,000, the impact of the development time of the knowledge base on the average duration of each test paper generated using the subsystem is negligible. It should be emphasised that the use for generating test questions is only one use of the knowledge base, and even if it is only used for generating test questions it can be seen from the evaluation result that it is profitable.

The above experimental results show that the developed subsystem can satisfy the conditions for practical applications.

V. Conclusion

An approach to automatic generation of test questions using knowledge bases and automatic verification of user answers according to the semantic description structures of the knowledge in an intelligent tutoring system developed using OSTIS Technology is presented in this article. A subsystem for automatic generation of test questions and automatic verification of user answers has been developed in the ostis-systems based on the proposed approach. Using the developed subsystem allows automation of the entire process from generation of test questions to grading of test papers, which greatly improves the learning efficiency of the users.

Finally the effectiveness of the developed subsystem was evaluated in terms of the availability of the generated test questions, the closeness between the automatic scoring and the manual scoring of the test questions in the discrete mathematics ostis-system and reduced time costs due to the use of subsystem. From the evaluation results, it can be seen that the developed subsystem can meet the conditions for practical application.

Acknowledgment

The work in this article was done with the support of research teams of the Department of Intelligent Information Technologies of Belarusian State University of Informatics and Radioelectronics. Authors would like to thank every researcher in the Department of Intelligent Information Technologies.

References

[1] Jiang. H., Zhao. Y., Li. X. Intelligent education system: A comprehensive review. IEEE Access, 2020, Vol. 8, pp. 201016-201026.

[2] V. V. Golenkov and N. A. Guljakina, "Proekt otkrytoj semanticheskoj tehnologii komponentnogo proektirovanija intellektual'nyh sistem. chast' 1: Principy sozdanija project of open semantic technology for component design of intelligent systems. part 1: Creation principles]," Ontologija proektirovanija [Ontology of design], no. 1, pp. 42–64, 2014.

[3] V. Golenkov, N. Gulyakina, "Next-generation intelligent computer systems and technology of complex support of their life cycle," in Otkrytye semanticheskie tehnologii proektirovanija intellektual'nyh sistem [Open semantic technologies for intelligent systems], V. Golenkov, Ed., BSUIR. Minsk, BSUIR, 2022, pp. 27–40.

[4] Wenzu. Li, "A semantics-based approach to automatic generation of test questions and automatic verification of user answers in the intelligent tutoring systems," in Otkrytye semanticheskie tehnologii proektirovanija intellektual'nyh sistem [Open semantic technologies for intelligent systems], V. Golenkov, Ed., BSUIR. Minsk, BSUIR, 2022, pp. 369–374.

[5] Wenzu Li. "An automated approach to checking user knowledge levels in intelligent tutoring systems," in Otkrytye semanticheskie tehnologii proektirovanija intellektual'nyh sistem [Open semantic technologies for intelligent systems], V. Golenkov, Ed., BSUIR. Minsk, BSUIR, 2023, pp. 269–278.

[6] Li. W., Qian L. Semantic Approach to User Answer Verification in Intelligent Tutoring Systems. Communications in Computer and Information Science, 2022, Vol. 1625, pp. 242–266.

[7] Li. H. Research on item automatic generation based on DL and domain ontology. Journal of Changchun University of Technology (Natural Science Edition), 2012, Vol. 33(04), pp. 461-464.

[8] (2024, Mar) Ims.ostis metasystem. [Online]. Available: https://ims.ostis.net

[9] Andreas. P., Konstantinos. K., Konstantinos. K. Automatic generation of multiple-choice questions from domain ontologies. In: IADIS International Conference e-Learning, 2008, pp. 427-434.

[10] Kaur. M., Batra. S. Ontology-based question generation: a review. Journal of Intelligent and Fuzzy Systems, 2020, Vol. 39(01), pp. 49-57.

[11] Arjun. S. B., Manas. K., Sujan. K. S. Automatic Generation of Multiple Choice Questions Using Wikipedia. International Conference on Pattern Recognition and Machine Intelligence, 2013, Vol. 8251, pp. 733-738.

[12] Chen. C. H. A comprehensive survey of text similarity measures. Journal of Information Science, 2019, Vol. 45(01), pp. 1-19.
[13] Singh. A., Singh. V. K. Text similarity techniques: a systematic literature review. International Journal of Computer Applications, 2019, Vol. 182(41), pp. 23-30.
[14] Mingyu. Ji., Xinhai. Zhang. A Short Text Similarity Calculation Method Combining Semantic and Headword Attention Mechanism. Scientific Programming, 2022.
[15] Anderson. P., Fernando. B., Johnson. M., Gould. S. Spice: Semantic propositional image caption evaluation. In: European Conference on Computer Vision, Springer, 2016, pp. 382-398.
[16] Longwei. Qian, "Ontological Approach to the development of natural language interface for intelligent computer systems," in Otkrytye semanticheskie tehnologii proektirovanija intellektual'nyh sistem [Open semantic technologies for intelligent systems], V. Golenkov, Ed., BSUIR. Minsk, BSUIR, 2022, pp. 217–238.
[17] Fujiwara. M., Kurahashi. T. Prenex normal form theorems in semi-classical arithmetic. The Journal of Symbolic Logic, 2022, pp. 1-31.
[18] Liu. J., Liu. Y., Wang. X., Wang. B. A survey on ontology integration techniques. Future Generation Computer Systems, 2020, Vol. 105, pp. 17-30.
[19] Pan. M., Ding. Z. A simple method for solving prenex disjunction (conjunction) normal forms. Computer Engineering and Science, 2008, Vol. 30(10), pp. 82-84.
[20] Ge. L., Li. X. An overview of the methods for automated theorem proving in geometry. Journal of Symbolic Computation, 2016, Vol. 74, pp. 303-335.
[21] Kryakin. A., Nguyen. P. Automated theorem proving in geometry using Gröbner bases. Journal of Automated Reasoning, 2014, Vol. 53(02), pp. 163-189.
[22] Davydenko, I. "Semantic models, method and tools of knowledge bases coordinated development based on reusable components," in Otkrytye semanticheskie tehnologii proektirovanija intellektual'nyh sistem [Open semantic technologies for intelligent systems], V. Golenkov, Ed., BSUIR. Minsk, BSUIR, 2018, pp. 99–118.

ПОДХОД К АВТОМАТИЗАЦИИ ВСЕГО ПРОЦЕССА КОНТРОЛЯ ЗНАНИЙ УЧАЩИХСЯ ОТ ГЕНЕРАЦИИ ТЕСТОВЫХ ВОПРОСОВ ДО ПРОВЕРКИ ОТВЕТОВ ПОЛЬЗОВАТЕЛЕЙ В ИНТЕЛЛЕКТУАЛЬНЫХ ОБУЧАЮЩИХ СИСТЕМАХ

Ли Вэньцзу

Данная статья посвящена вопросам автоматической генерации тестовых вопросов и автоматической проверки ответов пользователей в интеллектуальных системах нового поколения, основанных на смысловом представлении информации. В статье подробно рассмотрен подход к реализации автоматической генерации тестовых вопросов различных типов с использованием баз знаний и автоматической проверки ответов пользователей в соответствии со структурой семантического представления знаний.

Keywords—генерация тестовых вопросов, проверка ответов пользователей, Технология OSTIS, интеллектуальные системы на основе смыслового представления информации, семантическая структура, база знаний

Received 29.03.2024

Modeling the State of Information Security of a Smart Campus

Alexsander Sobol
Center of Information technologies
Belarussian State University
Minsk, Belarus
Email: sobolam@bsu.by

Viktor Kochyn
Vice-Rector for Academic Affairs and
Internationalization of Education
Belarussian State University
Minsk, Belarus
Email: kochyn@bsu.by

***Abstract*—The state of information security system in smart campuses is the main source of obtaining reliable events. This paper discusses the modeling of the state of information security system of smart campus. Special attention is paid to the smart campus infrastructure on which the state of the information security system depends, the interrelationships between the levels of the smart campus, and the subsystems of the information security system are highlighted.**

***Keywords*—smart campus, information security, state modelling**

I. Introduction

Technology has become ubiquitous in modern society, and education is no exception. A smart campus is an innovative educational institution that utilises advanced technologies to create a stimulating learning environment.

The goal of a smart campus is to provide an intelligent environment that supports active learning, encourages creative thinking, and provides access to advanced educational resources and technologies. A smart campus employs technologies such as the Internet of Things (IoT), artificial intelligence (AI), cloud computing, and data analytics to create an integrated and effective educational environment .

The objective of a smart campus is to create a learning environment that adapts to the needs of students and faculty, provides access to state-of-the-art educational resources and technologies, and supports a variety of learning and assessment methods. A smart campus provides innovative teaching methods, including online courses, virtual labs, and simulations, as well as various forms of feedback and learning assessment analysis based on machine learning [1], [2], [3].

However, the increasing number of interactive devices and networking technologies also brings about an increase in vulnerabilities and threats that pose information security issues for smart campuses. One example of such challenges is cyberattacks: Smart campuses may use Internet of Things (IoT) networks, which are vulnerable to cyberattacks such as DDoS attacks, IoT device hacking, or malware injection.

Additionally, there is a risk of sensitive or personal data leakage due to the large amount of data collected by smart campuses. Finally, physical security is also a concern. Unauthorized access to the physical infrastructure of a smart campus, such as surveillance cameras, sensors, and building management systems, can result in severe consequences, including vandalism, data theft, and privacy breaches.

Smart campus networks can also have vulnerabilities in their infrastructure that attackers can exploit for unauthorized data access or security breaches. Additionally, social engineering is a potential threat. Social engineering attacks may deceive smart campus personnel and allow unauthorized access to systems or information. Additionally, DDoS attacks can cause critical smart campus systems to experience denial of service, which can disrupt the learning process and campus operations. To address these issues, a comprehensive approach to security is required. This includes installing up-to-date information security tools, regularly updating software, providing staff training, ensuring physical security, and monitoring network activity to quickly detect and respond to threats. Initially, to build the information security of a smart campus, it is necessary to define the structure of both the smart campus itself and the components of the information security system [4], [5], [12], [13], [14] .

II. Components of a smart campus

A smart campus comprises several components, each of which plays a crucial role in creating an innovative educational environment. The following components can be identified for a smart campus.

A. Video surveillance systems

Video surveillance systems play an important role in providing physical security on campus. They provide constant surveillance of the campus and can be used to detect and prevent crime and ensure the safety of students and staff.

B. Access Control Systems

Access control systems provide access control to various areas of a smart campus. They can be used to control student and staff access to specific buildings, facilities, or resources.

C. Telecommunications Systems

Telecommunications systems facilitate communication and data exchange between different devices and systems on campus. They can provide Internet access, messaging, data sharing, online conferencing, and webinars [7].

D. Data processing systems

Data processing systems are utilised to analyse and process data collected by various systems in a smart campus. They can be used to analyse data on safety, student and staff activity, and to identify patterns and trends. Additionally, this system stores data on student and faculty performance [8].

E. Information Security Systems

Information security systems are employed to safeguard data and information in a smart campus. They are capable of detecting and preventing cyberattacks, protecting sensitive information, and securing the network.

"Fig. 1" presents the interconnection between the smart campus systems.

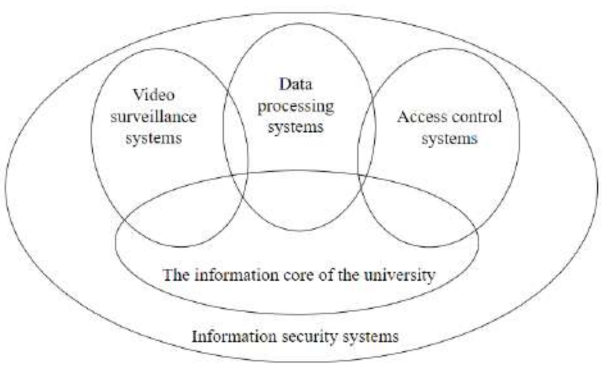


Figure 1. Relationship between smart campus systems.

The components of each of the systems are located at different interaction layers of the smart campus. Based on this, the smart campus can be visualised as a multi-layered architecture as shown in "Fig. 2".

The sensor layer provides event enrichment with additional information required for the intrusion detection and prevention system.

The access layer refers to the network resource access control layer and defines which users or devices are authorised to access certain network resources or functions.

The distribution layer provides connectivity between the various segments and subnets of the campus network.

The information core layer ensures the operability of business processes that exist in higher education institutions such as: distance learning system, university management, etc.

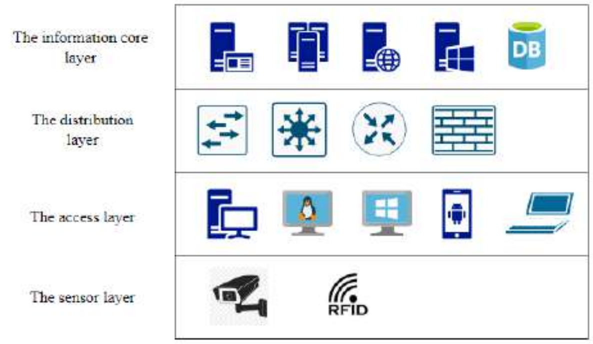


Figure 2. Multi-level structure of a smart campus

Developing a state model for a smart campus is a pressing task.

Each smart campus system can be in different states at any given time and may depend on the state of other systems. Assuming that $V(t)$ represents the state of the video surveillance system over time. the following formula at any given time:

$$\frac{dV(t)}{dt} = f(V, T, D, S) \tag{1}$$

where f a function describing the relationship between the telecommunications system T, the data processing system D , the video surveillance system V and the information security system S .

The state of the access control system at any point in time can be described by the following formula:

$$\frac{dA(t)}{dt} = j(A, T, D, S) \tag{2}$$

where $A(t)$ a state of access control system, which depends on time t, j a function describing the relationship between telecommunication system T , access control system A , data processing system D and information security system S .

The state of the telecommunication system at any point in time can be described by the following formula:

$$\frac{dT(t)}{dt} = h(T, S) \tag{3}$$

where $T(t)$ a state of the telecommunication system, which depends on time t, h a function describing the relationship between telecommunication system T and information security system S.

The state of the data processing system at any point in time can be described by the following formula:

$$\frac{dD(t)}{dt} = g(D, T, S) \tag{4}$$

where $D(t)$ a state of data processing system, which depends on time t , g a function describing the relationship between telecommunication system T, data processing system D and information security system S.

The state of the information security system at any point in time can be described by the following formula:

$$\frac{dS(t)}{dt} = l(A, V, D, T, S) \tag{5}$$

where $S(t)$ a state of the information security system, which depends on time t, l a function describing the relationship between the telecommunication system T, access control system A , data processing system D , the video surveillance system V and information security system S .

Based on the above equations (1) to (6), the state of the smart campus infrastructure at any point in time can be expressed by the following expression:

$$\frac{dCam(t)}{dt} = q(A, V, D, T, S) \tag{6}$$

where $Cam(t)$ is the state of the smart campus infrastructure at time t .

III. Components of the information security system

Information security system [8] is one of the key systems in the smart campus infrastructure. This system should perform:

- Personal data protection: The smart campus collects and processes large amounts of data about students, faculty and staff. This data may include a variety of personal data such as names, addresses, financial data, academic performance, etc. The information security system must protect personal data from unauthorized access and use.
- Intellectual Property Protection: A smart campus may contain intellectual property in the form of research data, scientific articles, software, etc. The information security system should protect this intellectual property from leaks and theft.
- Protection against cyber attacks: Smart Campus may be susceptible to various types of cyber attacks such as viruses, malware, phishing, etc. The information security system should provide protection against these attacks and respond to them in real time.
- Ensuring Continuity of Operations: The smart campus must be available and operational at all times. The information security system must ensure continuity of system operations and quick recovery from failures or attacks.
- Regulatory Compliance: A smart campus may have to comply with various regulatory and legislative requirements for information security, depending on the country of implementation of that campus. The information security system must ensure compliance with these requirements and provide for auditing and reporting.
- Protection against internal source threats: A smart campus may be vulnerable to threats from internal sources such as students, faculty, and staff. The information security system must protect against these threats and control access to sensitive data.

An information security system according to international standards namely ISO/IEC 27000 and NIST-800 [9], [10] should contain the following subsystems:

- Administration subsystem: This subsystem is responsible for managing all security aspects of the system. This includes creating and deleting user accounts, configuring access rights, managing network settings, managing security certificates, and other administrative functions. The administration subsystem is also responsible for educating users about information security and ensuring their compliance with established rules and policies.
- Collection and Filtering Subsystem: This subsystem is responsible for collecting data on security events such as login attempts, attempts to access protected resources, changes to system settings, etc. The collected data is then filtered to remove unnecessary information and prepare the data for further analysis.
- Data Analysis Subsystem: This subsystem is responsible for analyzing the collected security event data to identify potential threats and anomalies. The analysis may include the use of machine learning algorithms, statistical methods, comparison with attack patterns, etc. The analysis generates reports on potential threats and anomalies that require further investigation.
- Response Subsystem: This subsystem is responsible for responding to detected threats and anomalies. This may include automatic actions such as blocking access to protected resources, alerting the system administrator, etc. The response subsystem can also offer recommendations for further action to the system administrator.
- Security Audit Subsystem: This subsystem is responsible for logging and analyzing user activity on the system to identify potential security violations and ensure compliance with security standards. Security auditing may include checking compliance with established security policies, analyzing event logs, etc.
- Secure Operation Subsystem: This subsystem is responsible for ensuring the continuous and secure operation of the system. This includes protecting against failures, providing redundancy, monitoring system status, etc.
- Backup Subsystem: This subsystem is responsible for backing up data and setting up recovery mech-

anisms in case of data loss or corruption. This includes regularly backing up data, verifying its integrity and recovery availability.
- Testing subsystem: This subsystem is responsible for conducting penetration tests, vulnerability analysis, and other testing to identify potential threats and anomalies. This includes the use of specialized tools and techniques to detect vulnerabilities and verify system security.

Interrelationships between subsystems of the information security system can be represented using graph formalism. Each subsystem can be represented as a vertex of the graph, and the interactions between subsystems can be represented as edges of the graph. For example, if subsystem A interacts with subsystem B, there will be a graph edge between the vertices A and B.

Then let the administration subsystem be vertex A , the collection and filtering subsystem be vertex B , the data analysis subsystem be vertex C , the response subsystem be vertex D, the security audit subsystem be vertex E , the secure operation subsystem be vertex F , the redundancy subsystem be vertex G , and the testing subsystem be vertex H .

Let $G = (V, I)$ be a graph where V is the set of vertices and I is the set of edges. Then the set of vertices representing the subsystems will be as follows

$$V = \{A, B, C, D, E, F, G, H\} \tag{7}$$

And the set of vertices will look like $I = \{(A, B), (A, C), (B, D), (C, D), (C, E), (D, F), (E, G), (F, G), (G, H)\}$.

The smart campus information security system will depend on the state of each subsystem depending on current conditions and events.

It is important to note that the states of subsystems can change depending on current conditions and events, and system administrators must monitor the state of each subsystem and take the necessary steps to restore normal operation when necessary.

In order to develop a hardware and software system of information security, it is necessary to mathematically describe the states of each subsystem. This can be done using the theory of automata and finite states. Each subsystem can be represented as an automaton with a finite number of states, where each state corresponds to a certain functional state of the subsystem. Transitions between states can be triggered by external events such as subsystem failures, administrative decisions, changes in the environment, etc.

The following states can be defined for each subsystem:

- Active state (Active): The subsystem is functioning as defined and performing its functions.
- Inactive: The subsystem is temporarily disabled or not functioning due to technical problems or administrative decisions.
- Recovery state: The subsystem is in the process of recovering from a failure or shutdown.

Similar states can be defined for an information security system. Based on this, represent the information security system as a tuple:

$$(S, \Sigma, \delta, s_0, F) \tag{8}$$

where S a the set of states of the subsystem, Σ a the set of input events (e.g., failure events, administrative decisions, changes in the environment), δ a the transition function, which defines which events lead to a transition from one state to another s_0, a the initial state of the subsystem and F a the set of final states, which denote the successful completion of the subsystem.

The set of states of the subsystems of the smart campus information security system can be described as follows:

$$S = \{\mathit{Active}, \mathit{Inactive}, \mathit{Recovery}\} \tag{9}$$

The set of input events of the smart campus information security system can be represented as follows:

$$\Sigma = \{Failure, Administrationdescision, Change\} \tag{10}$$

The transition functions between the states of the information security system based on formulas (9) and (10) will be as follows:

$$\delta(\mathit{Active}, Failure) = \mathit{Inactive} \tag{11}$$

$$\delta(\mathit{Inactive}, \mathit{Recovery}) = \mathit{Recovery} \tag{12}$$

$$\delta(\mathit{Inactive}, Succesfulrecovery) = \mathit{Active} \tag{13}$$

Initial status of subsystems of information security systems of the smart campus:

$$s_0 = \mathit{Active} \tag{14}$$

The set of final states of the smart campus information security system that denote the successful completion of the subsystem:

$$F = \{\mathit{Active}\} \tag{15}$$

On the basis of the given mathematical descriptions it is possible to model the information infrastructure of the smart campus and information security systems. These models may allow to initially develop the infrastructure of the smart campus and modify the above expressions, as well as allow to develop the software implementation of the information security system.

Taking into account the fact that expression (5) is presented as a differential equation, and the information security system itself represents a graph with states described by expressions (7,8), the function describing the state of the information security system can be presented in the following form:

$$S(t) = \begin{cases} 1, \text{if the system in time } t \text{ Active} \\ 0, \text{if the system in time } t \text{ Inactive} \\ -1, \text{if the system in time } t \text{ Recovery} \end{cases} \quad (16)$$

However, this function is non-differentiable because it is not smooth and continuous. Therefore, using the Fourier transform, we present this formula in the following form.

IV. A model of an intruder in a smart campus

An intruder model is an abstract representation of potential threats that could be directed at an information system, network, application, or organization. It is a conceptual description of how attackers might attempt to penetrate a system or damage its operation or security.

The intruder model includes descriptions of the different types of attacks, intrusion techniques, vulnerabilities, and other factors that can be exploited by attackers. It can be used to assess the level of vulnerability of a system or network and to develop strategies to protect against potential threats.

The importance of an intruder model is that it helps in analyzing and understanding possible attack scenarios, which enables organizations to take measures to secure, prevent and respond to incidents when necessary. Such models are an important tool in cybersecurity and help improve the security of information systems and networks.

There are many types of intruders that may attempt to infiltrate a smart campus for different purposes. For smart campus infrastructure, the following intruders can be identified:

1) Hackers:
   - Ethical hackers: Information security professionals who use their skills to test systems for vulnerabilities and help organizations improve their defenses.
   - Unethical hackers: Attackers who attempt to infiltrate systems to cause damage, steal data, or engage in other illegal activities.
2) Scammers:
   - Phishing attacks: Attackers send false emails or create fake websites to trick users into accessing their sensitive information.
   - Social engineering: Attackers may use manipulation and deception to convince employees of an organization to grant them access to a system or sensitive information.
3) Internal Intruders:
   - Employees: Individuals with legitimate access to a system or data may abuse their privileges to gain unauthorized access or leak confidential information.
   - Compromised Accounts: Attackers can gain access to employee accounts through vulnerabilities or phishing attacks.
4) Spammers and Vulnerability Scanners:
   - Spammers: Attackers who use mass spamming to spread malware or phishing attacks.
   - Vulnerability scanners: Software tools that automatically scan networks for vulnerabilities for use in attacks.

Depending on the type of intruder, there may be the following popular attacks, which can correspond to several types of intruders at once:

1) Denial of Service (DoS/DDoS):
   - DoS (Denial of Service): An attacker may attempt to saturate the available resources of a smart campus network or servers by sending a large number of requests to a server, resulting in a denial of service for other users.
   - DDoS (Distributed Denial of Service): Multiple computers captured in a botnet simultaneously attack the smart campus network, causing it to become overloaded and unable to serve legitimate traffic.
2) Authentication and identification methods:
   - Login and password hijacking: Attackers may attempt to intercept smart campus user credentials through network traffic hijacking techniques or the use of malware such as spyware or keyloggers.
   - Cross-network spoofing (MITM): An attacker can use a MITM attack to interfere with communication between devices and intercept or alter transmitted data, including authentication data.
3) OS and application vulnerabilities: Attackers could exploit known or newly discovered vulnerabilities in operating systems or applications installed on smart campus devices to gain unauthorized access or perform malicious activities.
   These are only a fraction of the attacks that can be on a smart campus. Various modeling techniques can be used to design the information security system of a smart campus.

Various modeling techniques can be used to design the information security system of a smart campus.

V. Modeling information security systems

The expressions above, can allow modeling the information security system and building the virtual infrastructure of the smart campus. When modeling, it is necessary to consider the subsystems of the information security system as well as the relationship between the components of the smart campus. For each subsystem of the information security system and components of the smart campus, the following tools and approaches can be identified:

1) Simulators and modeling tools: There are specific software tools such as NS-3 or OMNeT++ that can be used to model networks and communication protocols including information security aspects [15].
2) Risk and Vulnerability Analysis Programs: These software tools help to identify vulnerabilities in the system and assess the likelihood and consequences of various cyber attacks.
3) Modeling and specification languages: Some programming languages such as Alloy or Promela can be used to formalize system models and perform formal security analysis of the system.
4) Visualization tools: Visualization tools such as data flow diagrams or attack diagrams can help in analyzing vulnerabilities and developing remediation measures.
5) Configuration management and monitoring systems: Using configuration management and monitoring systems such as Splunk or ELK Stack can help detect anomalies and attacks in real time.
6) Attack modeling tools: Some software tools, such as the Metasploit Framework, provide tools to simulate different types of cyberattacks and assess the security level of a system.

References

Building a mathematical model of information security system state is a very complex task and can be solved by various methods: differential equations, Markov chains, automata theory, etc.

For comprehensive information security in a smart campus, it is necessary to use semantic technologies that allow integrating different types of knowledge and problem solving models, and the corresponding approaches are developed within the framework of OSTIS technology. Also with the help of OSTIS technology [16] it is possible to create a database of vulnerabilities and promptly react to cyber incidents in a smart campus.

To combine these systems it is necessary to use the theory of hybrid systems. In addition to using this theory, in order to combine the different mathematical models of the information security system, it is necessary to consider the information data flows that are generated by the smart campus itself. In order to successfully create a technical interpretation of the mathematical model of the information security system, the threat model and the intruder model of the smart campus must be considered.

References

[1] A. Smith and B. Johnson, "Implementing Smart Campus Security Systems: Challenges and Opportunities," Journal of Smart Cities, vol. 5, no. 2, pp. 123-135, 2020.

[2] C. Chen and D. Wang, "Design and Implementation of a Smart Campus Security Monitoring System Based on IoT Technology," IEEE Access, vol. 7, pp. 100342-100352, 2019.

[3] H. Lee and S. Park, "A Review of Smart Campus Security: A Case Study Approach," International Journal of Security and Networks, vol. 13, № 3, pp. 137-150, 2018.

[4] R. Gupta and M. Chandrasekaran, Smart Campus Security: Solutions for the 21st Century. Springer, 2020.

[5] Y. Zhang and K. Lee, Building Smart Campus Security Systems: A Comprehensive Guide. Wiley, 2019.

[6] Sobol A., Kochyn V., Grakova N. Mathematical Methods for Assessing Information Security Risks / Open Semantic Technologies for Intelligent Systems. – 2023. – Iss. 7. – pp. 317-322.

[7] Kochyn, V.P. Designing a secure fail-safe cloud repository of paperworks of students and employees of educational institutions / V.P. Kochyn, A.V. Zherelo // Journal of the Belarusian State University. Mathematics and Informatics. – 2021. – № 3. – P. 104-108.

[8] Ponemon Institute, "Smart Campus Security: Trends and Challenges," Research Report.

[9] ISO/IEC 27001:2022, "Information security, cybersecurity and privacy protection. Information security management systems."

[10] National Institute of Standards and Technology (NIST), "Guidelines for Smart Campus Security Systems," NIST Special Publication 800-123.G. Eason, B. Noble, and I. N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955. (references)

[11] S. McClure, J. Scambray, and G. Kurtz, "Hacking Exposed". [Online]. Available: https://www.mheducation.com/highered/product/hacking-exposed-mcclure-scambray-kurtz/M0071780289.html. [Accessed: February 19, 2024].

[12] Y. Diogenes and E. Ozkaya, "Cybersecurity: Attack and Defense Strategies". [Online]. Available: https://www.oreilly.com/library/view/cybersecurity-attack-and/9781838827793/. [Accessed: February 21, 2024]..

[13] D. Kaafar, E. Panaousis, and G. Loukas, "Smart Cities Cybersecurity and Privacy". [Online]. Available: https://www.elsevier.com/books/smart-cities-cybersecurity-and-privacy/kaafar/978-0-12-815032-0. [Accessed: February 11, 2024].

[14] M. Brown and V. Sivaraman, "A survey of ns-3", in Proceedings of the 2011 ACM SIGCOMM Conference on Internet Measurement Conference, 2011, pp. 1-14, doi: 10.1145/2068816.2068841.

[15] C. Fischione and F. Santucci, Modeling and Simulation of Computer Networks and Systems: Methodologies and Applications. Cambridge University Press, 2016.

[16] Technology of complex support of the life cycle of semantically Compatible intelligent computer systems of the new generation / edited by A. A. Kuznetsov. ed. B. V. Golenkov – Minsk : Bestprint, 2023. – 1064 p. – ISBN 978-985-7267-25-5.

МОДЕЛИРОВАНИЯ СОСТОЯНИЯ СИСТЕМЫ ИНФОРМАЦИОННОЙ БЕЗОПАСНОСТИ УМНОГО КАМПУСА

Соболь А. М., Кочин В. П.

В статье рассматриваются компоненты умного кампуса и её многослойная архитектура. Описаны математические модели состояния системы информационной безопасности умного кампуса. Предложены программно-аппаратные средства для моделирования информационных потоков в умном кампусе, для проведения практических экспериментов.

Received 01.04.2023

Cyber-Physical System Improve by AI Models in Causal Approach for Energy Supply (on Transport) Systems

Alexey A. Bezrodniy*, Huang RuiQi*, Van Tszin†
*Belarusian State University, Minsk, Belarus
†Gubkin Russian State University of Oil and Gas, Moscow, Russian Federation

***Abstract*—With the development of control materials and information technologies, cyber-physical systems have become a pivotal area for the description and improvement of large-scale interdisciplinary objects. These systems represent the integration of computing, networking, and physical processes, where embedded computers and networks monitor and control the physical processes, often with feedback loops where physical processes affect computations and vice versa. In today's society, as technology continues to advance, the challenges we face also become increasingly complex. Understanding and improving complex systems, such as energy supply and traffic management, require not only technological support but also a deep understanding of the "semantics" of these systems-namely, the meanings and relationships between components of the system. This paper aims to explore how semantic models built on cause-and-effect relationships can help us better understand and optimize complex systems, and how these models can enhance our comprehension and improvement of such systems.**

***Keywords*—models, algorithms, cyber-physical large-scale inter-disciplinary systems, artificial intelligence, cause-and-effect relations, semantics of solutions**

I. Introduction

By definitions — ancient greek κυβερνητικη as a skill of a ship driver, φυσισ as Nature and συστημαas whole, combined of parts — cyber-physical-systems (CPS) assume efficient control by information technologies considering physical properties of controllables. It is seen as a goal stage of complex system models (after socio-economic, human-machine) at the integration of computing into objects. CPSs are now used to deal with diversative source Big Data problem solutions by decomposition till parts sufficient for decision-making processes - objects automation, 3D-printing of unique items, artificial intelligence (AI), etc.

In AI models, analyzing structure and composition of systems to synthesize their best variants (AI Up-to-Down) and processing big data with smaller attention to the object's nature and generation of new regularities (AI Down-to-Up) are identified. It is purposeful to join said approaches, extending known petroleum supply (PS) aspects for energy supply on transport (EST) [1].

The concept of semantics, or $\sigma\eta\mu\alpha\nu\tau\iota\acute{\kappa}\varsigma$ (meaningful), in the realm of cyber-physical systems, extends far beyond the mere categorization of data. It encompasses the interpretation and understanding of data within a given context(so — how a systems work with them), allowing for a more nuanced and effective control over complex systems. The semantics of solutions, therefore, becomes crucial in achieving precise and anticipatory control mechanisms. By leveraging the semantics within cyber-physical systems, it's possible to bridge the gap between raw data and actionable insights, fostering systems that are not only reactive but also predictive (pro-active) and adaptive to their environments. This semantic layer enriches the decision-making processes, ensuring that the actions are taken at deeper understanding of the system's state and its interaction with the physical world.

By incorporating a semantic understanding into the core of CPS, we align closer with the ancient Greek notion of κυβερνητικη , navigating the vast seas of digital and physical data with the skill and insight of a seasoned ship's captain. It is within this framework that our investigation into energy supply on transport, underpinned by a rich semantic foundation, seeks to unveil new paradigms for control, efficiency, and innovation in the age of information and interconnectedness.

II. General task to control and develop complex systems

A system may be described by pair $\{K, D\}$, where $D = \{Gr, G\}$, Gr — graphs of controlling and control systems (CS), K — quality indicator, G — restrictions.

Transition from verbal to formal description of large-scale inter-disciplinary objects, interesting for investigation, is purposeful to do by causal models [2]–[4].

General cause-and-effect (CE-) relation to develop CPSs is to achieve K by control as a structure formation (on strategical periods) and influences selection:

$$\begin{array}{ccc} \text{Cause: } K & & \text{Effect: } K* \\ > & A(H,P,X,U,F) & < \\ \text{Condition 1: } G, w, S & & \text{Condition 2: } G*, w*, \Gamma(X,U) \end{array}$$

Figure 1: General cause-and-effect relation to develop complex system

On Fig. 1 there is the following formalism (all sets except indexes):

- , A and F — functions, algorithms and control tasks on periods H;
- X — control means, U — relations between them, S — lifecycle stages;
- W — recourses (energy, control, material, techniques, finance, staff);
- G — also demands, influences of friendly and competitive environment;
- *, B — components after interaction (before — or without indexes).

In the general formulation a task usually is difficult to resolve due to high dimension, variety and diversity of components, non-linearity of their interaction, that provides necessity to create new and actualized known models.

III. Classical and cause-and-effect control model

CS models are determined by Cartesian product $F\colon C \times P \times H$, that elements match to number of organizational and technical control means $X(X_{pq}, p = 1..P, q = 1..Q)$ to resolve tasks $F_{ijk}\colon C_i \times P_j \times H_k$ in control loops (circuits) by execution of control functions $C(C_i, i = 1..I)$ for processes $P(P_j, j = 1..J)$ on $H(h_k, k = 1..K)$.

For comprehensiveness there should be build [1] inactive, control, decision-making, information, organization-technical subsystems structures resrectively $(G_r, G_{r1..4})$.

A task is resolved by formation of distributed system model, convolution on functions, periods, and processes (Ω-, C-, H-, P-synthesis), cutting useless (parasite) variants. For CPS it is considered control-by-control $F'\colon C_{cp} \times C \times H/G_{cp}$ with choosing C from $C^* = C \cup C_{cp}$, that provides P_j on H_k best by K.

To further refine this model, integrating semantic analysis into the classical and cause-and-effect control models introduces an advanced layer of understanding that transcends traditional control mechanisms. By incorporating semantics, the models gain the capacity to interpret complex interactions and causalities within the system, not just based on predefined rules but also on the contextual (at given conditions) meaning of actions and their impacts. This semantic enrichment allows for a dynamic adaptation of control strategies that are sensitive to the evolving state of the system and its environment.

Links between the model components are taken from theoretically proved, tested and/or intuitive CE-relations. They are structured on elementary cells (for operations), complexes (for processes) and Universum coverages (whole system).

Dimension of a task is decreased by an incident matrix like in table I.

On Fig. 2 for a 2-level CPS (object, net) in normal and abnormal modes there are relations in a CS, where Gr_2 of decision acts A_{ij} determines $Gr_{1,4}$ and Gr, Gr_3.

For EST-systems it is assumed, that within set of P_j ($j = 1..5$, 1 – supply and logistics, 2 – storage and

Table I: Matching of classical and cause-and-effect control model

Multi-loop model component	Cause-and-effect model component			
	G	C	W	S
C	C~G	Identity	C~F_{CW}(W)	C~ F_{CG}(G)
P	P~ F_{PG} (G)	P~F_{PC}(C)	P~F_{PW}(W)	P~ ∑S
H	H~F_{HG}(G)	H~C	H~W	H~S
X	X~ F_{XC} (G)	X~ C	X~ W	X~F_{XS} (W)
G	Identity	G depends on C, W, S weaker («control by objectives »)		
W	W~G	W~C	Identity	W~ F_{WS} (G)
N	N ~ G	N ~ C	N ~ W	N ~ S
S	S ~ G	S ~ C	S ~ W	Identity

operation, 3 – sales and marketing, 4 – support, service and maintenance, 5 – accounting/reporting) P_3 dominates (marketing paradigm, purpose-oriented approach) at typical set of C_i($i = 1..5$, 1 – data collection, 2 – situation identification, 3 – decision-making, 4 – execution, 5 – coordination).

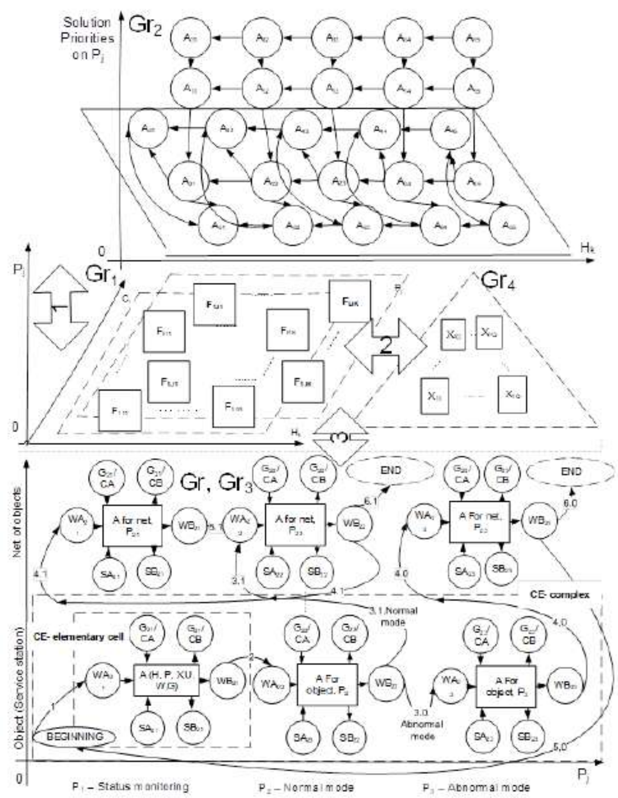


Figure 2: System structure at using of both classical and causal models

The synergy between classical control models and semantic analysis opens new avenues for developing intelligent cyber-physical systems. This combined approach not only enhances the precision and flexibility of control strategies but also ensures that systems are capable of understanding and reacting to their operational context in a more nuanced and effective manner. Consequently, it lays the groundwork for creating systems that are not only automated but also truly intelligent and responsive to the complex demands of modern infrastructure and societal needs.

At models co-using it is provided matching between

- elementary control tasks and cells of CE-relations F_{ijkpq} and $RC_{\text{phph}\to\text{phph}+1}$

$$F_{ijkpq} : C_i \times P_j \times H_k : X_{pq} \equiv \\ \equiv RC_{phph\to phph+1}(C, G, S, W), \quad (1)$$

- set of processes and CE-complexes

$$P_j \colon S_{ph1} \xrightarrow{F_{ijkpq}} S_{ph2} \cdots \xrightarrow{F_{ijkpq}} S_{phPH} \equiv \\ \equiv (RC_{A\to B}, U_{RC}), |RC|, |U_{RC}| \quad (2)$$

  $RC_{A\to B}$-set of elementary CE-cells, transforming system from A to B, URC – set of relations,$|RC|$ and $|U_{RC}|$ – matrixes of incidents and adjacency.

- structure models of sub-systems Gr Gr_{1-4}(graphs) and CE-components

$$P_j \colon S_{ph1} \xrightarrow{F_{ijkpq}} S_{ph2} \cdots \xrightarrow{F_{ijkpq}} S_{phPH} \equiv \\ \equiv (RC_{A\to B}, U_{RC}), |RC|, |U_{RC}| \quad (3)$$

  Order of resources w processing and status S changes as algorithmical solution finally determined by $Gr_5(X_5,\, U_5)$- coverage of Universum by links:

$$Gr_5(X_5, U_5) \colon \{(RC_{A\to B}, U_{RC}), |RC|, \\ |U_{RC}|, OA\}, j = 1..J, \quad (4)$$

  where $OA_{oa} \in \{OA_{oa}\}$ (oa=1..OA, 1 – beginning, 2 – operator, 3 – conditional pass, 4 – input-output). It is seen that number of models is decreased (Fig. 3).

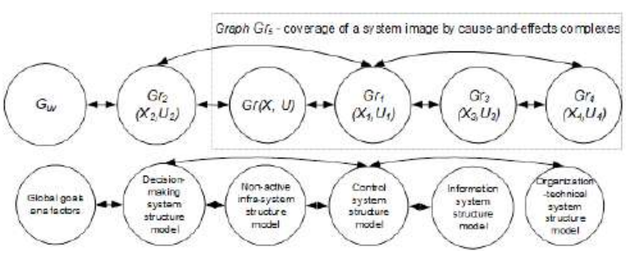


Figure 3: Dimension decrease of control (and system) model at algorithmical solution way

IV. Neural nets for causal system models

Artificial neural networks (NN) imitate human head mind structure as majority of interconnected elementary cells (neurons), executing many relatively simple operations in coordinated way. Information processing rules are given by LPR (leader of problems resolving), named also as decision-makers and experts.

At many-time using a NN is trained (control with information accumulation and analysis principle) with new regularities (knowledge) generation.

General structure of NN in modern understanding (e.g. of recurrent net with feedbacks and blocks of long short-time memory and convolution) may presented, accordingly to the opinion of authors as on Fig. 4.

Comparison and analysis of models, briefly presented on Fig. 2 and 4 shows about similarity of their structure and composition:

- elementary CE-cells (compatible NN nodes) with decomposition level, sufficient for LPRs for efficient control (layers and blocks) and coverage of a system Universum (whole NN);
- correction at new reports (data) coming, algorithmical solution (NN training);
- interrelated set of known models of subject field in Data (Knowledge) Base or D(K)B, forming the most general cause-and-effect relation (may be called as models of etalon processes), to which in NN match recurrent (RNN, modelling feedback), convolutional (CNN, dimension decrease with unproportionally lower quality lose), Long Short-time memory (LSTM, selection of alternative variants in conditions), Gated recurrent unit (GRU, variant of LSTM), etc.

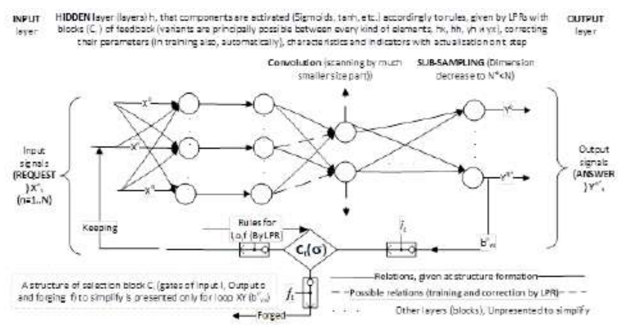


Figure 4: Structure model of a neural network with inter-element interaction

It permits to use developed mathematical support — software (and, sometimes, hardware for NN to model of operation and evolutional elaboration of CPSs by system cause-and-effect approach, that is good enough to investigate, control and improve of complex large-scale inter-disciplinary systems [2]–[4].

V. Task and methods to improve energy supply on transport

As a subject field one considers systems to supply energy on transport, on which development of State, Society and Man really depends one.

Most general cause-and-effect relations constituening semantics (σημαντιϰς,determening) are presented as follows:

- spatial distribution of the systems of high dimension and complexity;

- huge data arrays (sets) of diversative data (clients, suppliers, macroeconomics);
- alternative and renewable energy sources (ARES), requiring efficient control systems at restrictions G_{ARES} on street-and-road (S-n-R) nets;
- vehicle transportation car diversity (TC) with exponential growth of numbers;
- hierarchical structure model — filling (charging) station, net, company – at net object location on Street-and-Road (S-n-Rs) and program (goal) oriented CS structure models;
- miniaturization of controlling and controllable components (narrow case CPS);
- trend of automation control system units (ASU) wide implementation, AI, decision support systems (DSS), Internet-of-Things (IoF), etc.

In the intricate web of cyber-physical systems, the most general cause-and-effect relations not only serve as the foundational skeleton but also weave the intricate tapestry of semantics (why systems exist and what for) that imbue these systems with meaning and purpose. The semantics, derived from these relations, go beyond mere functional descriptions, embedding systems with an understanding of their environment, objectives, and the potential impact of their operations. This semantic layers (following, e. g., M.Mesarovitch stratas and echelons), informed by the multifaceted interactions and outcomes of cause-and-effect dynamics, equips systems with the capability to interpret, adapt, and respond to the complexities of real-world challenges.

By integrating these cause-and-effect relations into the semantic framework, we enable systems to comprehend not just the 'what' and the 'how' of their operations, but also the 'why' behind their actions. This deeper level of understanding is crucial for developing systems that are not only efficient and effective but also responsible and sustainable. It allows for the anticipation of potential consequences, the alignment of system operations with broader goals and values, and the adaptation to unforeseen challenges. Ultimately, the semantics constituted by these cause-and-effect relations transform cyber-physical systems from mere tools into intelligent agents capable of making informed decisions and contributing to the advancement of human society.

Actual tasks of EST-system improvement are presented in Table II.

Table II: Tasks of cyber-physical systems to supply energy on transports, models, methods

Actual task	Semantic description	Models, methods, examples
Improvement of alternative and renewable energy systems	Synthesis of optimal nets to supply ARES on transport	Models of ARES [5] to clarify Methodology of fuel supply net development [1]
Data (Knowledge) Bases Development	Added reality system form to improve human-machine DSS	Creation new/actualization known D(K)Bs with more efficient human-machine dialog
Control quality increase by genetical/evolutional schemes	System structures improve, using biological analogues	Known control system models, modified by evolutional theory causal relations
Service net optimization, using neural nets	Processing of diversative electronic sensorship Big Data	New regularities for service nets, obtained at NN use (CNN, GRU, LSTM, RNN)
Swarm intellect to model transportation flows	Client flow clarification s by adding activity properties	Group behavior models of living systems (ants, bees, etc.) vs. electronic sensorship
Smart objects (stations and terminals)	Extended rational automation use, considering accessibility	IoT, intellectual monitoring (energetics, alarm, spectroscopy, etc.), ASU of support
Robotic objects and sub-systems	Pilot-operation of fully automatic components	Out-door payment terminals, automatic markets, robotically cleaning, etc.

After results, already presented in [2], in the said paradigm, the following advancements were achieved in the last several years, further embedding semantics into the core of cyber-physical systems:

- optimal service net structure for alternative and renewable energy sources (on the example of Hydrogen and compressed natural gas) [5];
  The development of an optimal service net structure, especially for alternative and renewable energy sources such as Hydrogen and compressed natural gas, represents a significant leap in applying semantic models. By understanding the intricate cause-and-effect relationships governing energy distribution and consumption, these systems can optimize service networks in a way that maximizes efficiency and sustainability.
- maintenance of service nets by the control principle <satisfaction to requirements> [6], implemented in petroleum supply companies in some regions of Russian Federation and Belarus;
  The maintenance of service networks based on the principle of satisfying requirements showcases the application of semantics in ensuring systems can interpret and act upon complex criteria. This approach leverages semantic understanding to tailor maintenance activities, ensuring that operational parameters align with specific needs and conditions, enhancing overall service quality and reliability.
- ASU of guarantied petroleum product supply implementation in net of 84 stations and 4 terminal of a petroleum company in Belarus, decreased number of dangerous operations on objects practically to zero and increased operational efficiency on them;
  Implementing an Automated System of Control (ASU) for guaranteed petroleum product supply, which significantly reduced the risk of dangerous operations and enhanced efficiency, exemplifies the role of semantics in risk management and operational planning. By semantically analyzing data

from various sources, the system can predict potential risks and optimize operations to mitigate them effectively.

- co-projection and co-development of stations and terminals net [6], that in an application permitted to decrease quantity of terminals, necessary to serve a regional net, optimized there structure (minimal outages of both equipment and clients) and decreased cost of only construction on 12,5 %

  The co-projection and co-development of station and terminal networks highlight the semantic-driven approach to system design and optimization. Through a deep understanding of the system's goals and requirements, this approach enables the reduction of unnecessary infrastructure, leading to cost savings and increased efficiency without compromising service quality.
- investigation of general information-logical schemes to model energy supply on base of genetic algorithms, forecasting of petroleum product logistics by various kind back propagation NN models, added reality to improve interaction in human-machine decision-support systems, etc.

  The exploration of information-logical schemes and the application of genetic algorithms and neural network models for forecasting in energy supply systems underscore the importance of semantics in modeling and simulation. These semantic-rich models facilitate more accurate predictions, improving decision-making and interaction within human-machine decision-support systems by providing a deeper contextual understanding of operational data and trends.

In particular, only initial modeling by genetic algorithms (GAs) application accordingly to the scheme, presented on Fig. 5, efficiency of simulation and the subject field KPIs are increased on 5-10 % and more. Hence, GAs, with their ability to perform global searches through selection, crossover, and mutation, looking well-suited to identify optimal or near-optimal system configurations that balance performance requirements with operational stability and adaptability.

Furthermore, integrating semantic analysis into this process enhances the capability of GAs by providing a richer contextual understanding of the system components and their interactions. By incorporating semantics, the genetic algorithm can evaluate not only the quantitative performance metrics but also the qualitative aspects of system configurations. This involves analyzing the meaning and relevance of various system parameters and states in relation to the overall objectives, such as energy efficiency, sustainability, and user satisfaction. For instance, a semantic layer could help the GA prioritize configurations that align with sustainable practices or adapt to user preferences, beyond mere numerical optimization.

This enriched modeling approach, where semantics and genetic algorithms work in tandem, enables a more holistic optimization of cyber-physical systems. By understanding the "why" behind system behaviors and configurations, as informed by semantic analysis, and leveraging the "how" of achieving optimal configurations, as facilitated by GAs, we achieve a more robust, adaptable, and intelligent system design. This dual approach not only improves simulation efficiency and KPIs but also ensures that the system's evolution is aligned with strategic goals, user needs, and environmental considerations, thus contributing to a more sustainable and user-centric ecosystem.

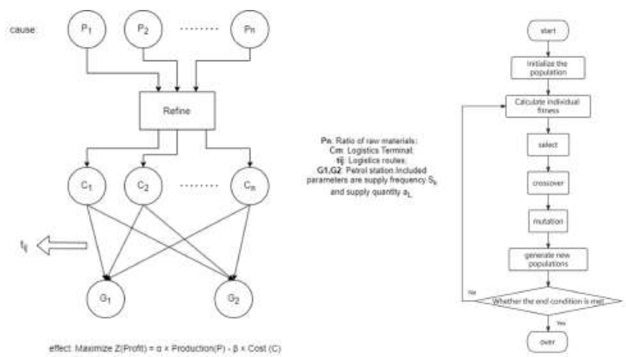


Figure 5: A structure of an elementary petroleum product net for genetical algorithm

So far, interaction of models related to energy supply systems (on transport) may pe presented as on Fig.6, where current tasks are marked by gray and/or bold.

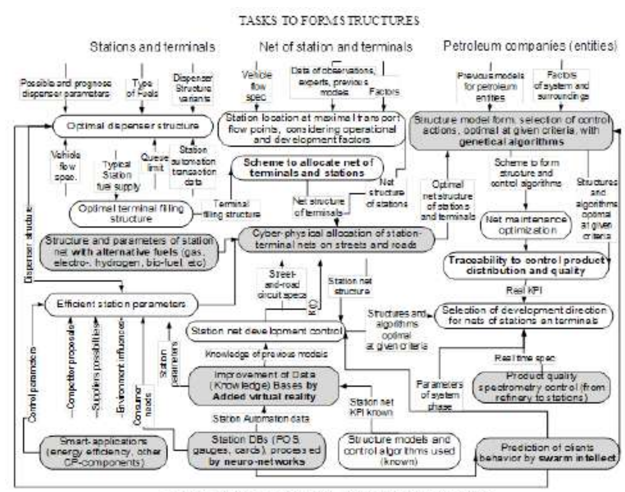


Figure 6: A structure of tasks and models for energy supply systems on transport

VI. Conclusion

1) Cyber-physical systems reflect a modern view on complex objects and their nets, that causes actuality

of their improvement to resolve efficient control tasks for large-scale inter-disciplinary cases.

2) Application of system causal structure models and artificial neural networks together permit to increase quality of investigations, control, and development of cyber-physical systems, due-to development of software packages for nerual networks and computer power.
3) Since the causal approach has a universal application, it may permit to investigated and improve systems of, practically. any kind, taking into consideration their inner languages, rather indefinite but important interactions between elements and verbal description of they characteristic, and control parameters, that constitutes semantics of object (events, phenomena and processes) images for their further modelling .formal description and mathematical modeling.
4) For systems to supply energy on transport as example, above models are being developed, tested or piloted, with such features as co-projection of controlling and controllable parts, wide cyber-physical component implementation, artificial intelligence using both in Up-to-Down and Down-to-Up variants and continued clarification at new reports coming, that is also applicable in other types of nets.
5) At synthesis of efficient control structures within so-called Principal component analysis dimension reduction by clustering/classification and summarizing there may be also applied some other methods of optimization tasks theory and practice – parcel solution trees, emergency combination of events, etc. as, probably future technique of AI, investigations to be continued.
6) Semantic analysis enabled us to identify hidden causal relationships and behavioral patterns within the system, offering a deep-level method of understanding systems. This understanding goes beyond mere data analysis to encompass grasping the meanings behind system behaviors, providing a new perspective for system design and optimization.
7) Utilizing semantic information, we demonstrated how to optimize decision-making processes in complex and dynamically changing environments. Mining the semantic layer allows the system to anticipate and adapt to future changes, achieving intelligent and adaptive decision-making, essential for building systems that are effective over the long term and highly adaptable.

References

[1] Models and ways to design the structure of multi-loop management information systems / Akinfiev V.K., Kostuk V.P., Rhezchikov A.F., Tsvirkun A.D. Automation and remote control. 1988. T. 49. № 3. C. 348.

[2] Alexey Bezrodny, Anatoliy Korolenok, Aleksandr Rezchikov. Complex human-machine petroleum system improves by causation approach //Open Semantic Technologies for Intelligent Systems. Research Papers Collection. Issue 5. Minsk, Belarus, 2021. Pp. 251-256.

[3] Mathematical models, algorithms and software package for the national security state of Russia. Yandybaeva N. and others. Studies in Systems, Decision and Control. 2019. T. 199. C. 646-659.

[4] The problem of preventing the development of critical combinations of events in large-scale systems. Bogomolov A.S. and others. Advances in intelligent systems and computing, vol. 1224, 2020, pp.274-280.

[5] Alexey Bezrodniy, Alexander Rezchikov, Oleg Dranko. Algorithms to De-velop H2 Filling Station Nets/ IEEE 10.1109/ MLSD52249.2021.9600214.

[6] Generalized algorithms for joint construction and improvement of oil deposits and filling stations systems/ Van Tsin, A.A. Bezrodniy. Proceedings of Gubkin Russian State University. 2022. № 1 (306). p. 60-74.

[7] Oliveira, Paulo, João Sequeira, and Joao Sentieiro. "Selection of controller parameters using genetic algorithms." Engineering Systems with Intelligence: Concepts, Tools and Applications (1991): 431-438.

[8] Fleming, P.J., & Purshouse, R.C. (2002). Evolutionary algorithms in control systems engineering: a survey. Control Engineering Practice, 10(11), 1223-1241.

[9] Bermúdez-Edo, María et al. “IoT-Lite: a lightweight semantic model for the internet of things and its use with dynamic semantics.” Personal and Ubiquitous Computing 21 (2016): 475–487pp.

[10] Goldberg, David E. Genetic Algorithms in Search, Optimization, and Machine Learning. Addison-Wesley Longman Publishing Co., Inc., 1989.

РАЗВИТИЕ КИБЕРФИЗИЧЕСКИХ СИСТЕМ С ПОМОЩЬЮ МОДЕЛЕЙ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА В РАМКАХ ПРИЧИННО-СЛЕДСТВЕННОГО ПОДХОДА НА ПРИМЕРЕ ОБЪЕКТОВ ОБЕСПЕЧЕНИЯ ЭНЕРГИИ (НА ТРАНСПОРТЕ)

Безродный А.А., Жуйци Х., Цзинь В.

С развитием управления, теории материалов и информационных технологий кибер-физические системы можно считать целевой парадигмой для описания и совершенствования сложных междисциплинарных объектов. Данного рода системы подразумевают интеграцию вычислительных мощностей управления с обратной связью непосредственно в сетевые распределённые структуры. В современном мире бурно развивающихся технологий их создание с требуемым качеством на примере обеспечения энергией на транспорте, требует не только надлежащего технического обеспечения, но и глубокого понимания смысловой составляющий – предназначения, структур и состава компонент и их взаимосвязей. Работа имеет цель показать как семантические модели на основе причинно-следственных связей могут помочь при оптимизации сложных систем, обеспечивая их комплексное представление и дальнейшее развитие.

Received 14.03.2024

Methods for Defending Deep Neural Networks Against Adversarial Attacks

Vitali Himbitski
Belarusian State University
Minsk, Belarus
gimbitskyvitaly@gmail.com

Angelina Varashevich
Belarusian State University
Minsk, Belarus
varvashevichangelina@gmail.com

Vasili Kovalev

Minsk, Belarus

***Abstract*—In this paper, Gradient-based adversarial attack methods on classification neural networks for image processing are discussed. And also the archeitecture of neural network for defense against adversarial attacks is proposed. For the proposed neural network architecture, the dependence of the preprocessing quality of the attacked images on the training time and the number of trained parameters has been studied.**

***Keywords*—Deep Neural Networks (DNN), Adversarial Attack, Autoencoder, Convolutional Neural Network**

I. Introduction

In modern society, deep neural networks (DNNs) are widely used in various fields such as voice assistants, smart homes, security systems, and medical equipment. They face significant challenges, such as performing financial transactions upon voice assistant requests, autonomous vehicles participating in traffic, and medical equipment detecting diseases based on analysis and images. However, what happens if these systems fail to perform their intended tasks correctly? This could occur due to insufficient system testing or intentional interference. For example, based on data analysis, it may be recommended for almost everyone to visit a doctor, even if they are healthy. This would lead to overload in medical facilities and a decrease in attention from healthcare professionals.

Since DNNs can be easily attacked, it is important to consider the possibilities of preventing attacks on neural networks, especially in situations where the consequences of distorted operations can be extremely serious, such as informing a patient that they are healthy when they are not. Analytical work has been conducted to create protective mechanisms based on the fundamental principles of attacks and defense. The article "Intriguing Properties of Neural Networks" [1] proposed a process for creating adversarial input data in the field of image processing, known as the L-BFGS method. It also hypothesized the instability of DNNs to adversarial images due to their nonlinearity.

Later, the article "Explaining and Harnessing Adversarial Examples" [2] was published, which extensively discusses the problem of adversarial images and their impact on DNNs and linear models. The research concluded that the instability of DNNs is specifically due to the linear nature of their components. The article introduced the FSGM attack, which is computationally lighter compared to the previously proposed L-BFGS method. The influence of this attack on various networks, including linear models trained on the MNIST dataset for classification, was described. An alternative approach to data augmentation and improving resilience to adversarial images was also proposed — training the network on generated adversarial images.

The book "Reliability of Neural Networks: Strengthening AI's Resistance to Deception" [3] provides information about various attacks on artificial intelligence, describes the conceptual foundations of generating adversarial input data, and explains the computational methods used to create them. The book also presents arguments about protecting DNNs from attacks.

The article "Adversarial Attacks on Reinforcement Learning" [4] presents different attacks on DNNs as well as various defense methods against such attacks.

The structure of this article is as follows. Section 2 describes the datasets used in this research. Then, in Section 3, the method for generating adversarial attacks is described. Section 4 briefly outlines the adversarial training method. Finally, Section 5 describes the defense methods against adversarial attacks proposed in this work and analyzes the results of protecting DNNs using these methods.

II. Training Data

In this paper, adversarial attacks were performed on classification neural networks for image processing. The following input datasets were used for training:

- CT slices of people with corresponding classes: shoulders, heart, liver. The size of the dataset is 1801 and the classes are balanced. The images are proposed in RGB (red, green, blue) format with size 512*512.
- Histologic breast cancer image set composed of images from 2 different hospitals in the Netherlands. The set was used in the CAMELYON-2016 competition, and includes the corresponding classes: normal

(NRM), tumor (TUM), epithelial tissue (EPI). The dataset size is 18000 and the classes are balanced. The images are proposed in RGB (red, green, blue) format with a size of 256*256.

Figures 1 and 2 are an example of the contents of CT slice set and histology set, respectively.

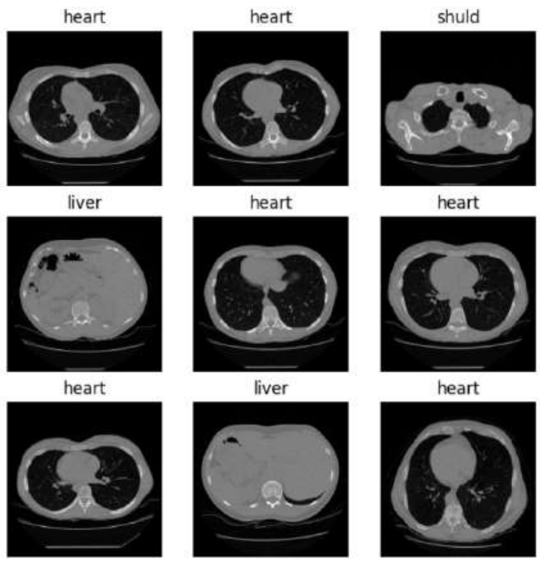


Figure 1. Example of images from the CT scan dataset.

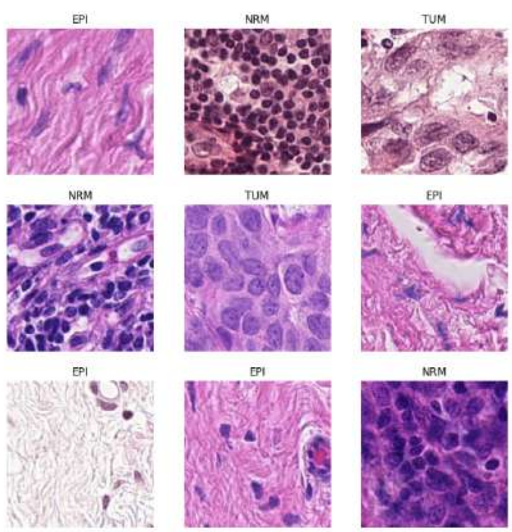


Figure 2. Example of images from a histology set.

III. Adversarial attaks

Let's consider an algorithm that classifies images. Let its processing be defined by the function:

$$f : X \to y \quad (1)$$

Here, XX is the set of images with a size equal to the input size of the classification algorithm, and yy is the set of numbers corresponding to class labels. Now, suppose there exists an attacking algorithm whose processing is defined by the function:

$$g : X \to X' \quad (2)$$

The sets of images XX and X'X' have different distributions, but the images in these sets have the same size.

If the attacking algorithm works effectively, then the following condition holds:

$$f(x) \neq f(g(x)), \quad x \in X \quad (3)$$

under the condition that:

$$\|x - g(x)\| < \epsilon \quad (4)$$

A. Gradient-based adversarial attacks

Gradient-based adversarial attacks are used to misclassify an image by a neural network. They can be directional and non-directional. Gradient-based adversarial attacks work as follows. The attacking algorithm generates noise from some distribution. The noise should have the same dimensions as the image itself. This noise is added to the image forming a new transformed image. The transformed image is fed to the input of the classification network. If the adversarial attack is successful, the neural network misclassifies the transformed image. Next, we will describe some of the most common methods of gradient attacks. They will also be used later in this paper.

B. Fast Gradient Sign Method (FGSM)

The Fast Gradient Sign Method (FGSM) **??** is an adversarial attack method that utilizes the gradients of a model to create adversarial examples. It was proposed in 2014 by researchers Ian Goodfellow, Jonathon Shlens, and Christian Szegedy.

The FGSM method defines the function (4) as:

$$g(x) = x + \epsilon \cdot \text{sign}(\nabla_x l(x, y_{\text{true}})) \quad (5)$$

Here, ϵ is the magnitude, and $l(x, y_{true})$ is the loss function of the classification neural network with respect to the true label.

The main idea of FGSM is to utilize the gradient information of the model's loss function with respect to the input data to create adversarial examples. Gradients indicate the direction in which the loss function will change the most when the input data is perturbed. The FGSM attacking algorithm utilizes this information to create adversarial examples that maximize the loss function, resulting in the misclassification of the image by the neural network.

The FGSM method, as presented in the implementation (5), can be used for untargeted attacks. Since this implementation increases the loss function with respect to the true label, the classification neural network is more likely to predict the wrong class.

By using a different implementation of the attacking function in FGSM, a targeted attack can be achieved. To do this, it is sufficient to define the attacking function implementation as:

$$g(x) = x - \epsilon \cdot \text{sign}(\nabla_x l(x, y_{\text{target}})) \quad (6)$$

Here, ϵ is the magnitude, and $l(x, y_{target})$ is the loss function of the classification neural network with respect to the target label. This implementation minimizes the loss function of the classification neural network with respect to the target label, resulting in the neural network more likely to predict the target class for the given image.

C. Iterative Fast Gradient Sign Method (I-FGSM)

The Iterative Fast Gradient Sign Method (I-FGSM) is an adversarial attack method that is implemented by repeating the FGSM method a certain number of times. In I-FGSM, the attacking algorithm is implemented using the following functions:

$$g(x) = x_n \quad (7)$$

$$x_{t+1} = \text{clip}_\epsilon(x_t + \alpha \cdot \text{sign}(\nabla_x l(x_t, y_{\text{true}}))) \quad (8)$$

Here, $x_0 = x$, α is the step size for adapting x_t, $clip$ is a function that ensures $x_t \in (x-\epsilon, x+\epsilon)$, and $l(x_t, y_{true})$ is the loss function of the classification neural network with respect to the true label.

In the above implementation, the I-FGSM algorithm generates an untargeted adversarial attack. It increases the loss function of the classification neural network with respect to the true label, causing the neural network to more likely predict the wrong class for the given image.

This method can be modified to implement a targeted adversarial attack. For this, the attacking algorithm should be defined using the following function:

$$g(x) = x_n \quad (9)$$

$$x_{t+1} = \text{clip}_\epsilon(x_t - \alpha \cdot \text{sign}(\nabla_x l(x_t, y_{\text{target}}))) \quad (10)$$

Here, $l(x_t, y_{target})$ is the loss function of the classification neural network with respect to the target label. This implementation minimizes the loss function of the classification neural network with respect to the target label, resulting in the neural network more likely to predict the target class for the given image.

The I-FGSM attack method is stronger than FGSM. In experiments conducted in this study, it has been found that for the same magnitude of attacks, the I-FGSM method leads to a higher number of incorrect predictions by the considered neural networks compared to the FGSM method.

D. Generation of adversarial attacks

The experiment investigated the effect of FGSM on ResNet50, MobileNetV2 and DenseNet169 neural networks, which are widely used image classification models. Different attack scenarios were studied, including changing the input data, changing the noise magnitude ϵ, and applying defense techniques such as Adversarial Training and using an additional neural network for image preprocessing.

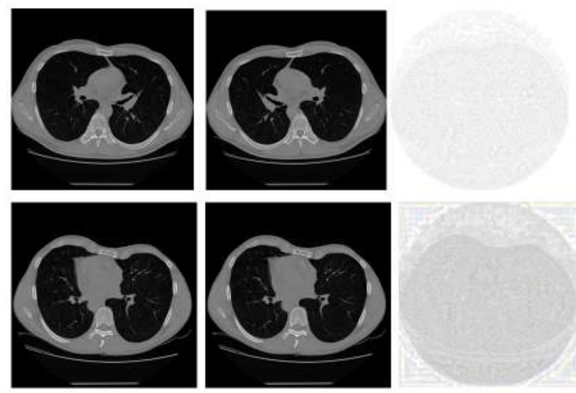


Figure 3. Example of adversarial attack when $\epsilon = 0.004$, (first and second rows, respectively). From left to right in the row — the original image, the malicious image, and the attack itself.

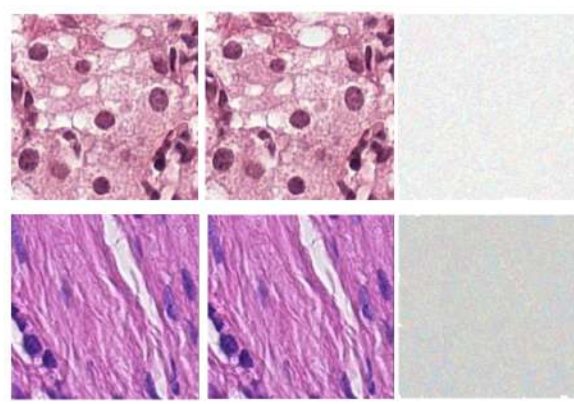


Figure 4. Example of adversarial attack when $\epsilon = 0.004$ (first and second rows, respectively). From left to right in the row — the original image, the malicious image, and the attack itself.

Figures 3, 4, 5 and 6 show examples of the original images, malicious images with added noise from the FGSM attack, and the attack itself. The values of each pixel have been magnified by a factor of 30 to visualize the differences between the images. In each figure, the first row shows the best classification quality result obtained at $\epsilon = 0.004$, while the second row shows the best result at $\epsilon = 0.01$. From left to right, each figure

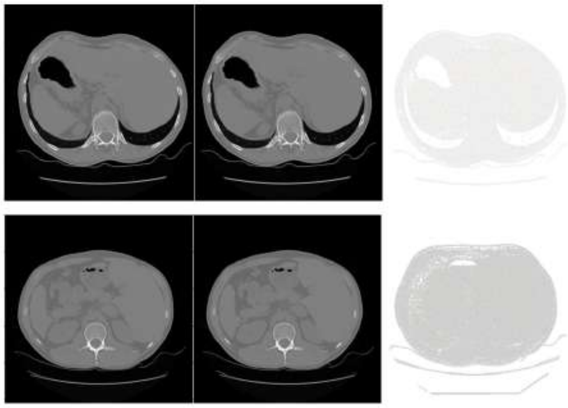


Figure 5. Example of adversarial attack when $\epsilon = 0.01$, (first and second rows, respectively). From left to right in the row — the original image, the malicious image, and the attack itself.

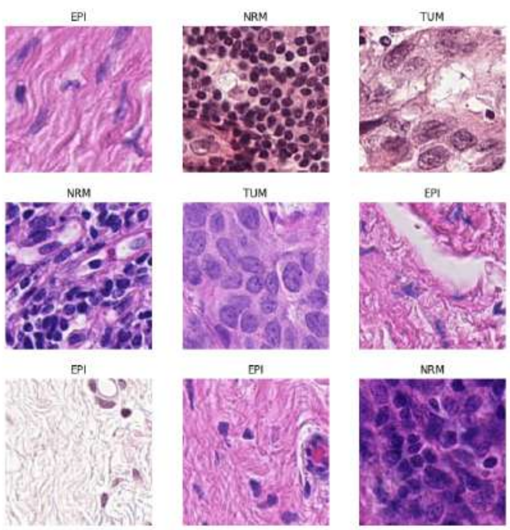


Figure 6. Example of adversarial attack when $\epsilon = 0.01$, (first and second rows, respectively). From left to right in the row — the original image, the malicious image, and the attack itself.

shows the original image, the malicious image, and the attack corresponding to that image.

E. Application of semantic technologies for image classification

Semantic approaches can also be used for the task of classifying attacked images. One such approach is to use semantic segmentation, which divides the image into semantic regions or objects. This allows the model to more accurately analyze the content of the image and classify it based on the different segments. Semantic segmentation can be achieved through a variety of techniques, including using convolutional neural networks with architectures specialized for segmentation, or combining deep learning with image processing techniques.

In addition, semantic techniques can be used to improve the interpretability of image classification. This is especially important in tasks where explanation and understanding of the decisions made are critical. Semantic technologies allow models to explain their classification decisions based on high-level concepts and semantic attributes. For example, one can use activation visualization techniques in neural networks to identify important image regions that influence classification, or to show relationships between classes and semantic concepts.

IV. Adversarial training

Adversarial training defense is as follows: the input dataset that learns to classify a DNN is augmented because the training dataset will now include adversarial images, a kind of augmentation. This form of data augmentation utilizes input data that is unlikely to be naturally acquired, but which reveals the shortcomings of the DNN. With the generated dataset, the model tries to minimize the loss function. In this way the DNN becomes less sensitive to small distortions. However, if malicious images produced by a different model but trained on a subset of the original input data, i. e., the model approximates the target DNN, are given as input, their behavior will be similar. That is, incorrect results may be obtained again, more or less different from the results obtained on the malicious images generated by the original network (from [2]).

In order to improve the robustness of the network using Adversarial training, the target DNN can be trained not only on malicious images generated by the original model, but also on other models. Since the model weights will be adjusted to the combination of malicious image groups, the DNN will become more robust to malicious images and it will be more difficult for an attacker to attack it.

V. Defense against adversarial attacks

In the Adversarial training procedure of the classification neural networks considered in this paper, we used the $SparseCategoricalCrossentropy$ loss function, $Adam$ optimizer with a training step of 10^{-3}. Since the classes in the input datasets are balanced, to assess the quality of the model we will use the accuracy quality metric, which characterizes the proportion of correct predictions relative to the number of all objects.

In this paper, a neural network with autoencoder architecture was used to implement the method of preprocessing attacked images. For the proposed neural network, the classification quality of the images preprocessed using it was measured.

The disadvantage of the autoencoder architecture is the significant reduction in the amount of information in the latent space relative to the amount of information of the image itself. To solve this problem, layers with information transfer from the previous state were added to the neural network architecture to protect against adversarial attacks. The architecture of the proposed neural network is described in Table 3.

The neural network for image preprocessing was trained using a composite loss function consisting of image restoration loss function and classification loss function.

Also during the research of this image preprocessing method, the dependence of the classification quality of preprocessed images on the number of training epochs of neural networks for defense against adversarial attacks was measured.

The results of the conducted experiments are presented in Tables II and III.

From the results presented in the tables in Tables II and III, the following conclusions can be drawn:

1) The implementation of the attack depends on the choice of the dataset on which the target deep neural network (DNN) has been trained. When the same DNN is used to classify histology and computed tomography (CT) related datasets, the results after the FSGM attack do not match the accuracy of the original classification proportionally.
2) Attacks on different neural networks show different results. For example, the MobilenNetV2 neural network is found to be more resistant to attacks. This property is also manifested at different levels of attack magnitude. The results obtained depend primarily on the number of trained parameters in the neural network. The neural network with less number of trained parameters is less susceptible to attacks.
3) When the considered defense methods are applied on malicious images generated on the same networks, the classification accuracy on both datasets is recovered, and this recovery depends on the number of trained parameters. Although the MobileNetV2 neural network proved to be more resistant to attacks, its accuracy is recovered slower compared to the Resnet50 and Dencenet169 neural networks.
4) The amount of noise added to the original data to create malicious images affects the classification accuracy. The more noise added, the greater the deviation of the prediction vector from the original vector. Hence, the amount of noise required to create a malicious image depends on the particular network and the input dataset used.
5) The considered neural networks on a set of CT images showed high classification accuracy, indicating that the added noise is not enough to create malicious images.
6) When more noise is added to the original images at $\epsilon = 0.01$, the attack significantly reduces the accuracy of the classifier on the malicious image sets, except in the case of the CT image set for MobileNetV2, where the added noise at this value of ϵ was insufficient to turn the images into malicious images.

Thus, the classification accuracy of a DNN after applying FGSM to it is affected by the original input dataset, the architecture of the DNN, and the amount of noise superimposed on the images. The considered methods of protection against adversarial attacks showed high quality on all pairs of network-dataset type. This indicates the effectiveness of this type of protection to ensure the network resistance to FSGM attack.

VI. Conclusion

In this paper, a comprehensive analysis of gradient-based attacks is performed, taking into account the influence of the Deep Neural Network (DNN) architecture, the input dataset used, and the amount of added noise ϵ to form malicious images. Adversarial training and image preprocessing using neural networks were applied as defense techniques against the FGSM attack. These approaches achieved high classification accuracy on the original FGSMs, but did not fully recover the original accuracy level.

The results emphasize the importance of security techniques in deep neural networks. The study of attacks on such networks is of great importance for the development of effective defense mechanisms. The considered problematic requires further in-depth research, especially in light of the wide application of deep neural networks in various fields, including medicine, where classification accuracy is of particular importance and the potential consequences of errors are unacceptably high.

References

[1] Ch. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, R. Fergus, “Intriguing properties of neural networks,” arXiv preprint arXiv:1312.6199, 2013.

[2] I. Goodfellow, J. Shlens, Ch. Szegedy, Explaining and Harnessing Adversarial Examples, InternationalConference on Learning Representations, ICLR 2015, San Diego, CA, USA, 2015.

[3] Warr K, “Reliability of neural networks: strengthening AI robustness to deception,” SPb: Peter, 2021, pp. 272.

[4] Weilin Xu, David Evans, Yanjun Qi, “Feature Squeezing: Detecting Adversarial Examples in Deep Neural Networks,” arXiv preprint arXiv:1704.01155v2, 2017.

[5] Seungju Cho, Tae Joon Jun, Byungsoo Oh, Daeyoung Kim, “DAPAS : Denoising Autoencoder to Prevent Adversarial attack in Semantic Segmentation,” arXiv preprint arXiv:1908.05195v4, 2020.

[6] Baoyuan Wu, Shaokui Wei, Mingli Zhu, Meixi Zheng, Zihao Zhu, Mingda Zhang, Hongrui Chen, Danni Yuan, Li Liu, Qingshan Liu, “Defenses in Adversarial Machine Learning: A Survey,” arXiv preprint arXiv:2312.08890v1, 2023.

[7] Boxi Wu, Heng Pan, Li Shen, Jindong Gu, Shuai Zhao, Zhifeng Li, Deng Cai, Xiaofei He, Wei Liu, “Attacking Adversarial Attacks as A Defense,” arXiv preprint arXiv:2106.04938v1, 2021.

[8] Ergezer M., Duong Ph., Green Ch., Nguyen T., Zeybey A., “A One Noise to Rule Them All: Multi-View Adversarial Attacks with Universal Perturbation,” arXiv preprint arXiv:2404.02287, 2024.

Table I
Results of the experiments with $\epsilon = 0.004$

Data $\varepsilon = 0.004$	No attack	FSGM	Adversarial training	Preprocessing data	Pretrained
Histology	90.11	71.78	77.6	83.5	no
CT	100	100	100	100	no
Histology	94.65	60.73	91	93.6	yes
CT	100	86.12	100	100	yes

Table II
Results of the experiments with $\epsilon = 0.01$

Data $\varepsilon = 0.004$	No attack	FSGM	Adversarial training	Preprocessing data	Pretrained
Histology	90.11	36	83	85.5	no
CT	100	90	99.2	100	no
Histology	94.65	9.98	94.1	94.5	yes
CT	100	5.59	99.6	100	yes

Table III
Neural network architecture for image preprocessing

Layers	Input Shape → Output Shape	Layers Information
Input Layer	(1, 224, 224) → (32, 112, 112)	Conv2d
Encoder	(32, 112, 112) → (64, 56, 56) (64, 56, 56) → (128, 28, 28) (128, 28, 28) → (256, 14, 14)	Conv2d, LeakyReLU Conv2d, LeakyReLU Conv2d, LeakyReLU
Hidden Layer	(256, 14, 14) → (256, 14, 14)	Reshape, Linear, Reshape, Skip Connection
Decoder	(256, 14, 14) → (128, 28, 28) (128, 28, 28) → (64, 56, 56) (64, 56, 56) → (32, 112, 112)	Conv2dTransform, LeakyReLU, Skip Connection Conv2dTransform, LeakyReLU, Skip Connection Conv2dTransform, LeakyReLU, Skip Connection
Output Layer	(32, 112, 112) → (1, 224, 224)	Conv2dTransform, ReLU

[9] Yu-An Liu, Ruqing Zhang, Jiafeng Guo, Maarten de Rijke, Yixing Fan, Xueqi Cheng, “Multi-granular Adversarial Attacks against Black-box Neural Ranking Models,” arXiv preprint arXiv:2404.01574, 2024.
[10] Amirian M., “Deep Learning for Robust and Explainable Models in Computer Vision,” arXiv preprint arXiv:2403.18674v1, 2024.

МЕТОДЫ ЗАЩИТЫ ГЛУБОКИХ НЕЙРОННЫХ СЕТЕЙ ОТ ВРАЖДЕБНЫХ АТАК

Гимбицкий В. В., Варашевич А. Г., Ковалёв В. А.

В данной статье рассматриваются методы градиентной атаки на классифицирующие нейронные сети для обработки изображений. А также предложена архитектура нейронной сети для защиты от злоумышленных атак. Для предложенной архитектуры нейронной сети исследована зависимость качества предварительной обработки атакуемых изображений от времени обучения и количества обучаемых параметров.

Received 14.03.2024

Semantic Notation of Access Control Technology based on eID Identification, FIDO2-Authentication and Attribute-Based Authorization in Digital Environment

Anton Zhidovich and Alexei Lubenko and Iosif Vojteshenko
Belarusian State University
Minsk, Belarus
{anton.zhidovich, alexeilubenko02}@gmail.com, voit@bsu.by

***Abstract*—The paper outlines the semantic OSTIS notation of a user identification system based on ICAO electronic documents, FIDO2 authentication and attribute-based access control. Anonymous credentials, which greatly enhance user privacy, are being considered for implementation in the system. Obtained results will allow to increase the efficiency of the joint use of FIDO2 authentication, modern methods of identification and authorization.**

***Keywords*—FIDO2 technology, ABAC, eID, ICAO passwordless authentication, OSTIS technology, biometrics**

I. INTRODUCTION

A growing trend in information security is the use of passwordless authentication methods, in particular the FIDO2 specification, which enhances the user security and privacy compared to traditional password methods.

It's advisable to use semantic notation to describe the process of setting user identity and privileges based on the joint use of the three components mentioned above. This will increase their level of integration, which is particularly important in the context of multi-agent systems where each component may have its own security system.

In the further presentation the universal language of semantic representation of knowledge in the memory of ostis-systems — SC-code and the language of external graphical representation of SC-code constructions — SCg-code will be used to record knowledge [1].

II. FIDO2 AUTHENTICATION

In [2], [3] a semantic approach to designing systems using FIDO2 authentication was proposed and the benefits of FIDO2 were highlighted.

FIDO2-authentication is a modern, secure and convenient phishing-resistant method based on open standards and implemented in browsers and operating systems. The method provides ease of use by allowing users to register their device with a given online service through the selection of a local authentication mechanism [2].

A. Potential drawbacks of FIDO2 authentication

Despite its significant benefits, FIDO2 authentication has a few of potential drawbacks. Access to the hardware or software token loss can make it difficult to recover an account. In most cases, developers solve this problem by adding an additional authentication method, such as password-based.

Another solution is recovery codes. This approach is often used in various systems outside of FIDO2 authentication and in this case all responsibility for storing passwords rests with the user.

III. ATTRIBUTE-BASED ACCESS CONTROL

A. Definition and key benefits

Attribute-based access control
⇒ *acronym**:
[ABAC]
:= [An authorization model where attributes (or characteristics) are used to determine user permissions]

The main benefits of attribute-based access control include:

1) Possibility of building rules close to domain business terms from the modeled subject area.
2) No restrictions on the complexity of rules, which increases the flexibility of the security system.
3) Possibility of supporting rules with dynamic parameters.
4) Possibility of filtering data the user has access to.

Therefore, attribute-based access control provides the ability to create flexible rules that more closely match the requirements of a particular task, given the context and characteristics of the environment. Thus, ABAC as the basis of an authorization system allows the creation of rules that promote semantic interoperability between security components.

B. Attribute-based access control system elements

Attribute-based access control

= {• *Subject attribute*
  ⇒ *synonym**:
    [User attribute]
  := [the characteristics of the user trying to access the resource.]
  ⇒ *example**:
    {• *identifier*
     • *departmental affiliation*
     • *age*
    }
 • *Resource*
  := [the object the subject is trying to gain access.]
  ⇒ *example**:
    {• *file*
     • *application*
     • *API endpoint*
     • *server*
    }
  ∋ *Resource attributes*
    ⇒ *example**:
      {• *resource creation date*
       • *resource owner*
       • *resource type*
      }
 • *Action*
  := [what the user intends to do with the resource.]
  ⇒ *example**:
    {• *create, read, update, delete (CRUD)*
     • *fulfilment*
     • *replication*
    }
 • *Environment*
 := [the context in which the action request is created and processed.]
 ∋ *environmental attributes*
  ⇒ *example**:
    {• *time and location of the action request*
     • *subject device*
     • *network protocols used*
    }
}

C. Usage examples for attribute-based access control

ABAC is widely used in enterprise systems as well as in user applications and IoT devices to improve the security and efficiency of access control to information and resources. It allows a wide range of access problems to be solved with minimal administrative control [4].

Consider examples of ABAC usage:

1) In financial systems to control access to transactions and financial operations.
2) In [5], an ABAC model is proposed to manage access to emergency patient data.
3) ABAC is also used in identity and access management services on such platforms as Amazon Web Services, Google Cloud Platform, Microsoft Asure and Okta.

D. Joint use of FIDO2 and ABAC

The joint use of FIDO2 authentication and attribute-based access control can significantly increase the flexibility and scalability of a security system. It is important to recognise that these are two different processes and that the underlying FIDO2 specification (specifically WebAuthn) does not provide for such integration. Consequently, configuring ABAC and FIDO2 together may require additional effort and resources specific to each system or task.

IV. USE OF ANONYMOUS ATTRIBUTES

By anonymous credentials, we mean a way of implementing attribute-based access control where the matching of user characteristics is proven with Zero Knowledge Proof (ZKP). For example, proving that an age matches a required value without revealing any personal identity.

Anonymous credentials can greatly enhance user privacy, as well as provide authorization with unlinkability. This means that different authorization attempts by a user cannot be linked by a relying party. To achieve this, ZKP is used to prove the user's compliance with the required policy.

Therefore, to implement anonymous credentials in the developed system, it is required that the client (reader) has the software capability to issue zero knowledge proof and the relying party has the capability to verify it.

V. SUBJECT ATTRIBUTES STORAGE

One of the key issues in implementing an attribute-based access control system is the choice of the storage for user attributes. The classic solution is to store this data on the information system side, for example in the database of the authorization service.

In [6] to use the ABAC model in conjunction with the OAuth 2.0 authorization protocol is proposed. This is a logical solution, but in this case there is no way to verify user credentials provided by a service provider such as Google or GitHub. A suitable option is to use electronic documents that store the signature of the issuing party.

A. Electronic identifiers

Let us introduce the definition of electronic identification means in the form of semantic code.

Electronic identification mean
⇒ *synonym**:
[electronic identifier, eID]
:= [an identification document containing up-to-date biographic and biometric information about the owner, the issuer's signature and having built-in data protection mechanisms.]
⇒ *interoperability levels**:
{• *legal*
:= [The use of eIDs is in accordance with government and issuer legal requirements and standards.]
• *organisational*
:= [The use of eID requires adherence to standards and organisational principles, for example in the issuance, implementation and updating of identification tools.]
• *semantical*
:= [eID enables the exchange of structured owner information between participants in a digital environment. This facilitates the establishment of a common meaning and understanding of the data transferred between systems.]
• *technical*
:= [The use of eIDs requires compliance with technical standards such as security specifications, communication protocols and data formats.]
}

One form of eID is an electronic identification card or smart card with an embedded RFID microchip, similar to those used in biometric passports. The chip contains an electronic means of biometric identification with personal data of the biometric document holder in accordance with the requirements of the International Civil Aviation Organisation (ICAO) [7].

In 2005, ICAO's 188 member states approved a new standard requiring all states to begin issuing machine-readable travel documents (MRTDs) in accordance with Doc 9303. The MRTD contains mandatory visual data and separate mandatory summary data in a machine-readable format, as well as a contactless integrated circuit for biometric holder identification [8].

B. eID data structure

To ensure a global level of interoperability and semantic compatibility, ICAO standardized the logical data structure (LDS) that stores information about the electronic identifier holder. It is divided into 16 data groups, each contains specific information about the holder required to identification. (Table I) [9].

Table I: eID Logical Data Structure.

Data Group	Data Element
DG 1	Document Details
DG 2	Encoded Headshot
DG 3	Encoded Face
DG 4	Encoded Fingerprint
DG 5	Encoded Palm print
DG 6	Encoded Iris biometrics
DG 7	Displayed Portrait
DG 8	Reserved for Future Use
DG 9	Signature
DG 10	Data features
DG 11-13	Additional Details
DG 14	CA Public Key
DG 15	AA Public Key
DG 16	Persons to Notify
SOD	Security Data Element

C. eID data protection protocols

To ensure a high level of security and integrity of the data presented on an electronic document, the ICAO standard provides the following security protocols.

Basic Access Control (BAC) is designed to ensure that card data is only accessed when the card is physically present. To do so, a key based on the document number, the holder's date of birth and the document's expiry date is generated. When attempting to access the eID data, the generated key is compared with the stored one, and if they match, access is granted (Figure 1).

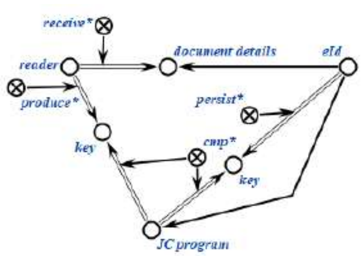


Figure 1: Basic Access Control

Passive Authentication (PA) is a mandatory verification protocol defined in the ICAO standard. Its name refers to the fact that no additional computation is required from the eID chip. The reader retrieves a Document Security Object (DSO) containing a signed hash value of the data. The signature is then verified using the issuer's public key. The hash value of the eID data is also calculated and compared to the value retrieved from the DSO. If these two steps are successfully completed, the passive authentication is passed (Figure 2).

Active Authentication (AA) is designed to protect document data from modification and cloning. It does this by generating a pair of public and private keys when issuing an electronic document. The public key is transmitted with the data to the reader and the private key is stored in the secured eID memory. During active authentication, the received public key is checked against the private key. To do this:

1) The reader sends a cryptographically random string to the eID.
2) The eID signs the string using the stored private key and sends it to the reader.
3) The reader verifies the signature using the received public key. [10] (Figures 2, 3).

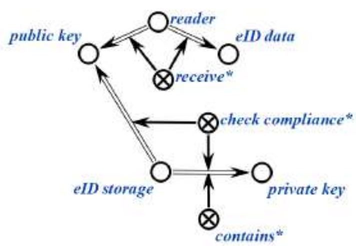


Figure 2: Active Authentication

Chip authentication is an alternative to active authentication, i. e. it allows the terminal (reader) to verify the authenticity of the chip on the electronic document [7].

Electronic documents that comply with the ICAO standard therefore include security mechanisms that allow the authenticity of the document to be verified and prevent the data contained in the document from being stolen or cloned. This standard carefully takes into account modern technological advances and information protection requirements, ensuring a high level of confidence and reliability in the use of documents for personal identification.

VI. PASSWORDLESS AUTHENTICATION WITH ATTRIBUTES AND MEDIATOR (PAwAM)

This chapter considers the implementation of a security system based on identification using ICAO eID, FIDO2 authentication and attribute-based access control.

In [10], a comprehensive and indusrtry-ready solution for the use of anonymous credentials with local or remote attestation is proposed in the form of the FIDO-AC framework, which is an extension of the basic FIDO2 specification. An evaluation of the security and privacy provided by the resulting system and the realisation of a working prototype are also presented.

The essence of the obtained solution is the creation of an additional party — a mediator, which is responsible for the validation of the user data obtained from the eID document.

The FIDO-AC system has the following parties highlighted:

1) The FIDO server (relying party), which forms requests to the FIDO client according to the FIDO2 specification and verifies the response received.
2) The FIDO authenticator used to operate the underlying FIDO2 specification.
3) A client (browser / mobile application with WebAuthn API support) to interact with the FIDO server, the authenticator and the FIDO-AC application.
4) The FIDO-AC application is responsible for reading data from the user's eID. It performs basic access control.
5) The Mediator is responsible for verifying the data received from the eID. It performs active and passive authentication.
6) ICAO eID.

The FIDO-AC application can be specific to each platform and operating system. For example, most modern mobile devices have a built-in NFC sensor with a uniform operating system-level interface, while desktop operating systems have a wide range of different NFC readers from different vendors, the software interaction with which can vary widely. Therefore, Android is proposed as an operating system for FIDO-AC application in [3].

A. *FIDO-AC parties interaction stages*

According to the [10], the work of the FIDO-AC system starts with a user pre-reading the data from the electronic identifier. It is suggested, but not mandatory, to cache the data in the memory of the mobile device for further performance improvements and to reduce the response time of the system.

Then, when attempting to perform an action on an information system resource, the FIDO server generates a signature request that is received by the FIDO client. The request is within the standard WebAuthn specification [11] and contains a cryptographically random bytes buffer — challenge.

On the side of FIDO client, the FIDO-AC application intercepts the request. The FIDO-AC application generates a request containing: eID data hash and signature generated with the issuer's private key for passive authentication; eID public key for active authentication; challenge buffer for unique mediator signature generation. When performing PA and AA, the mediator verifies that the data recorded on the eID card is valid and that the card contains a valid issuer signature.

Once the mediator has successfully verified data 'liveliness', it issues a signature based on the challenge buffer, which also provides additional privacy to the user, as multiple different mediator interactions with the same eID document cannot be linked.

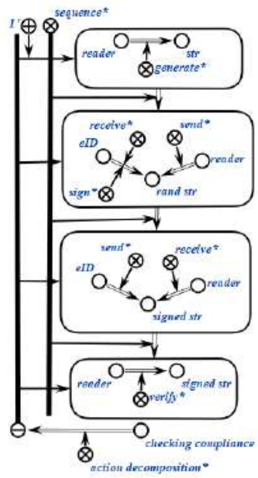


Figure 3: Active Authentication. Checking compliance.

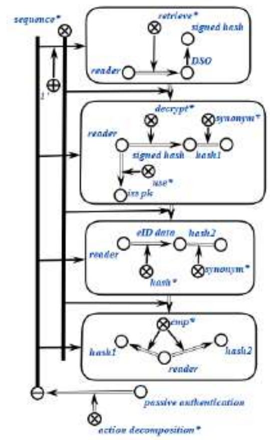


Figure 4: Passive Authentication.

The FIDO-AC application using the ZKP proposed in IV proves, on the one hand, the possession of the subject's attributes and, on the other hand, the conformity of the attributes with the policy required by the FIDO server. The signature issued by the mediator and the proof of the FIDO-AC application are appended to the challenge buffer and sent to the relying party.

At the end of the FIDO-AC process, in contrast to the basic WebAuthn specification, the FIDO server verify not only that the received challenge buffer matches the sent one (except for the attached part), but also the mediator's signature with the proof of FIDO-AC application. After all checks, the FIDO-AC process is considered complete.

The described process can be represented in the form of the SCg-code shown in Figure 3.

B. Implementation options

To implement a FIDO server, it is reasonable to use the ASP.NET Core framework, which has wide deployment and scaling capabilities. As a component allowing the server to act as a relying party, you can use, for example, the FIDO2 component from Rock Solid Knowledge [12].

Web hosting services such as Somee and Google Cloud Platform are offered to test and debug the web application. In addition, most deployment scenarios involve Docker containerisation. In real-world scenarios, the application is deployed on the company's enterprise server.

Currently, all modern desktop and mobile browsers have built-in WebAuthn API support, so any of them can act as a client.

Since FIDO-AC does not require physical separation of the mediator and FIDO-AC application sides, they can be joined together. In this case, there is no need to provide a secure connection between the mediator and the FIDO-AC application.

The Android operating system can be selected as the application platform. In this case, the device must be equipped with a built-in NFC sensor to read the data from the electronic ID.

To achieve correct and proper interaction between an Android device and a biometric document, a JMRTD component can be used [13]. JMRTD provides the ability to connect to the electronic identifier, retrieve data from it and perform the verification described above, supporting ICAO standard documents.

"AVEST" CC, the leading manufacturer of electronic document management security systems in Belarus, provided a functional model of an electronic identifier equivalent to the one used in Belarus as part of the technical assistance for the implementation of this project. It is used for debugging the system and contains the necessary data.

C. Potential vulnerabilities

Let's take a look at the attacks that can be carried out on the mediator / application side of FIDO-AC. One of the key assumptions made by the authors of FIDO-AC is the integrity of the underlying hardware of the device, in our case a mobile phone. This requires the correct operation of the scanner and the trusted platform module, which stores the private keys of the mediator and the platform-dependent authenticator. The latter means that third party access to the device's protected memory is impossible.

The [14] describes that L1 level FIDO2 is vulnerable to timing attacks based on analysing the time taken to execute a cryptographic algorithm depending on the input data. In this way, data such as the private key can be found. A power attack works in a similar way, measuring the power consumed by the device rather than the time taken to complete an operation.

One possible attack is a microarchitecture side-channel attack linked to processor vulnerabilities. For example, a vulnerability in Intel and ARM processors called Meltdown [15] became known some time ago. It exploits

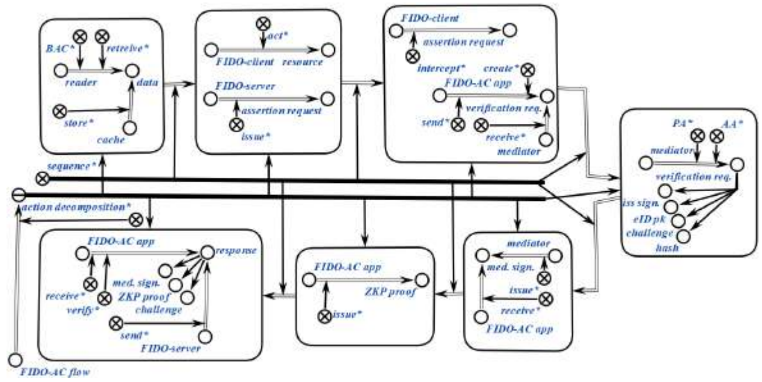


Figure 5: FIDO-AC flow

a bug in the implementation of speculative instruction execution, which causes the processor to ignore page access rights when speculatively executing instructions that read from memory.

In summary, while the mediator side is protected against most software attacks and data interception, it is not protected against side-channel attacks, which are inherently exotic, often require additional equipment, and do not apply to ordinary users of information systems.

Acknowledgment

The authors thank the teams of the Department of Software Engineering of the Belarusian State University for their help and valuable comments as well as "AVEST" CC for technical support for further developments.

References

[1] V. Golenkov, Ed., *Tekhnologiya kompleksnoj podderzhki zhoznennogo cikla semanticheski sovmestimyh intellektual'nyh komp'yuternyh sistem novogo pokoleniya*. Bestprint, Minsk, 2023.

[2] A. Zhidovich, A. Lubenko, I. Vojteshenko, and A. Andrushevich, "Semantic Approach to Designing Applications with Passwordless Authentication According to the FIDO2 Specification," *Otkrytye semanticheskie tehnologii proektirovanija intellektual'nyh sistem [Open semantic technologies for intelligent systems]*, pp. 311–316, 2023.

[3] A. Zhidovich, A. Lubenko, and I. Vojteshenko, "About Enhanced Access Control using FIDO2 Authentication and Attributes," *Informacionnie tehnologii i sistemi 2023 [Information Technologies and Systems 2023 (ITS 2023)]*, pp. 88–89, 2023.

[4] (2024, Jan) Company Okta's Blog. [Online]. Available: https://www.okta.com/blog/2020/09/attribute-based-access-control-abac/

[5] M. T. de Oliveira, Y. Verginadis, L. H. Reis, E. Psarra, I. Patiniotakis, and S. D. Olabarriaga, "Ac-abac: Attribute-based access control for electronic medical records during acute care," *Expert Systems with Applications*, vol. 213, p. 119271, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417422022898

[6] A. Belovodov and O. Laponina, "Ispol'zovanie Upravleniya Dostupom na Osnove Atributov v Protokole OAuth 2.0," *International Journal of Open Information Technologies*, vol. 11, no. 6, pp. 182–189, 2023.

[7] (2024, Feb) The International Civil Aviation Organization. [Online]. Available: https://www.icao.int

[8] (2024, Feb) Doc 9303. Machine Readable Travel Documents: Part 2. [Online]. Available: https://www.icao.int/publications/pages/publication.aspx?docnum=9303

[9] V. K. Kumar, B. Srinivasan, and P. Narendran, "Efficient Implementation of Electronic Passport Scheme Using Cryptographic Security Along With Multiple Biometrics," *International Journal of Information Engineering and Electronic Business*, vol. 4, no. 1, pp. 18–24, 2012.

[10] W.-Z. Yeoh, M. Kepkowski, G. Heide, D. Kaafar, and L. Hanzlik, "Fast IDentity online with anonymous credentials (FIDO-AC)," in *32nd USENIX Security Symposium (USENIX Security 23)*. Anaheim, CA: USENIX Association, Aug. 2023, pp. 3029–3046. [Online]. Available: https://www.usenix.org/conference/usenixsecurity23/presentation/yeoh

[11] (2022, Jan) Webauthn Specification. [Online]. Available: https://w3c.github.io/webauthn/

[12] (2023, Dec) FIDO2 for ASP.NET | Unphishable second factor and passwordless authentication for ASP.NET Core. [Online]. Available: https://www.identityserver.com/products/fido2-for-aspnet

[13] (2024, Jan) JMRTD: An Open Source Java Implementation of Machine Readable Travel Documents. [Online]. Available: https://jmrtd.org/

[14] M. Kepkowski, "How not to handle keys: Timing attacks on fido authenticator privacy," *22nd Privacy Enhancing Technologies Symposium*, pp. 705–726, 2022.

[15] C. Metz and N. Perlroth. The New York Times: Researchers Discover Two Major Flaws in the World's Computers. [Online]. Available: https://www.nytimes.com/2018/01/03/business/computer-flaws.html

СЕМАНТИЧЕСКАЯ НОТАЦИЯ ТЕХНОЛОГИИ УПРАВЛЕНИЯ ДОСТУПОМ НА ОСНОВЕ EID-ИДЕНТИФИКАЦИИ, FIDO2-АУТЕНТИФИКАЦИИ И АТРИБУТИВНОЙ АВТОРИЗАЦИИ В ЦИФРОВОЙ СРЕДЕ

Жидович А. А., Лубенько А. А., Войтешенко И. С.

В работе изложена семантическая OSTIS-нотация системы установки личности пользователя и его привилегий, основанной на идентификации с использованием электронных документов, соответствующих стандартам ICAO, беспарольной FIDO2-аутентификации и управления доступом на основе атрибутов. Рассмотрен вариант реализации в системе анонимных учетных данных, значительно повышающих конфиденциальность пользователя.

Полученные результаты позволят повысть эффективность совместного использования FIDO2-аутентификации с современными методами идентификации и авторизации.

Received 13.03.2024

Intelligent Urban Management System by Minibus

Vasili Shuts, Andrei Kazinski, Evgenii Prolisko
dept. Intelligent Information Technologies
Brest State Technical University
Brest, Belarus
lucking@mail.ru
kaa1964@bk.ru
prolisko@mail.ru

***Abstract*—The results of the analysis of the problems of passenger transportation by urban public transport are presented. More attention is paid to transportation using fixed-route taxis. To improve the efficiency of passenger transportation, an automated irregular transport management system developed at the Brest State Technical University is proposed. The article describes the structural components of the system and their functionality. The source of the development is the observation of the behavior of drivers, passengers and other road users. The exchange of information between the subjects of the movement is based on unified knowledge about the nature of the movement. This approach requires the use of semantic technologies in passenger transportation management systems. This publication describes the first stage of the project, designed to collect information and its semantic analysis**

***Keywords*—city minibus, passenger transportation management system, transportation on request, semantic model**

I. Introduction

Currently, public transport infrastructure is actively developing in cities around the world. This is due to population growth, as well as the fact that public transport is an ecological alternative to personal transport.

The main task of organizing the movement of urban transport is to ensure high quality passenger transportation at a minimum cost. The quality of passenger transportation is assessed by the following indicators: ensuring the regularity of movement, the size of the interval on the route, passenger occupancy, the total time spent by the passenger on the trip, the speed of communication and the comfort of transport services.

The system of scheduling public transport is not optimal. According to scientific research carried out at the company "Altimeta" [1] many management systems rely on relational databases, document-oriented information technologies, procedural algorithmic programming paradigms, and stack architecture. All these aspects reflect the features of the automation object and are determined by its features. The use of semantic models can bring productivity gains and other useful changes. A possible approach is to apply a semiotic approach to the creation of automation systems. An example of such an implementation is the Smart Stop mobile application [2]. The application of a semiotic approach to passenger transportation is aimed at integrating technical means and management methods in design solutions that meet the needs of all participants. This combination allows you to have a set level of satisfaction of both the needs of passengers and the optimization of the costs of carriers. In the proposed semiotic approach, passenger transport is analyzed in the context of the movement of material and information (symbolic, symbolic) resources. The urban passenger transport network in a semiotic context can be represented as a system of multilevel hypertext. At the same time, passenger traffic flows are expressed in the form of a system of symbolic communications. Such symbolic communications in automated systems can act not only as an accompaniment to traffic flows. An important feature of modern systems based on the semiotic approach is that it is the traffic flows that are motivated by symbolic communications. Such communications satisfy the informational, symbolic and symbolic needs of passengers. At the same time, it is the needs of passengers expressed by semiotic means that should be the primary source of urban passenger traffic flows. The implementation of the semiotic approach proposed by the authors [2] requires a different approach to the logistics system of passenger transportation. With this approach, significant connections are identified that determine the target functionality of the system as a whole. The main objectives of such links should be to provide each passenger with an appropriate quality service for transportation on request. The request must reflect the time of transportation, route and other properties. It should be noted that the level of service quality should also be determined by the cost of transportation with continuous adaptation to changing environmental conditions. In the proposed system, consideration of semiotic aspects is implemented by means of the user interface, personalization procedures. These tools can significantly improve two aspects that determine the basic needs of the passenger: the correspondence of the information received about possible transportation to the information requests of passengers (persistence of response to search queries), the degree of adequacy of the results of transportation to the expectations of the passenger (relevance of passenger transportation). In this paper, the main content is devoted to another author's research, which is a continuation of the above. The project of an intelligent management system for urban minibuses allows you to generate more meaningful information. Such information will be the basis for the implementation of an intelligent management system using semantic technologies [2].

Before describing the project, we will pay attention to the conceptual provisions that form the basis for building a semantic model of the city's passenger transport system. This order generally corresponds to the correlation of the concepts of semiotic and semantic modeling implemented by the authors in the studies of passenger traffic flows in the city of Brest.

It should be pointed out that the analysis of research on the application of neural network technologies for semantic analysis tasks has been carried out in the works of a fairly wide range of scientists.

For example, in the source [3] provides a list of research areas that use neural network technologies to solve semantic analysis problems. Among these areas are listed:

- tasks of text classification;
- brain activity in pattern recognition based on signals processed by different areas of the cortex;
- building models of language learning based on sentence processing without prior knowledge of word semantics and others.

Quite interesting results are presented based on the results of the development of the OSTIS Project [4] These include:

- development of semantically compatible computer knowledge management systems for distributed solution of complex problems and their applications;
- practice-oriented training of students and undergraduates based on OSTIS technologies;
- clarification of the main provisions defining the interaction of science, education, engineering and business in the market of knowledge-driven computer systems.

Among other examples, the works of V. can be given. Golovko, A. Kroshchanka and others [5] D. Shunkevich and others [6], [7], I. Davydenko [8], I. Koronchik [9], [10] and many other studies.

It should be noted that the works listed above are devoted to fundamental areas of research in the field of semantic technologies. The main applications of research are their theoretical aspects. However, the practice of using semantic technologies for applied research cannot be ignored. As a possible application, we have selected projects based on knowledge processing, used to solve applied problems in the field of passenger transportation. An example of such a task is a semantic approach for designing automated urban passenger transportation management systems.

The most acute problem of traffic management is for a minibus. The route taxi schedule is more based on the set departure time from the final stopping points of the route. There is no timetable for intermediate points along the route. Minibus drivers receive information about the number of passengers waiting at stops and the density of passenger traffic by exchanging voice messages among themselves.

The information provided by the driver of the minibus in front affects the speed of movement along the route, allows you to adjust the time of departure from the final stops. However, this does not solve the general problems of the transport system's capacity.

The frequency of traffic does not change during the day. This indicator does not correspond to possible changes in the intensity of passenger traffic. Passenger traffic, in turn, depends on various indicators (weather conditions, working hours of passengers during the day and days of the week, etc.). The lack of funds to regulate the route taxi schedule leads to collisions. Collisions in the movement of a minibus taxi should be considered the occurrence of situations in which there is a sharp contradiction between the interests of the carrier and the passenger. So, in case of an increase in passenger traffic, situations arise in which public transport is loaded to the maximum, but there are passengers at the stops who are denied transportation. A decrease in passenger traffic leads to the movement of "half-empty" passenger vehicles. The so-called automatic transport gives special hopes for improving the situation, while still in the initial stage of development [**?**].

The problem of resolving collisions in passenger transportation can be solved by optimizing the operation of a minibus. On the one hand, such optimization should be carried out in the direction of increasing the carrier's profit. On the other hand, it is necessary to improve the comfort of the transport network for passengers. Currently, automated systems are not used in Belarus to solve the problems of improving the efficiency of transportation by minibus. However, there are similar solutions for another type of taxi. An example of a solution is the "Yandex Taxi" application.

II. Functioning Of The Transport System According To The "Transport On Demand Model"

The authors propose an optimized model of urban transport based on the "Transport on demand" model. This utility model is described in the patent "System for calling a minibus taxi, management and control of passenger transportation", obtained by the authors [11].

The model is developed based on the analysis of existing systems.

A. Features of well-known systems used to control passenger transportation

A device is known for indicating empty seats in fixed-route taxis (patent for utility model BY 4186 u 2008.02.28), including sensors for detecting the presence of a passenger, by installing contact devices under each seat of a minibus, a data generation and processing unit, an indication unit. The task of the utility model is to create a device for indicating empty seats in fixed-route taxis, which makes it easier for both its driver and passengers inside the cabin and at the bus stop to receive information about the availability of free seats in vehicles.

The disadvantage of this device is the low functionality, which consists in the fact that the minibus is relatively small (15-25 seats) and the driver is in the mirror condition of the cabin. He can visually determine the availability of seats. Also, passengers waiting for transport at the bus stop do not need external information about available seats in the cabin. The taxi has stopped, which means there are places, and passengers enter the vehicle. If there are no available seats, the minibus does not stop at the stop at the passenger's request. Thus, the information content of this device does not justify the cost of its implementation.

The GeoBus automated system is known (GeoBus program for Android OS, State registration of programs No. RU2021660693) for booking and ordering tickets (seats) in public transport by passengers; managing free and occupied places in public transport by route drivers; dispatching control by public transport route drivers; geolocation of drivers and passengers; tracking public transport routes in real time; providing information about public transport routes; calculation of the approximate time of arrival of public transport at the stop.

The disadvantage of such a system is the need for dispatching transport management and the inability to view the number of boarding requests for the entire route. A route is understood as a pair: the route number and the direction of travel (the starting and ending stopping points of the route).

Analogues and prototypes for fixed-route taxis have not been identified anymore. But there is a large class of devices and systems for taxis that are close in functionality to the proposed utility model for calling a minibus, managing and controlling passenger transportation.

So there is a known method and system for managing taxi orders using automated dispatch control systems for taxi operation (patent for invention No. RU 2297042, G07C 5/00, 2007.04.10) in which the system contains at least one external terminal for accepting orders, a server of a single database of orders and one terminal in each taxi company. In this case, the order placed by the customer via telephone lines and/or from an external terminal is entered by the operator into a single database of orders. In this case, the order criteria are specified. An order entered into a single database is processed by a system that sends a specific order to a taxi company selected by rating, after which the taxi company operator processes the order and determines whether it can be fulfilled, while there is a possibility of canceling the order. In this case, the order is sent back to the unified system, which is sent by the program to the next taxi company for processing and execution.

The disadvantage of this system is the complexity of order fulfillment and inefficient operation of the order acceptance and processing system.

There are known systems for collecting and transmitting messages about vehicles containing a navigation signal receiver installed on a vehicle, a control unit, memory blocks with information about the vehicle, a message transmission and reception device for transmitting and receiving information from a control room. At the control room, known systems contain a message transmission and reception device, a message allocation unit, a database and a display tool (RU 2113013 C1, G08G 1/01, 06/10/1998; RU 2143745 C1, G08C 19/12, 11/27/1999 and RU 2173885 C1, G07C 5/00, 09/20/2001).

The disadvantage of these systems is the use of dispatchers exclusively, when the quality of service provision directly depends on the human factor.

A system is also known (patent for invention No. RU 2173885 C1, G07C 5/00, 09/20/2001) for controlling and monitoring the operation of vehicles, containing on each vehicle a device including a message generator, to which a GPS receiver, a current time counter, a status code generator are connected, the output of the message generator via a transceiver is connected to the transceiver of the control room device, on which the transceiver is located, connected via the parameter selection block to the data memory, designed to store current data on the parameters of all operating vehicles. The well-known system provides an audio communication unit between the dispatcher and the drivers.

The disadvantages of this solution include the use of expensive GPS equipment in the system, which requires the modernization of the technical service park of each taxi, while the quality of voice communication involved in the system is limited by the bandwidth of voice communication channels, which leads to errors when receiving information by system participants.

An automated system for ordering taxi services and monitoring the operation of a taxi is also known according to utility model patent No. RU 103952, G07C 5/00, 08/23/2010, containing at least one web server of the system connected via communication channels to the control system of remote dispatching taxi services and a user communication device. At the same time, all means of communication and the central computer are designed to function in the environment of a pre-installed program that provides online connection of all these elements of the device to each other, support for the functions of the control panel and display necessary for viewing and editing information contained in memory using each of the elements of the software module installed in the operating system, as well as the formation and transmission of data and the conduct of information exchange of data.

Among the disadvantages of this system is the need to install specialized software on the means of communication, which ensures the operability of the system, while all operations, including accounting for trips, settlement operations and other operations for the control of transportation are carried out in this program. This requirement may not be applicable for small taxi companies and/or individual drivers who keep records under simplified programs. Besides. The installation of this software requires the retrofitting of remote control rooms and driver workstations with communication devices compatible with this software.

The closest in terms of its operational characteristics to the claimed technical solution is a taxi calling system, management and control of passenger transportation according to utility model patent No. RU 126493, G07C 5/00, 06/9/2012, including at least one web server connected via communication channels to the control system of remote dispatching taxi services and user communication devices.

The Web server of the system contains interconnected, by means of system interface trunks and multiplex channels of information exchange, control and data processing units with a memory block containing at least one, implemented in a software and hardware manner, a consolidated database of stored order processing procedures, data on vehicles registered in the system, data about the user and the remote dispatch services registered in the system, and is connected to the dispatching services system and user communication devices by means of a hardware-software control module of the API system via an Internet gateway with feedback, with the ability to accept and process order data from user communication devices, select the order parameters of the vehicle registered in the system and transfer order information to the remote control system in the control room, with subsequent control of the order execution via the feedback line and the API module.

The disadvantage of this system (prototype) and all other above-mentioned systems is the presence of a dispatch service, which makes the system automated, not automatic.

The second disadvantage of this system is the presence of many unnecessary functions that are needed for a taxi, but are absolutely not required for a minibus that operates on a given, fixed route of the city. The second feature of a minibus taxi is the large number of passengers transported compared to a taxi. These two differences require different control algorithms while maintaining many matching hardware.

Thus, the complexity, high hardware costs, the lack of direct interaction between the driver of the vehicle and the passenger, as well as the lack of functions that are needed for optimal transportation management, namely by minibus, are the disadvantages of the prototype device.

The objective of this utility model is to create an automatic system for calling a minibus taxi, managing and controlling passenger transportation, free from the above disadvantages of systems of this type.

B. Description of the utility model

At the moment, there is no such thing as "calling a minibus" in route transportation. The passenger, having approached the stop, patiently waits for the vehicle. In this system, he reads the QR code of the stop at the bus stop with his smartphone. From that moment on, it is initialized in the system, the request for its service was received in it, as in a taxi. There is an individual order in the taxi. In this system, this is a collective order. As soon as a sufficient number of passengers accumulate, the process of servicing them begins immediately.

The passenger is the first participant (subject) of the system. The second participant is a minibus driver. If the passenger's goal is to minimize the waiting time, then the driver's goal is to maximize the loading of the vehicle interior. At the moment, when leaving the initial (final) point of the route, the driver does not know how many passengers are on the route and whether there are any at all. Therefore, the moment of departure on the route is important for him. Hence the dilemma: I left early, then I drove empty; I left late, then I couldn't take passengers at some stops because of the overcrowding of the cabin.

At the moment, there is no such thing as "calling a minibus" in route transportation. The passenger, having approached the stop, patiently waits for the vehicle. In this system, he reads the QR code of the stop at the bus stop with his smartphone. From that moment on, it is initialized in the system, the request for its service was received in it, as in a taxi. There is an individual order in the taxi. In this system, this is a collective order. As soon as a sufficient number of passengers accumulate, the process of servicing them begins immediately.

The passenger is the first participant (subject) of the system. The second participant is a minibus driver. If the passenger's goal is to minimize the waiting time, then the driver's goal is to maximize the loading of the vehicle interior. At the moment, when leaving the initial (final) point of the route, the driver does not know how many passengers are on the route and whether there are

any at all. Therefore, the moment of departure on the route is important for him. These situations are examples of collisions in which the interests of the driver and passenger collide. So, if the driver left earlier than the required number of applications was received, then the cabin will not be filled efficiently on the route. In case of late departure, the number of requests for transportation will exceed the cabin capacity. Passengers will receive a denial of service.

The proposed technical solution makes it possible to exclude both of these "bad cases", since the driver at any given time sees from his smartphone how many passengers are waiting for him on the route, as well as at each stop separately. Moreover, from the experience (statistics) of previous trips and the time of day, the system itself offers the driver the optimal time to set off on the route.

The technical result of the claimed utility model is to simplify the process of interaction between users of the system (passenger, driver) among themselves, improve the quality of service for minibus passengers due to the minimum delivery time of a minibus to the client, optimize passenger transportation by creating a single database of orders and optimal time for departure on the route.

The main features of the proposed minibus calling system are:

- the inclusion of a server for query management, a database and a system for managing this data in the system;
- the inclusion of a passenger client and a driver client in the system, made for the Android operating system;
- the presence of a server in the system, which is connected using the Internet;
- using client-server technology for system development;
- connecting users to the system through their existing smartphones;
- inclusion of QR codes of route stops in the system.

When developing the system, the authors took into account some of the features of the systems proposed earlier. These include: a method for determining the degree of occupancy of the cabin, the presence of a dispatching transport control to determine the number of requests from passengers for service, the presence of geolocation systems to determine the location of public transport, calculating the time of arrival of public transport at a stop, and others. So, in some cases, to determine the number of available seats, it is proposed to use special devices – sensors for indicating empty seats. External terminals are used to manage orders using automated systems. In such terminals, as a rule, registration of a service request is used using a telephone or a dispatcher accepting the request. There are also known vehicle control systems using devices for generating messages using GPS navigators, current time counters and other expensive equipment. Such equipment includes a sound communication unit between the dispatcher and the drivers.

The closest in terms of its operational characteristics to the solution proposed by the authors is the taxi calling, passenger transportation management and control system, which includes web servers that connect remote taxi dispatching services and user communication devices. Web servers of such systems, as a rule, contain a database with information about incoming orders, data on registered passenger vehicles on routes, data on users of the system and dispatching services.

Among the disadvantages of the existing systems, we have identified the following:

- the complexity of registration and execution of orders;
- low efficiency of the registration and order processing systems;
- the need to install additional equipment on vehicles (receivers of navigation signals, control units, memory blocks with information about the vehicle, information exchange devices with a dispatch point, and others);
- the presence of functions in the system that complicate its operation and use by users.

One of the main disadvantages of the existing passenger transportation management systems is the lack of direct communication between the passenger and the driver.

All these factors served as the basis for setting the task of creating a useful model of a minibus taxi calling system. For a more detailed acquaintance with the utility model, you should refer to the content of the authors' patent [11].

Based on the utility model, an automated system is being developed to optimize the operation of a minibus taxi, which has the following functionality:

- creating an application for a passenger waiting at a bus stop;
- informing the driver about the busy route;
- informing the driver about the number of passengers who enter and exit the taxi cabin at each subsequent stop.

Passenger traffic data is stored and transmitted in the form of requests to the server. Each request is a set of data defining the route, starting and ending stops. The route, in turn, is determined by an ordered sequence of stops.

The diagram of the interaction of the components of the automated system is shown in Figure 1.

The interaction of the system components takes place according to the REST (Representative state transfer) architecture.

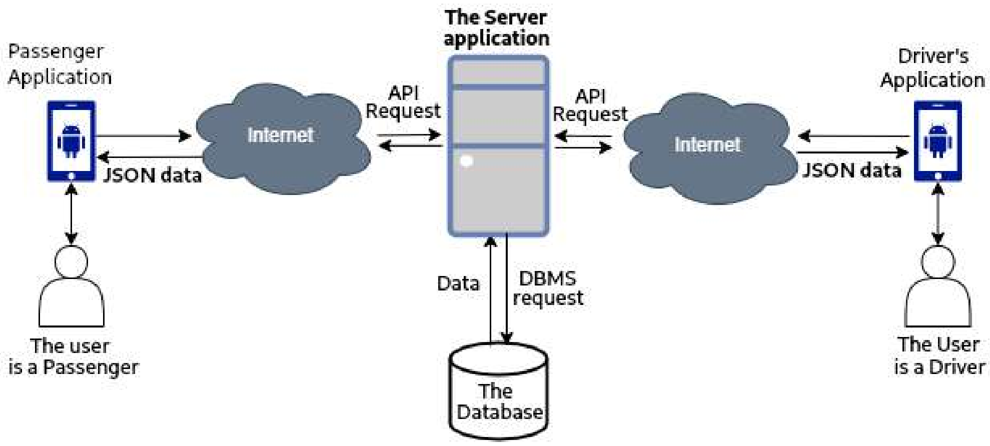


Figure 1. Flowchart of the passenger client application.

According to the information received by the server, the driver will receive an activation signal that indicates the departure time from the current stop point. Departure takes place at the time corresponding to the estimated time of maximum passenger occupancy at one of the stops on the route.

The components of the client-server system are applications for the two main participants of minibus transportation – the passenger and the driver, as well as the software implementation of the server. The user part of the system is represented by two separate client-server applications. Both applications are primarily focused on use on mobile devices.

The Passenger client application allows you to send a request for transportation along the route. The application guarantees the time interval during which the specified transport will arrive. Also, the passenger user has the opportunity to track the location of the called car. The passenger client application allows you to send a request to the server from a stop, which indicates: the location on the route, the selected route, the starting and ending points of passenger transportation. The server processes the received information and accumulates data. When a situation arises that is necessary and sufficient to send a transport along the route, the server sends an activation signal to the application for the driver client. The generation of an activation signal based on accumulated data on passenger requests can significantly reduce the number of collisions in passenger transportation or avoid them altogether. The absence of collisions leads to lower transportation costs and improved passenger service quality.

III. The Server Of The Automated Taxi Management System

The server acts as the central element of the system. Tasks solved by the server: providing data flow management, providing a service for processing data requests from client components, ensuring the storage of information in the database, interacting with the database management system (DBMS).
The JavaScript programming language was used to implement the server, in particular the “Node frameworks.js”, “Express.js” and “Mongoose”. Node.js is a software platform that transforms JavaScript from a highly specialized language into a general–purpose language. In the developed system, the web server is “Express.js” serves as the standard framework for “Node.js” and implements the mechanisms of the server component working with requests. Mongoose introduces a special ODM library for working with MongoDB. The ODM library allows you to map class objects and collection documents from a database.

The storage of information about passengers, drivers, stops, routes and other data is provided by the MongoDB database management system.

Tasks solved by the server component:

- registration of a request from a passenger client for transportation along the route between selected stops;
- sending information about passengers waiting at the stop;
- sending information about passengers leaving the vehicle at a stop;
- accounting for the dynamics of all incoming requests and requests from mobile and static customers;

- collecting statistical information and making recommendations to improve the efficiency of passenger transportation;
- maintaining a database of an automated route transport management system;
- registration of drivers in the system;
- providing a sequence of stops and routes on request.

IV. Passenger Client Application

The passenger and driver client applications were created in Java in the Android Studio environment for mobile devices running on the Android operating system. The JSON standard is used to exchange data with the server.

The passenger client application provides the user with the opportunity to generate requests in two modes: the mode of choosing the desired route or the mode of choosing stops. The flowchart of the passenger client application is shown in Figure 2.

In the case of the first mode, the user selects the route number, starting and ending stops. At the end of the selection, the passenger's request is generated and sent to the server.

Selecting the second mode allows the user to specify only the starting and ending stops. Based on the choice of route points, the application suggests a suitable route or reports the lack of direct communication between them.

Instead of specifying the point of departure, the passenger is asked to scan the QR code placed at the stop.

Tasks solved by the passenger-client component:

- identification of the stop where the user is located by means of a QR code;
- providing complete information on the routes passing through this stop;
- formation of requests for transportation and sending them to the system server.

V. Client Application For The Driver

The flowchart of the driver client application is shown in Figure 3.

To use the driver client application, the user must register and log in to the system.

The application for the driver client provides the user with information about the current route number, the nearest stop point, and the current number of passengers in the minibus. The number of passengers is recorded on the server by the client application. However, not every passenger uses it. For this reason, the user of the driver client has the opportunity to adjust the number of passengers.

The implementation of the driver client also allows you to keep statistics on the number of users of the application for the passenger client.

Tasks solved by the client-driver component:

- displaying information about the number of passengers in the vehicle with the possibility of its adjustment;
- display information about the number of passengers leaving the vehicle at the next stop;
- display information about the number of passengers waiting for the vehicle at the next stop;
- registration and authorization of a minibus driver in the system.

VI. Conclusion

The presented automated management system for irregular transport makes it possible to effectively solve the tasks set: improving the efficiency of transportation and the quality of passenger service, eliminating collisions of passenger flows. The system can also be used to optimize the interaction of traffic participants and improve the efficiency of the urban transport system as a whole. Currently, an automated management system for irregular transport is being implemented on individual routes. In case of positive results in the future, the system can be implemented for the entire transport network of the city. Using such a system will increase profits and improve the quality of passenger transportation.

References

[1] Why semantics and why now? [Electronic resource]. – Mode of access:http://www.ultimeta.ru/technologies/semantic_technologies/pochemu-semantika-i-pochemu-seychas/. – Date of access 03.04.2024

[2] Zhogal, A.N., Kontsevoy M.M. Development of a mobile application for the Smart Stop information system / A.N. Zhogal, M.M. Kontsevoy // Materials of the International Scientific and Technical Conference "Artificial Intelligence. Intelligent Transport Systems (Be-Safe 2016)", dedicated to the 50th anniversary of the Brest State Technical University, May 25-28, 2016, Brest, 2016 – P. 93-96.

[3] Averkin A.N. Neural networks in semantic analysis / A.N. Averkin, S.Y. Yarushev [Electronic resource]. – Mode of access:https://libeldoc.bsuir.by/bitstream/123456789/7004/1/Averkin_Neyronnyye.PDF. – Date of access 03.04.2024

[4] Golenkov V.V. Semantic technology of component design of systems, managed by knowledges / V.V. Golenkov, N.A. Guliakina [Electronic resource] – Mode of access:https://libeldoc.bsuir.by/bitstream/123456789/3926/1/Golenkov_Semanticheskaya.PDF. – Date of access 03.04.2024

[5] Integration of artificial neural networks and knowledge bases / V. Golovko, A. Kroshchanka, V. Golenkov, V. Ivashenko, M. Kovalev, V. Taberko, D. Ivaniuk // Open semantic technologies for designing intelligent systems, 2018, vol. 8, issue 2, pp. 133-147.

[6] The methodology of component design of knowledge-driven systems / D. V. Shunkevich, I. T. Davydenko, D. N. Koronchik, A.V. Gubarevich, A. S. Boriskin // Open semantic technologies for the design of intelligent systems (OSTIS-2015) : proceedings of the V International Scientific and Technical conf., Minsk, February 19-21, 2015 / Belarusian State University of Informatics and Radioelectronics. – Minsk, 2015. – pp. 93-110.

[7] Shunkevich, D. Ontology-based design of knowledge processing machines / D. Shunkevich // Open semantic technologies for the design of intelligent systems. – Minsk, 2017. – Issue 1.– P. 73–94.

[8] Davydenko, I. Ontology-based knowledge base design / I. Davydenko // Open semantic technologies for the design of intelligent systems. – Minsk, 2017. – Issue 1. – P. 57–72.

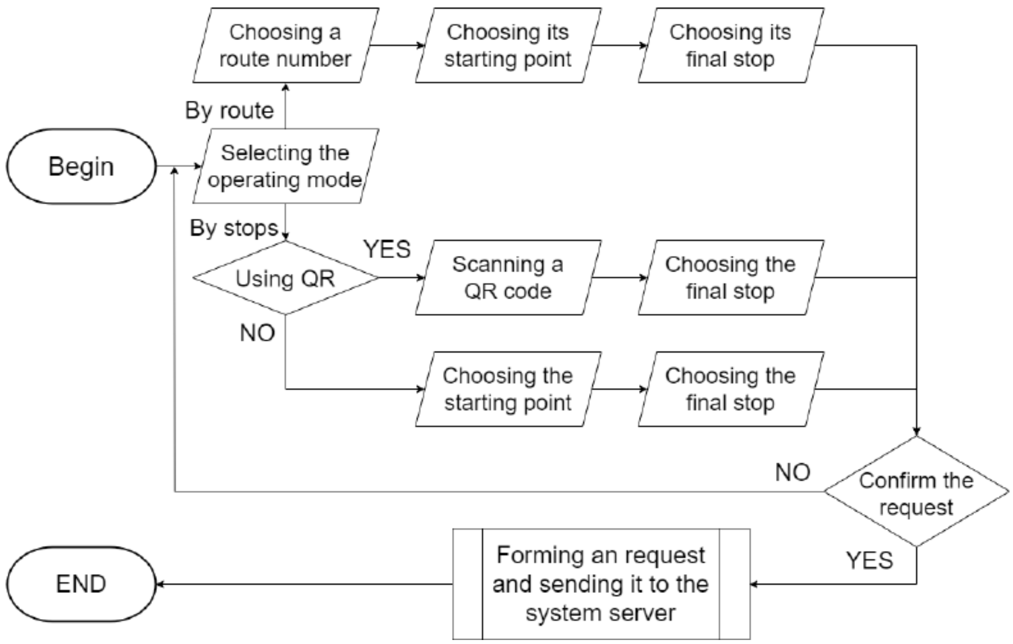


Figure 2. Flowchart of the passenger client application.

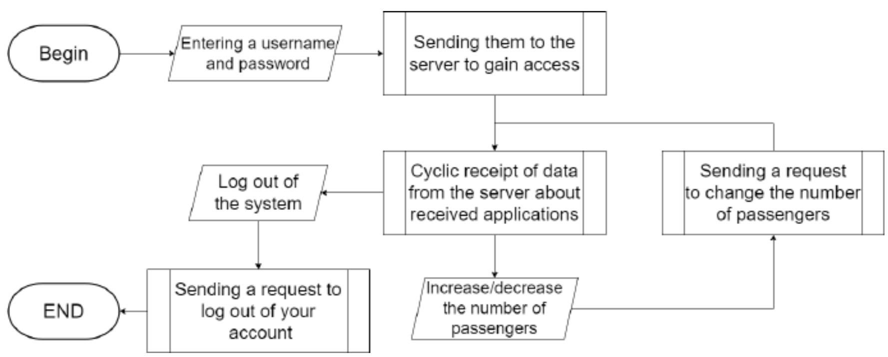


Figure 3. Flowchart of the client application for the driver.

[9] Koronchik, D. N. Implementation of web-platform for systems based on knowledges / D. N. Koronchik // Open semantic technologies for designing intelligent systems (OSTIS-2015) : materials of the V International Scientific and Technical conf., Minsk, February 19-21, 2015 / Belarusian State University of Informatics and Radioelectronics. – Minsk, 20-15. – pp. 89-92

[10] Koronchik, D. N. Implementation of the repository of unified semantic networks / D. N. Koronchik // Open semantic technologies for designing intelligent systems (OSTIS-2013) : materials of the III International Scientific and Technical conf., Minsk, February 21-23, 2013 / Belarusian State University of Informatics and Radioelectronics. – Minsk, 2013. – pp. 125-129.

[11] The system for calling a minibus, managing and controlling passenger transportation : utility model patent No. 13345 Rep. Belarus G 07C 5/00 / V. Shuts, A. Kazinski, A. Levchuk, A. Zhogal ; applicant Brest State Technical University – No. u 20230114 ; announced 02.06.23. – Mode of access: https://search.ncip.by/database/index.php?pref=mod&lng=ru&page=3&target=24258. – Date of access 03.01.2024

ИНТЕЛЛЕКТУАЛЬНАЯ СИСТЕМА УПРАВЛЕНИЯ ГОРОДСКИМИ МАРШРУТНЫМИ ТАКСИ

Шуть В., Козинский А., Пролиско Е.

Представлены результаты анализа проблем перевозки пассажиров городским общественным транспортом. Большее внимание уделено перевозкам с использованием маршрутных такси. Для повышения эффективности перевозки пассажиров предлагается автоматизированная система управления нерегулярным транспортом, разработанная в Брестском государственном техническом университете. В статье описаны структурные компоненты системы и их функционал. Источником разработки являются наблюдения за поведением водителей, пассажиров и других участников движения. Обмен информацией между субъектами движения опирается на унифицированные знания о характере движения. Такой подход требует использования семантических технологий в системах управления пассажирскими перевозками. Данная публикация описывает первый этап проекта, предназначенный для сбора информации и ее семантического анализа.

Received 08.03.2024

AUTHOR INDEX

АВТОРСКИЙ УКАЗАТЕЛЬ

Научное издание

Открытые семантические технологии проектирования интеллектуальных систем

Open Semantic Technologies for Intelligent Systems

Сборник научных трудов

Основан в 2017 году

Выпуск 8

В авторской редакции

Ответственный за выпуск *В. В. Голенков*
Компьютерная вёрстка *Н. В. Гракова*

Подписано в печать 11.04.2024. Формат 60*84 1/8.
Бумага офсетная. Гарнитура «Таймс». Ризография.
Усл. печ. л. 45,34. Уч.-изд. л. 39,97.
Тираж 105 экз. Заказ 46.

Полиграфическое исполнение
УП «Бестпринт». Свидетельство о государственной регистрации издателя, изготовителя, распространителя печатных изданий №1/160 от 27.01.2014.
Ул. Кальварийская, д. 25, каб. 116, 220026, г. Минск