Том 73
Выпуск 1

ИСА РАН

ISSN 2079-0279
2023

ТРУДЫ

ИНСТИТУТА СИСТЕМНОГО АНАЛИЗА РОССИЙСКОЙ АКАДЕМИИ НАУК

http://www.ISA.ru

Информационные технологии

Интеллектуальный анализ данных

Методы и модели в естественных науках

Компьютерный анализ текстов

POLY
PRINT
SERVICE

Труды Института системного анализа Российской академии наук (ИСА РАН)

http://www.isa.ru/proceedings/

Журнал основан в 2002 году
Журнал выходит ежеквартально

Учредитель
Федеральное государственное учреждение «Федеральный исследовательский центр «Информатика и управление»
Российской академии наук» (ФИЦ ИУ РАН)
119333, г. Москва, ул. Вавилова, д. 44, корп. 2, тел.: 8 (499) 135-62-60, e-mail: frccsc@frccsc.ru, http://www.frccsc.ru/

Главный редактор
Попков Ю.С., академик РАН
117312, г. Москва, проспект 60-летия Октября, д. 9
тел. 8 (499) 135-42-22, e-mail: popkov@isa.ru

Адрес редакции
117312, г. Москва, проспект 60-летия Октября, д. 9,
E-mail: tisa@isa.ru, тел. 8 (499) 135-24-33, 8 (926) 490-87-87

Научный журнал «Труды Института системного анализа Российской академии наук (ИСА РАН)» публикует материалы по широкому кругу фундаментальных проблем развития методологии системного анализа и ее приложений к решению разнообразных задач в сфере науки и практики. Журнал предназначен для научных сотрудников, инженеров и исследователей, работающих в рамках этой проблематики, а также для политиков, сотрудников государственных и муниципальных органов, специалистов предприятий и представителей общественных организаций. Правила подготовки статей для публикации в журнале, а также правила их рецензирования приведены на сайте журнала.

Журнал зарегистрирован в Федеральной службе по надзору в сфере связи, информационных технологий и массовых коммуникаций (Роскомнадзор). Свидетельство о регистрации средства массовой информации ПИ № ФС77-65718 от 20 мая 2016 г.
Подписной индекс журнала – 36089 в электронном каталоге ГК «УРАЛ-ПРЕСС» (http://ural-press.ru/).
Журнал включен в перечень ведущих рецензируемых научных изданий, рекомендованных ВАК для публикации основных научных результатов диссертаций на соискание ученых степеней доктора и кандидата наук.

ПОЗДРАВЛЯЕМ В.Н. ЛИВШИЦА
С ПРИСУЖДЕНИЕМ СЕРЕБРЯНОЙ МЕДАЛИ
ВОЛЬНОГО ЭКОНОМИЧЕСКОГО ОБЩЕСТВА РОССИИ!

Вениамину Наумовичу Лившицу, главному научному сотруднику Федерального исследовательского центра «Информатика и управление» Российской академии наук, заслуженному деятелю науки Российской Федерации, профессору, доктору экономических наук присуждена Серебряная медаль Вольного Экономического Общества России за вклад в развитие экономической науки и участие в деятельности общественных институтов нашей страны.

Вениамин Наумович – выдающийся ученый в области математического моделирования социально-экономических процессов, методов оптимизации решений динамических задач в экономике, оценки эффективности инвестиционных и инновационных проектов, обоснования рациональных стратегий развития производственной инфраструктуры, реформирования естественных монополий, экономики транспорта. Результаты его исследований опубликованы в ведущих российских и международных журналах: «IEEE Access», «Автоматика и телемеханика», «Экономика и математические методы» и других. В.Н. Лившиц является лауреатом премии РАН по экономике им. академика В.С. Немчинова, лауреатом конкурса Международного научного фонда экономических исследований академика Н.П. Федоренко.

В течение многих лет Вениамин Наумович ведет большую преподавательскую работу в различных вузах, в том числе в МФТИ ГУ, ВШЭ и Государственном университете «Дубна». Под его руководством защищено более 60 кандидатских и 10 докторских диссертаций. В 2001 году В.Н. Лившиц награжден серебряной медалью им. академика Н.Н. Моисеева за заслуги в сфере образования.

Он является членом научных, научно-технических и специализированных советов, а также многих международных научных организаций.

В настоящее время В.Н. Лившиц уделяет пристальное внимание формированию новой парадигмы государственного управления экономикой России. Им разработан механизм перехода к новой системной парадигме социально-экономического развития страны, предусматривающей принципиальные изменения в ключевых областях государственной социально-экономической политики – кадровой, инвестиционной, финансовой и макроэкономической.

Профессионализм ученого, высокий уровень работоспособности, творческое отношение к делу, чуткость и внимательное отношение к людям снискали любовь и уважение коллег, друзей, ученых и специалистов к Вениамину Наумовичу Лившиц в нашей стране и за рубежом.

Том 73, выпуск 1, 2023 г.

Содержание

Vol. 73, Issue 1, 2023

PROCEEDINGS OF THE INSTITUTE FOR SYSTEMS ANALYSIS OF RUSSIAN ACADEMY OF SCIENCES

IN THIS ISSUE

Информационные технологии

Classes of objects and relations in the Common Digital Space of Scientific Knowledge*

N.E. Kalenov, I.N. Sobolevskaya, A.N. Sotnikov

Joint Supercomputer Center of the Russian Academy of Sciences — Branch of Federal State Institution “Scientific Research Institute for System Analysis of the Russian Academy of Sciences” (JSCC RAS — Branch of SRISA), Moscow, Russia

Abstract. All over the world there are many both global and local information systems focused on solving various problems. As an integrator that allows you to solve complex information problems at the intersection of sciences and application areas of existing information systems. to the maximum extent using the information resources accumulated in them, the Common Digital Space of Scientific Knowledge (CDSSK) can be considered. The article provides the structure of the CDSSK, the requirements for its functionality and the structure of the software shell, corresponding to the principles of the Semantic WEB. All objects reflected in the CDSSK are divided into two classes − universal and local. Relationships between objects are also divided into two groups − universal and specific. The paper proposes a list of universal classes of objects, defines universal types of relations between them, gives examples of specific relations and approaches to identifying local classes and subclasses of objects in a particular field of science.

Keywords: *common Digital Space of Scientific Knowledge, Semantic WEB, codification, objects classes, object reference, object relations, entity reference, relationships between objects, ontology.*

DOI: 10.14357/20790279230101

Introduction

The Common Digital Space for Scientific Knowledge (CDSSK) is being formed with the aim of supporting and developing services in the field of science and education in the modern digital environment. [1 -4] The CDSSK includes heterogeneous information objects tested by the world scientific community. All over the world and in Russia, in particular, there are many both global and local information systems focused on solving various problems. In this regard, CDSSK should be considered as an integrator that allows you to solve complex information problems at the intersection of sciences and areas of application of existing information systems. maximum use of information resources accumulated in them. In particular, if we talk about Russia, then there are and are developing a number of state information systems in the digital environment, for example, the Russian Encyclopedia [5], the National Electronic Library [6], the Common State Register of Legal Entities [7], State catalog of geographical names [8], Russian Science Citation Index [9] (Fig. 1).

* The research is carried out by Joint SuperComputer Center of the Russian Academy of Sciences – Branch of Federal State Institution “Scientific Research Institute for System Analysis of the Russian Academy of Sciences” within the framework of a state assignment topic FNEF-2023-0014.

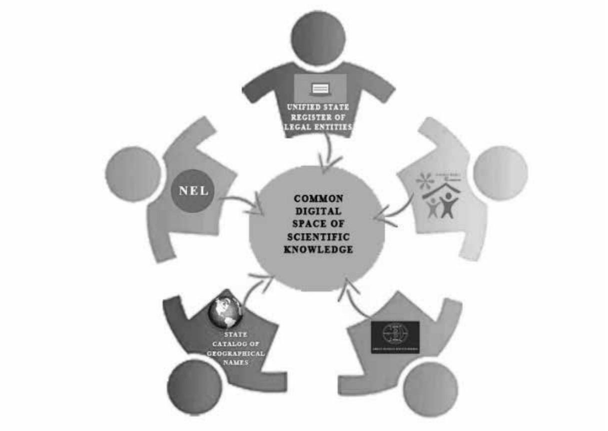

UNIFIED STATE REGISTER OF LEGAL ENTITIES
NEL
COMMON DIGITAL SPACE OF SCIENTIFIC KNOWLEDGE
STATE CATALOG OF GEOGRAPHICAL NAMES

Fig. 1. CDSSK is an integrator for scientific purposes of the state information systems

1. The common digital space of scientific knowledge object classes

Formation of CDSSK as a semantic WEB-space includes the following tasks:

- selection and structuring of scientific objects presented in existing information systems and containing reliable and comprehensive information about scientific achievements in various fields of knowledge;
- metadata profiles formation of objects presented in information systems;
- link start-up and registration of various kinds connections between dissimilar objects;
- formation of RDF triplets within the chosen field of knowledge.

Using OWL and RDF objects representations, their properties and relationships, and SPARQL-based data manipulation tools, you can build an information system containing multifaceted scientific information, backed by citations from reliable, time-tested, information-based sources. systems that are constantly being updated [10-12].

2. The common digital space of scientific knowledge general ontology construction

When constructing CDSSK common ontology, it is necessary to implement the following steps:

1. Allocation of universal classes of objects.

Currently, these include:

- Persons;
- Groups of persons (people united by a certain criterion, for example: “high school students”, “students studying in a given specialty”, “geologists”, “residents of a given country or city”, etc.);
- Publications. This object contains subclasses such as: monographs, collections, serials, etc .;
- Qualification works (dissertations and abstracts, copyright certificates);
- The documents. This class contains physical units such as a specific book, handwritten materials, archival documents;
- Museum items. In particular, rare editions should be treated as museum pieces with appropriate relations;
- Events;
- Location (geographic characteristics);
- Time characteristics;
- Organizations;
- Scientific directions;
- Thesauri (subject ontologies);
- General laws of nature from all scientific fields (the law of universal gravitation, the three laws of Newtonian mechanics, the laws of Lomonosov, the laws of thermodynamics, Zipf’s law, etc.).

2. Development of metadata profiles of objects of each class (subclass), including:

a. Formation of a list of metadata elements;
b. Define the characteristics and acceptable values of metadata, including:
   - type of data – text, number (range of numbers), date (dates range), link to another object, e-mail address, URL of an external object;
   - mandatory or optional;
   - unique or repetitive;
   - presentation format;
   - selection from permissible values linear table (single or multiple);
   - choice from a hierarchical structure (single or multiple);
   - free value (in accordance with the established view) with possible formal control within the view (text – according to dictionaries, number – for the validity of characters, date – for the established format, link – for checking the existing id, URL – for structure and accessibility).

3. Relations type establishing between objects. I.e., normalized tables formation of relations values between objects inside and outside each class;
4. Development of the data warehouse structure and the relations organization between them;
5. Development of a customizable administrator interface that implements the attributes list formation, tuning tables, data types selection, control type;
6. Development of an operator interface for entering metadata objects;
7. Development of the system internal organization (the formation of the name space of objects and relations (URN), the space of identifiers (URI), the formation of RDF triplets).

When defining metadata profiles for universal classes objects, it makes sense to focus on semantic relations with international and domestic organizations (ResearchGate, ORCID, RSCI – for persons, RAR and USRLE for organizations, RSCI and WEB of Science for publications, SCGN for geographic objects) [13, 14].

3. Universal relations types

1. Equivalence. For persons it means different spellings of surnames and first names;

- for publications – translated versions of one publication, reprints of books (stereotyped);
- for organizations – different names of the same organization. For example, Moscow State University – Lomonosov Moscow State University – Moscow University, etc.;
- for temporal characteristics;
- for geographical names. For example, RF Russian Federation – Russia.

Each object has required attributes:

- a unique identifier;
- name and relations of a given type with other objects.

Synonyms in subject ontologies. It is possible to select one of the equivalent objects as the base (descriptor), while the rest of the equivalent objects have enough three attributes – id, the name and the relationship of the type "equivalent" with the base.

2. "To be part of" (subordination of terms in subject ontologies). For organizations it means subordination of subdivisions;

- for publications: article – journal, collection;
- for geographical objects: country – continent, city – country, street – city, sea – ocean;
- for museum items: object – collection;
- for archives: document – inventory.

3. Contains (the prototype of "To be part of").
4. Intersects (subject ontologies are intersection of classification indices, country and natural zones, desert on the territory of several countries; rivers and countries; international organizations and countries, etc.).

4. Specific relations

Specific relations exist in both generic and local classes [15].

Examples of specific connections.

"Publication" – "Person":

- author;
- editor;
- compiler;
- interpreter;
- painter;
- sponsor;
- contains information about a person (in the library terminology "about him");
- reviewer;
- other roles (technical editor, proofreader, etc.)

"Publication" – "Organization":

- author (collective author – in library terms);
- publishing house;
- contains information about the organization;
- sponsor.

"Qualification work" – "Person":

- author;
- scientific adviser;
- opponent / reviewer.

"Qualification work" – "Organization":

- place of work performance;
- leading organization

"Document" – "Person";

- author;
- owner;
- contains information about the person;
- the document contains notes of this person.

"Document" – "Organization":

- author;
- owner + location specification (for archival documents – the number of the document, case and inventory, for the library the storage code, for the museum – the inventory number).
- mentioned in the document
- sponsor.

"Museum Item" – "Person":

- author – manufacturer;
- collection author (for natural science collections)
- source of income (donor / seller)
- restorer
- is associated with this person (photography, film-video, audio recording, etc.).

"Museum item" – "organization":

- author – manufacturer;
- owner (+ clarification of location – inventory number)
- source of income (donor / seller)
- is associated with this organization (photography, film-video, audio recording, etc.).

5. The common digital space of scientific knowledge subspace structure

The factographic basis of each thematic CDSSK subspace (its subject ontology) is structured encyclopedic concepts related to each other and to objects of universal classes. The structure of subject ontology can be based on the sections of existing heading lists of scientific information. For example: UDC (for general scientific) [16], INIS (for nuclear physics) [17], etc.

Below is an example of structuring the subject ontology of the thematic subspace "Astronomy" on the State Rubricator of Scientific and Technical Information [18] basis.

ASTRONOMY
General problems of astronomy
Theoretical astronomy. Celestial mechanics
Astrometry
Astrophysics
Solar system
Sun
Stars
Nebulae. Interstellar medium
Star systems
Cosmology

Observatory. Instruments, devices and methods of astronomical.

Each of the 11 sections highlighted at the second level of the hierarchy is subdivided into subsections of the third level. In particular, the following subsections are highlighted in the "Solar system" section:

Solar system
General problems of solar system research
Structure and origin of the solar system
Planets and their satellites
Moon. Lunar eclipses
Comets
Meteors. Zodiacal light. Interplanetary environment.
Meteorites

Within each subsection of the third level, subsections of the next level or individual objects are allocated. Each section (subsection) of a subject ontology is an object of the CDSSK. For each object, universal and specific connections are established with other objects of this subspace, other subspaces and with objects of universal classes. For astronomical objects, these can be connections with persons of the form "discovered", "described", "calculated"; with publications -relations of the form "first published", "textbook for school", "the most complete monograph", etc .; with objects from the subspace "Mathematics" – connections of the type "described by equations", etc. Objects of the subclass "Astronomical observatories" included in the last section of the second level of the subject ontology of the software program "Astronomy" are connected with objects of the "Location" class by the obligatory link "located in", etc.

Location (geographic objects) as a universal class contains general information about an object, not detailed from the point of view of geography, but allowing to determine the location with varying accuracy (from the mainland to the house number and coordinates with an accuracy of seconds). The purpose of distinguishing this universal class is to process generalized queries such as "archaeological excavations in Peru," or "herbaria collected in Altai," or "astronomical observations carried out in Chile," and so on. despite the fact that the description of the object of archaeological finds may indicate "Machu Picchu" or "Easter Island", when describing the herbarium, the surroundings of Biysk were indicated, and in astronomical observations, the Atacama Desert was indicated.

Objects of the universal class "location" are associated with elements of the thematic subspace "Geography", which contains comprehensive descriptions of geographic objects. The location class includes the subclasses Land and Water, which in turn include the following subclasses.

Land.

- continent
- part of the world;
- natural area;
- part of the land that has a geographical name;
- country
- subject of the country
- locality (city, town, village)
- the named part of the settlement (district, street, square, etc.)
- address
- coordinates

Water space.

- oceans;
- seas;
- lakes;
- rivers;
- other bodies of water that have a name (waterfalls, swamps ...).

Along with universal connections, specific connections of the type "washed" (connection between a continent or country and the sea or ocean), "is an inflow" (connection between rivers), "stands on" (connection between a city and river), etc.

Conclusions

When software development for a particular scientific direction, it is necessary to move along the path of identifying classes and subclasses of objects, forming objects metadata profiles of each subclass, establishing relations between objects of this class, within this software and with objects of universal classes. The result of the design should be a set of RDF triplets, which will allow implementing mechanisms for finding answers to complex queries based on the SPARQL language.

References

1. *Antopolskij A.B., Kalenov N.E., Serebryakov V.A., Sotnikov A.N.* O edinom cifrovom prostranstve

nauchnyh znanij // Vestnik Rossijskoj akademii nauk, 2019. – T. 89, – № 7. – S. 728-735. DOI 10.31857/S0869-5873897728-735

2. *Savin G.I.* Edinoe cifrovoe prostranstvo nauchnyh znanij: celi i zadachi // Informacionnye resursy Rossii, 2020. – № 5. – S. 3-5. DOI: 10.51218/0204-3653-2020-5-3-5
3. *Nikolay Kalenov, Gennadiy Savin, Alexander Sotnikov.* Fundamentals of Common Digital Space of Scientific Knowledge Building // CEUR Workshop Proceedings (CEUR-WS.org) , 2021. – Vol. 2990. – P. 93-99. DOI: 10.51218/1613-0073-2990-93-99
4. *Olga Ataeva, Nikolay Kalenov, Vladimir Serebryakov, Alexander Sotnikov.* Informational Infrastructure of the Common Digital Space of Scientific Knowledge // CEUR Workshop Proceedings (CEUR-WS.org) , 2021. – Vol. 2990. – P. 1-10. DOI: 10.51218/1613-0073-2990-1-10
5. https://bigenc.ru/ (the last access 12.2021)
6. https://rusneb.ru/ (the last access 12.2021)
7. https://fedresurs.ru/?attempt=1 (the last access 12.2021)
8. https://cgkipd.ru/science/names/reestry-gkgn.php (the last access 12.2021)
9. https://elibrary.ru/project_risc.asp? (the last access 12.2021)
10. *Millar D., Braines D., D'Arcy L., Barclay I., Summers-Stay D., Cripps P.* Embedding Dynamic Knowledge Graphs based on Observational Ontologies in Semantic Vector Spaces // Artificial intelligence and machine learning for multi-domain operations applications III. Vol. 11746., № 117461O. (2021).
11. *Wang Q., Ji YD., Hao YS., Cao J.* GRL: Knowledge graph completion with GAN-based reinforcement learning // Knowledge-based systems. Vol. 209., № статьи 106421. (2020).
12. *Hansen C., Hotz I., Ynnerman A.* Visualization in Public Spaces // Ieee computer graphics and applications. 40 (2). pp. 16-17. (2020).
13. *Piplai A., Ranade P., Kotal A., Mittal S., Narayanan SN., Joshi A.* Using Knowledge Graphs and Reinforcement Learning for Malware Analysis // 2020 IEEE international conference on big data (big data). pp. 2626-2633. (2020).
14. *Dessi D., Osborne F., Recupero DR., Buscaldi D., Motta E.* Generating knowledge graphs by employing Natural Language Processing and Machine Learning techniques within the scholarly domain // Future generation computer systems-the international journal of escience. Vol. 116. pp. 253-264. (2021).
15. *Nikolay Kalenov, Irina Sobolevskaya, Alexander Sotnikov.* Hierarchical Representation of Information Objects in a Digital Library Environment // Communications in Computer and Information Science. Vol. 1093. pp. 93-104 (2019).
16. UDC: https://udcc.org/index.php/site/page?view=factsheet (the last access 12.2021)
17. INIS: https://www.iaea.org/sites/default/files/19/09/en-2019-09.pdf (the last access 12.2021)
18. SRSTI: https://grnti.ru (the last access 12.2021)

Kelenov N.E. DSc. Joint Supercomputer Center of the Russian Academy of Sciences — Branch of Federal State Institution "Scientific Research Institute for System Analysis of the Russian Academy of Sciences", 119334, Moscow, Leninsky av., 32 a, Russia, e-mail: nkalenov@jscc.ru

Sobolevskaya I.N. PHd. Joint Supercomputer Center of the Russian Academy of Sciences — Branch of Federal State Institution "Scientific Research Institute for System Analysis of the Russian Academy of Sciences", 119334, Moscow, Leninsky av., 32 a, Russia, e-mail: ins@jscc.ru (correspondent author)

Sotnikov A.N. DSc. Joint Supercomputer Center of the Russian Academy of Sciences — Branch of Federal State Institution "Scientific Research Institute for System Analysis of the Russian Academy of Sciences", 119334, Moscow, Leninsky av., 32 a, Russia, e-mail: ASotnikov@jscc.ru

Curation of bibliographic metadata of the institutional repository on the Invenio-JOIN^2 platform

I.A. Filozova, T.N. Zaikina, G.V. Shestakova, R.N. Semenov

Joint Institute for Nuclear Research, Dubna, Moscow Region, Russia

Abstract. Content filling of the institutional repository and keeping the entered data “up to date” is a very resource-intensive task that requires organizing the coordinated actions of operators to enter data into an information system (IS). To resolve one helps the curation of bibliographic metadata — a set of actions and measures aimed for updating, managing and preserving digital objects throughout their life cycle in educational and the scientific interests of the community. This work considers the issues of bibliographic descriptions curation of publications by JINR (Joint Institute for Nuclear Research) employees, their enrichment of metadata entered into the JINR institutional repository from external sources: the Scopus bibliographic and abstract database, the Web of Science search Internet platform, the information platform in High Energy Physics INSPIREHEP. The development of information services for solving the problem of current accounting of the publication activity of JINR staff is described.

Keywords: *Open Access (OA), Institutional Repository (IR), CRIS & OAR, software platform Invenio-JOIN^2, Utility, Application Programming Interface (API), Authority Record, Publication Activity.*

DOI: 10.14357/20790279230102

Introduction

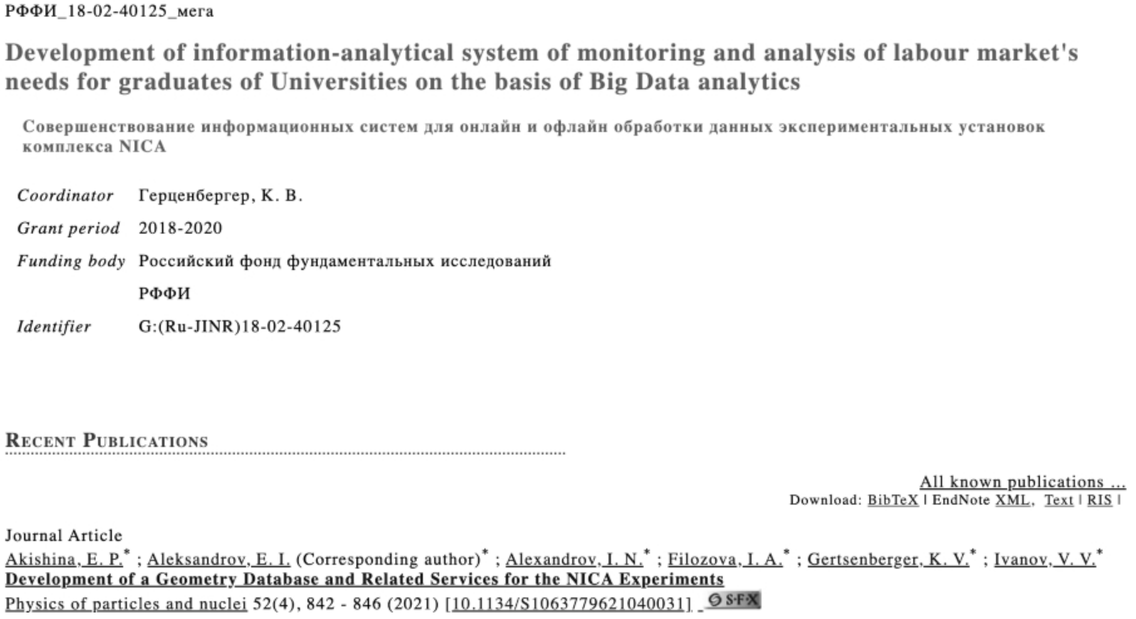
Keeping the content of the institutional repository up to date is an important task that requires significant time and human resources. The quality of the content of any information system is one of the key factors that makes it attractive to end users and able to meet their changing information needs. The way to organize such laborious work is to ensure the process of continuous curation of bibliographic metadata (the main digital objects of such kind information systems) and include a set of actions and measures aimed at updating, managing and preserving this metadata throughout their entire life cycle. The tasks of curation include: search and analysis of missing bibliographic metadata (not included in the repository at the current time for some reason); search and analysis of missing metadata of bibliographic descriptions uploaded to the repository; tracking changes of the publication’s status (made earlier preprint can be transformed into an article published in a peer-reviewed journal); detection of input errors; updating user account data; updating data on the structural divisions of the organization, etc.

Currently, JINR is implementing an institutional repository based on the Invenio-JOIN^2 software platform [1; 2]. Closed beta testing is underway now.

This paper considers the issues of identifying missing bibliographic descriptions of JINR staff publications, adding bibliographic descriptions of reference books (grants, experiments, persons); enrichment of metadata (their clarifications and additions) entered into the JINR institutional repository with data from external sources: the bibliographic and abstract database Scopus, the information platform in the field of high energy physics INSPIREHEP [3]. These tasks can be partially automated. As a solution, a set of utilities and services has been developed to facilitate the implementation of some specialized standard curation processes for the JINR Publications Server.

1. Approaches and Practices for the Curation of Institutional Repositories

Most literature on the research topic presents the implementation of IRs from the viewpoint of end users. Thus the description is restricted by the front-end component of such IS, that is only “iceberg tip” and does not give a complete picture of the complexity of such information systems. Detailed structured descriptions of activ-



РФФИ_18-02-40125_мега

Development of information-analytical system of monitoring and analysis of labour market's needs for graduates of Universities on the basis of Big Data analytics

Совершенствование информационных систем для онлайн и офлайн обработки данных экспериментальных установок комплекса NICA

Coordinator	Герценбергер, К. В.
Grant period	2018-2020
Funding body	Российский фонд фундаментальных исследований РФФИ
Identifier	G:(Ru-JINR)18-02-40125

Recent Publications

All known publications ...
Download: BibTeX | EndNote XML, Text | RIS |

Journal Article
Akishina, E. P.* ; Aleksandrov, E. I. (Corresponding author)* ; Alexandrov, I. N.* ; Filozova, I. A.* ; Gertsenberger, K. V.* ; Ivanov, V. V.*
Development of a Geometry Database and Related Services for the NICA Experiments
Physics of particles and nuclei 52(4), 842 - 846 (2021) [10.1134/S1063779621040031]

Fig. 1. Authority record of grant "РФФИ 18-02-40125_мега"

ities related to the management of research data, the needs and actions of various categories of users, used tools are a good basis for developing best practice guidelines and infrastructure templates for IRs [4]. These knowledge tools can then be used by institutions that are currently implementing institutional repositories. In many respects, the approaches and practices of IRs content curation depend on their specifics and the specifics of the institution that holds the content [5]. The resources management of an institutional repository includes the task of curation [6; 7]. The JOIN² project pays a lot of attention to the issue of curation. All instances provide updating authority records Persons and Institutes in automatic or semiautomatic mode [2]. One more example: curation process to enrich data in JOIN² repository with Metadata from PubMed. The script produces several output-files that are uploaded in batch and make updating records [2]. The implementation of curation processes depends on the set of protocols and authentication tools used by the institution. The project ISTINA [8] uses back-end algorithms to find sustainable teams of authors, research performance evaluation, authorship disambiguation and so on. The paper [4] describes data curation and use activities in IRs, their structures, roles played, skills needed, contradictions and problems present, solutions sought, and applied approaches.

2. Utilities and Services for JDS-JOIN² Prototype

To solve the above tasks, the following set of useful utilities has been developed: New Grants, People&Institutes Corrections, Inspire Corrections, HAC List. Services are implemented as a backend solution for the JINR institutional repository and are executed as server processes with a given frequency — tasklets. Publications Server is carried out as an Open Access archive of JINR scientific output [9].

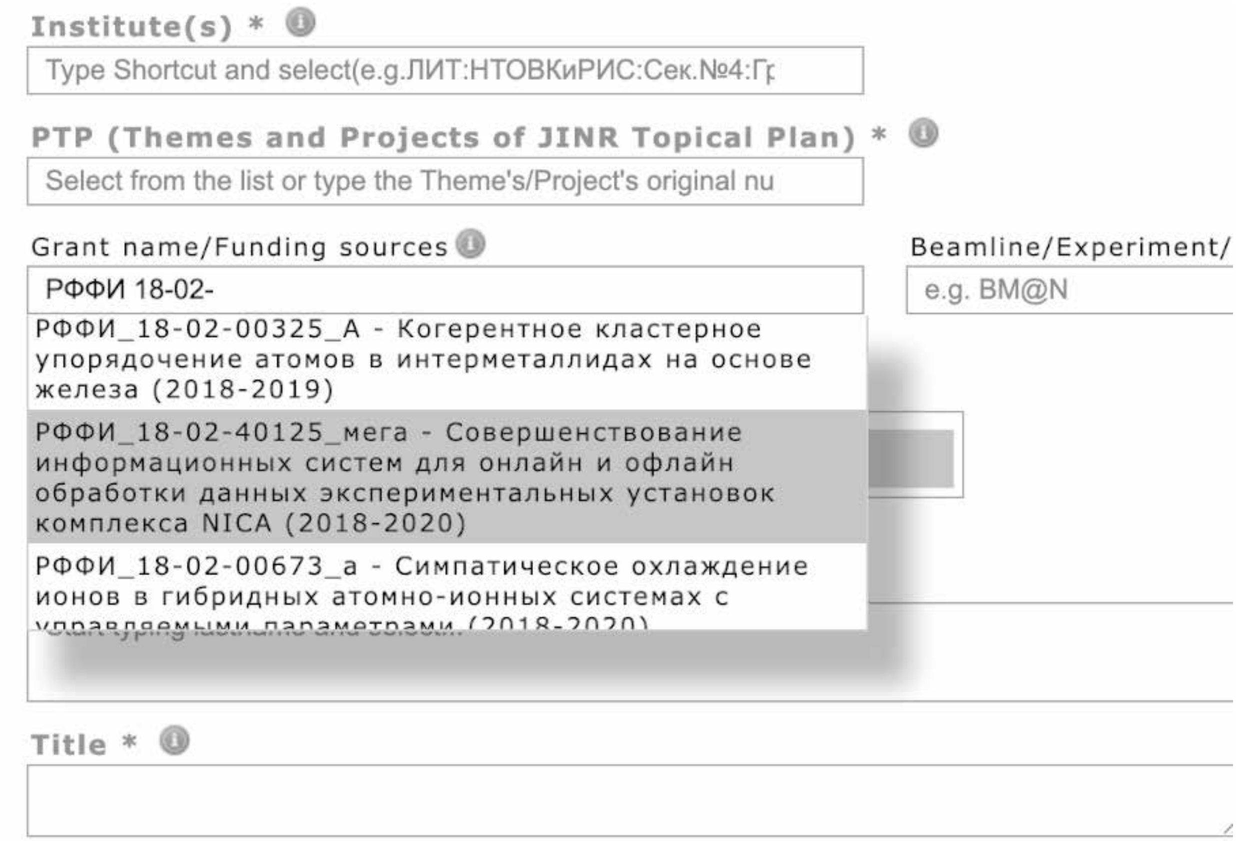
2.1. Utility New Grants

To reflect the information about grant support of the research work, resulting or describing in a publication, it is necessary to add the appropriate metadata to bibliographic description of this publication. Grant metadata is stored in the Authority Collection Grants (fig. 1) – catalog of research funding sources (JINR Themes/Projects, RFBR and RSF grants).

The Authority record of the Grant "РФФИ 18-02-40125_мега" contains the title, coordinator name, grant period, as well as a list of publications loaded into the system and linked (logical connection of the publication belonging to this funding source) to it. To specify the relationship between publication and Grant, the submitter should enter the necessary metadata in publication through the web-Submit tool (fig. 2).

The input subsystem, using the import module by the values of digital identifiers (in this case, DOI), recognizes public bibliographic metadata and distributes them among the form fields. In the Grant name/ Funding sources section, the user selects the desired grant from the drop-down list (it must first be loaded into the system).

Information about JINR projects, funded by the Russian Foundation for Basic Research and the



Institute(s) *

Type Shortcut and select(e.g.ЛИТ:НТОВКиРИС:Сек.№4:Гр

PTP (Themes and Projects of JINR Topical Plan) *

Select from the list or type the Theme's/Project's original nu

Grant name/Funding sources

РФФИ 18-02-

РФФИ_18-02-00325_А - Когерентное кластерное упорядочение атомов в интерметаллидах на основе железа (2018-2019)

РФФИ_18-02-40125_мега - Совершенствование информационных систем для онлайн и офлайн обработки данных экспериментальных установок комплекса NICA (2018-2020)

РФФИ_18-02-00673_а - Симпатическое охлаждение ионов в гибридных атомно-ионных системах с управляемыми параметрами (2018-2020)

Beamline/Experiment/

e.g. BM@N

Title *

Fig. 2. Web-form for entering a publication

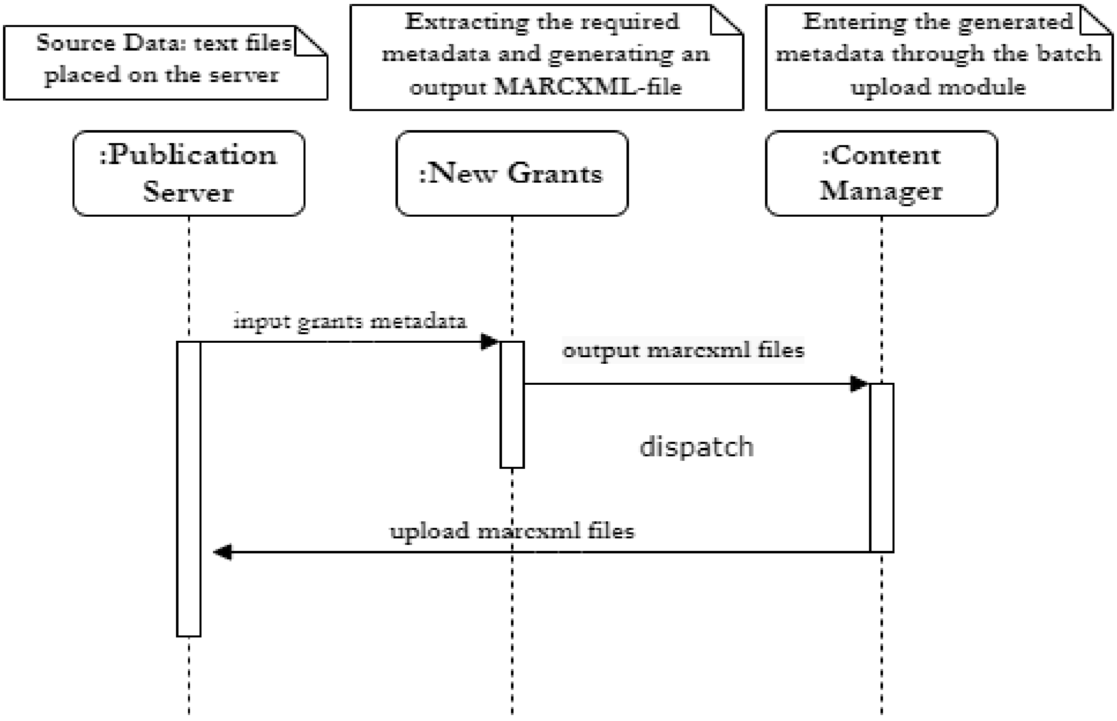

Source Data: text files placed on the server
Extracting the required metadata and generating an output MARCXML-file
Entering the generated metadata through the batch upload module
:Publication Server
:New Grants
:Content Manager
input grants metadata
output marcxml files
dispatch
upload marcxml files

Fig. 3. Interaction of the server, module New Grants and content manager of institutional repository in the scenario of generating and loading metadata about new grants

Listing 1. Example Authority Record Grant РНФ 19-75-20121

```
<record>
<datafield ind1=”7” ind2=” ‟ tag=”024”>
<subfield code=”a”>G:(Ru-JINR)19-75-20121</subfield>
<subfield code=”0”>I:(Ru-JINR)_400000</subfield>
<subfield code=”d”>19-75-20121</subfield>
</datafield>
<datafield ind1=” ‟ ind2=” ‟ tag=”035”>
<subfield code=”a”>G:(Ru-JINR)19-75-20121</subfield>
</datafield>
<datafield ind1=” ‟ ind2=” ‟ tag=”150”>
<subfield code=”a”>Новые гибридные и углеродные аэрогели – синтез и анализ структуры методами малоуглового рассеяния (на ИБР-2)</subfield>
<subfield code=”y”>2019-2022</subfield>
</datafield>
<datafield ind1=” ‟ ind2=” ‟ tag=”371”>
<subfield code=”a”>Горшкова Ю.Е.</subfield>
<subfield code=”0”>P:(Ru-JINR)P002369</subfield>
</datafield>
<datafield ind1=” ‟ ind2=” ‟ tag=”450”>
<subfield code=”a”>РНФ 19-75-20121 Проведение исследований на базе существующей научной инфраструктуры мирового уровня</subfield>
<subfield code=”y”>2019-2022</subfield>
<subfield code=”w”>d</subfield>
</datafield>
<datafield ind1=”1” ind2=” ‟ tag=”510”>
<subfield code=”a”>Российский научный фонд</subfield>
<subfield code=”0”>I:(DE-588b)1228775-1</subfield>
<subfield code=”b”>РНФ</subfield>
</datafield>
<datafield ind1=” ‟ ind2=” ‟ tag=”980”>
<subfield code=”a”>G</subfield>
</datafield>
<datafield ind1=” ‟ ind2=” ‟ tag=”980”>
<subfield code=”a”>AUTHORITY</subfield>
</datafield>
</record>
```

Russian Science Foundation is the source of data for the New Grants module. This information is presented on the official website of JINR (http://www.jinr.ru/grants/) [11]. The New Grants module receives a text file with data about grants as input, and generates an output file in MARCXML format, which is loaded into the system in batch loading mode by the content manager of the JINR publications server (fig. 3).

This utility implements the following functions:

- Extract grant metadata from source file.
- Checking for duplicates (whether there is an entry with an identical number in the Grants authority records collection).
- Generation of a marcxml file with a description of the grant for subsequent loading into the system (see Listing 1. — an example of the description of the Grant authority record for the RSF project 19-75-20121 in marcxml format).

Using the New Grants utility, 270 records of RFBR grants for the period 2013-2022 and 60 entries about RSF projects for the period 2014-2022 were uploaded to the JINR Publications Server.

2.2. Service People&Institutes Corrections

This application updates the authority records of People and Institutes – collections – directories of JINR employees (potential authors of scientific publications) and JINR structural divisions of all levels of the hierarchy, respectively (fig. 4).

The main functions of *People&Institutes Corrections*:

- Updating data about working employees: modifying email, adding SSO login, adding identifiers of various external author identification systems like Inspire, ORCID, Scopus (if available).
- Deleting authority records of non-working (dismissed) employees.
- Making authority records of new (hired) employees.

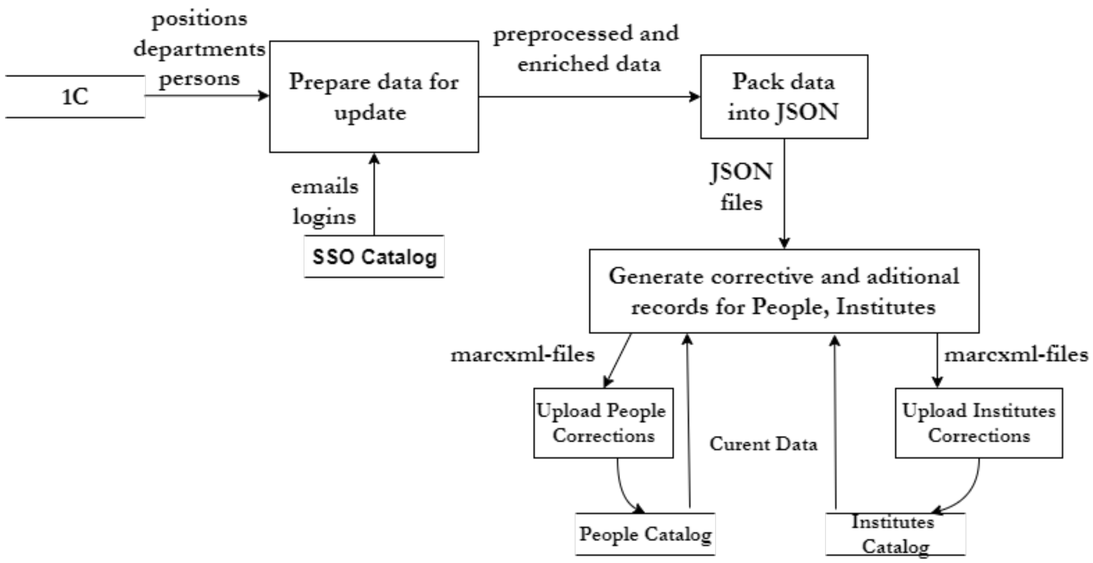


Fig. 4. Update People and Institutes Catalogs

- Updating data when an employee moves from one department to another.
- Updating data on current structural divisions (renaming, transferring to another level of hierarchy, etc.).
- Deleting authority records of liquidated structural divisions.
- Entering authority records of new (newly created) structural units.

Deletion of the record happens after attaching a specific tag, the record becomes inactive (invisible to the user) but is not destroyed from the database.

This service is currently running weekly. On average, the volume of corrective entries is: 25-35 updates (including dismissed and hired people) per week, 3-4 updates for departments per month. When the institutional repository is set to production mode, the service will be launched daily.

2.3. Service Inspire Corrections

INSPIRE (https://inspirehep.net/) is the trusted information hub for the high energy physics community [3]. It serves as a one-stop information platform for the HEP community, including 8 interconnected databases of literature, conferences, institutions, journals, researchers, experiments, jobs, and data. INSPIREHEP works in collaboration with CERN, DESY, Fermilab, IHEP, IN2P3 and SLAC and has been serving the scientific community for nearly 50 years.

Experience has shown that reports on publication activity often lack publications for a given reporting period from subject repositories such as INSPIREHEP. This information platform has collected more than 35 thousand publications of JINR employees for the period from 1956 to the present (fig. 5).

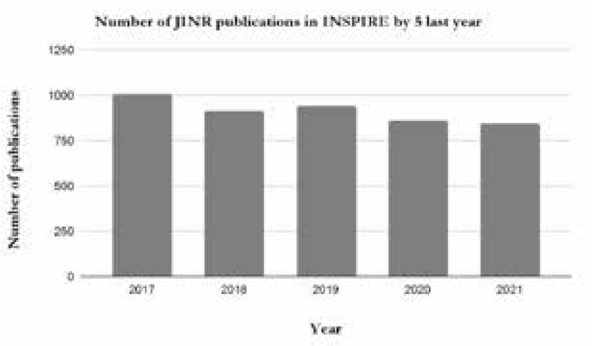
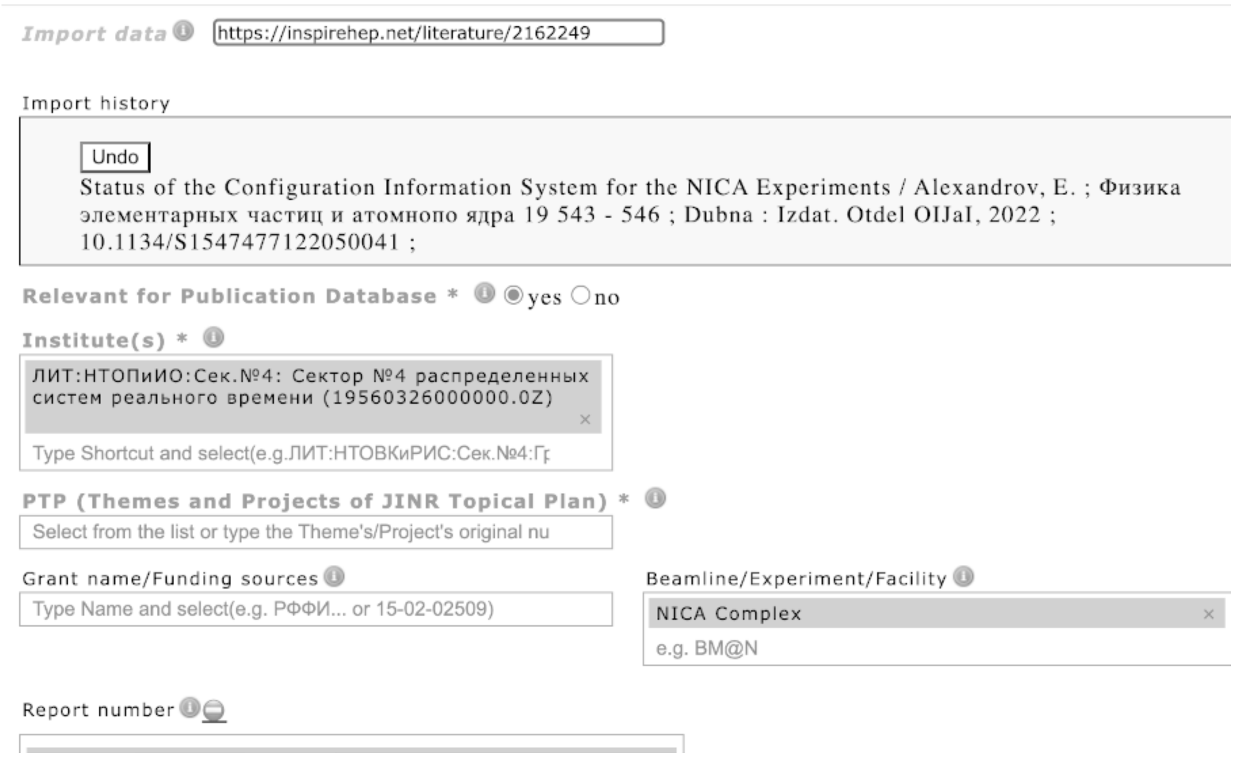


Fig. 5. JINR Publications in INSPIREHEP (2017-2021)

Over the past five years, the number of publications of JINR staff members stored in INSPIREHEP has been at the level of 850 ÷ 1000 per year.

To identify missing (missing descriptions in the institutional repository) INSPIREHEP publications, the *Inspire Corrections* service was developed, which performs the following functions:

- Retrieve bibliographic metadata from the INSPIREHEP hub using the public API (https://github.com/inspirehep/rest-api-doc) for a specified period, as an example: for the last month (configurable options).
- Duplicate Check: Checks for the publication metadata retrieved from INSPIREHEP and stored at the JINR Publication Server.
- Sending an email notification to the co-author (or responsible person performing archiving by proxy) about unloaded INSPIREHEP publications with references and INSPIRE's IDs. Using this information, an employee can easily submit these publications to the JINR Publication Server by using the *Import Data* function of the web-submit module (fig. 6).



Import data https://inspirehep.net/literature/2162249

Import history

Undo

Status of the Configuration Information System for the NICA Experiments / Alexandrov, E. ; Физика элементарных частиц и атомнопо ядра 19 543 - 546 ; Dubna : Izdat. Otdel OIJaI, 2022 ; 10.1134/S1547477122050041 ;

Relevant for Publication Database * ● yes ○ no

Institute(s) *

ЛИТ:НТОПиИО:Сек.№4: Сектор №4 распределенных систем реального времени (19560326000000.0Z) ×

Type Shortcut and select(e.g.ЛИТ:НТОВКиРИС:Сек.№4:Гр

PTP (Themes and Projects of JINR Topical Plan) *

Select from the list or type the Theme's/Project's original nu

Grant name/Funding sources

Type Name and select(e.g. РФФИ... or 15-02-02509)

Beamline/Experiment/Facility

NICA Complex ×

e.g. BM@N

Report number

Fig. 6. Web-form for new submissions

2.4. Utility HAC List

The experience of pre-production prototype of institutional repository "JINR Publications Server" showed that one of the frequent types of user queries is the search for publications in journals registered in various international and national catalogs and databases (Database Coverage). For example, the search of publications published in the Directory of Open Access Journals (DOAJ) indexed by Scopus, or journals included in the List of Higher Attestation Commission a list of leading peer-reviewed scientific journals included by the Higher Attestation Commission of the Russian Federation, recommended for publishing the main scientific results of a dissertation for the degree of candidate and doctor of science) [12].

The functionality of the Invenio-JOIN[2] software platform, on which the JINR Publications Server institutional repository prototype is deployed, allows the filtering of publications described above. To do this, the Statistics key – special Authority Record, should be created and associated with the necessary journal. Then all publications of this journal entered into the system will be indexed as belonging to this Statistics key, and can be found by the search attribute StatID = value. Authority Record with internal value **2002** was added to the Statistics keys catalog. The aim of it is to indicate that the article is/was published in a journal from the LIST of the peer-reviewed scientific publications, approved by the Higher Attestation Commission under the Ministry of Science and Higher Education of the Russian Federation.

The HAC List published at the official website in a pdf file (the latest up-to-date version – 04/27/2022 includes 2679 journal titles). The file contains a table with columns: *No.; Name of the journal; ISSN; Scientific specialties and branches of science corresponding to them, in which academic degrees are awarded; The date of inclusion of this journal in the List.* The content of the pdf-file in its original form is not suitable for machine processing. An application was implemented

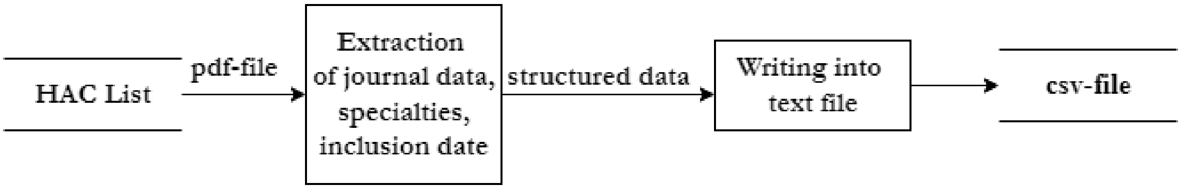


Fig. 7. Extraction of structured data from source HAC List

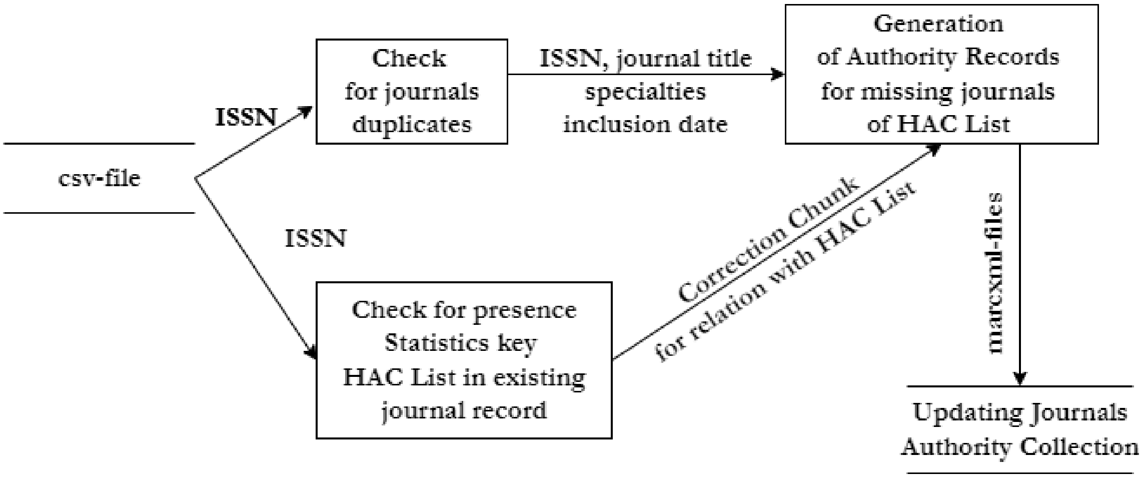

csv-file
ISSN
ISSN
Check for journals duplicates
ISSN, journal title specialties inclusion date
Generation of Authority Records for missing journals of HAC List
Check for presence Statistics key HAC List in existing journal record
Correction Chunk for relation with HAC List
marcxml-files
Updating Journals Authority Collection

Fig. 8. Indication journals records by statistics key HAC List

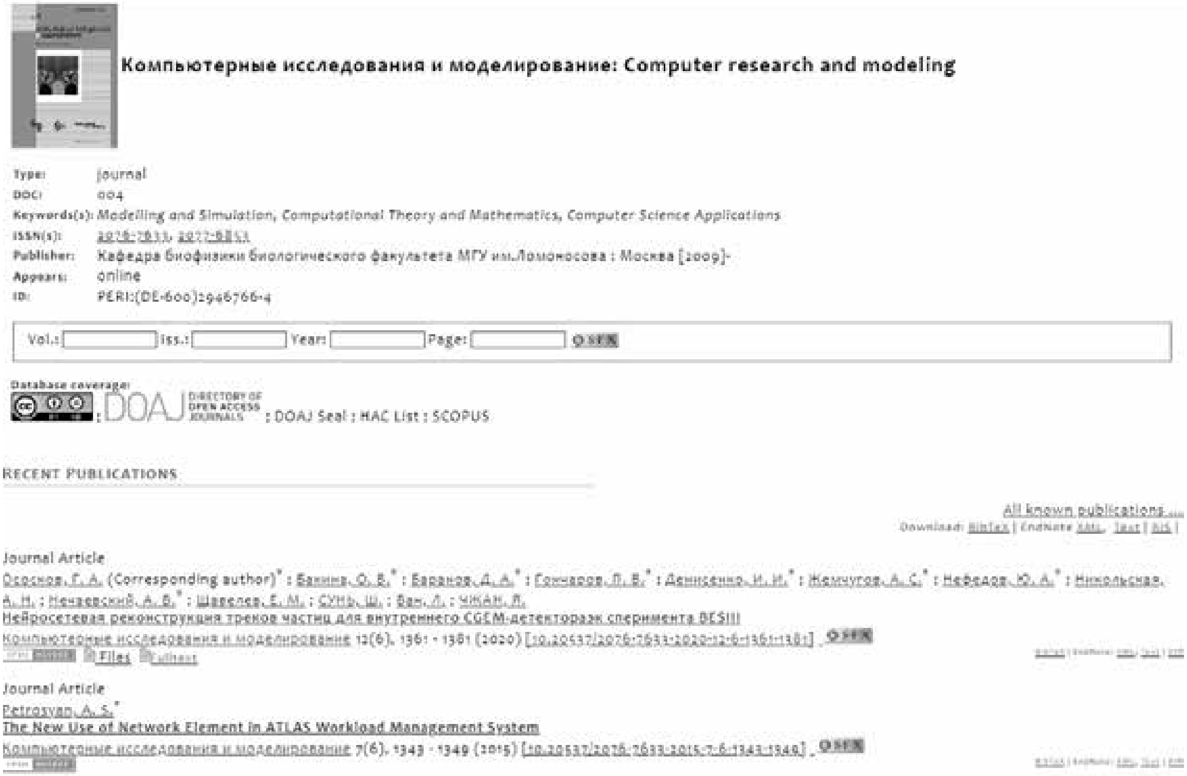
that extracts data from this pdf-file and structures the data according to the above columns, and generates a text csv file (fig. 7).

The resulting csv file is fed to the input of another application execute the following functions:

- Checking the journal for duplicates by ISSN: whether there is a record with an identical ISSN value in the Journals authority collection.
- Checking if an authority record has a *HAC List* Statistics key label, which means that the journal is identified in the system as belonging to the HAC List.
- Creating the corrective chunk with *HAC List* Statistics key in its absence**.**
- Generation of journals authority records metadata absened in a system, in marcxml format for uploading with Statistics key HAC List**.**

Fig. 8. illustrates it.

An example of this approach is presented in fig. 9. The authority record of the journal *Computer research and modeling* displays a list of publications published in it. The Database Coverage block displays the *HAC List* label.


Компьютерные исследования и моделирование: Computer research and modeling

Type: journal
DOC: 004
Keywords(s): *Modelling and Simulation, Computational Theory and Mathematics, Computer Science Applications*
ISSN(s): 2076-7633, 2077-6853
Publisher: Кафедра биофизики биологического факультета МГУ им.Ломоносова : Москва [2009]-
Appears: online
ID: PERI:(DE-600)2946766-4

Vol.: Iss.: Year: Page:

Database coverage:
: DOAJ Seal : HAC List : SCOPUS

RECENT PUBLICATIONS

All known publications ...
Download: BibTeX | EndNote XML, Text | RIS |

Journal Article
Ососков, Г. А. (Corresponding author)* ; Бакина, О. В.* ; Баранов, Д. А.* ; Гончаров, П. В.* ; Денисенко, И. И.* ; Жемчугов, А. С.* ; Нефедов, Ю. А.* ; Никольская, А. Н. ; Нечаевский, А. В.* ; Щавелев, Е. М. ; СУНЬ, Ш. ; Ван, Л. ; ЧЖАН, Л.
Нейросетевая реконструкция треков частиц для внутреннего CGEM-детектора эксперимента BESIII
Компьютерные исследования и моделирование 12(6), 1361 - 1381 (2020) [10.20537/2076-7633-2020-12-6-1361-1381]
Files Fulltext

Journal Article
Petrosyan, A. S.*
The New Use of Network Element in ATLAS Workload Management System
Компьютерные исследования и моделирование 7(6), 1343 - 1349 (2015) [10.20537/2076-7633-2015-7-6-1343-1349]

Fig. 9. Authority record of journal *Computer research and modeling*

Conclusion

Updating the content of an institutional repository is a complex and multifaceted task that requires continuous curation by the accompanying staff. Curation is organized on the basis of typical business processes, some of which can be automated. Automation consists in the implementation of a set of specialized utilities (auxiliary scripts that complete the functionality of the software platform in order to perform typical curation tasks) and information services. The work shows the usage of this functionality on the JINR Publication Server as an example. In particular, the *New Grants* Utility, Service *People&Institutes Corrections,* Service *Inspire Corrections* have been implemented and included in the pre-production version of the system. The Utility *HAC List* is implemented with limited functionality and is still under development. The relevance to continue the development of *Scopus Corrections* Service with functionality similar to *Inspire Corrections* Service is discussed.

References

1. The official web-site Invenio project. Available at: https://invenio-software.org/ (accessed November 23, 2022).
2. The official web-site JOIN² Project. Available at: https://join2.de (accessed November 23, 2022).
3. Information platform for HEP community. Available at: https://inspirehep.net/ (accessed November 23, 2022).
4. *Lee DJ, B. Stvilia.* 2017. Practices of research data curation in institutional repositories: A qualitative view from repository staff. PLoS ONE 12(3): e0173987. Available at: https://doi.org/10.1371/journal.pone.0173987 (accessed November 23, 2022).
5. *Kidney A., C. R. Borges, D. Molodenskiy, et al.* 2020. SASBDB: Towards an automatically curated and validated repository for biological scattering data. Protein science 29(1): 66 – 75. doi:10.1002/pro.3731.
6. *Redkina N.S.* 2022. The libraries and Open Science: vectors of interaction. Scientific and technical libraries 3:105–126. doi:10.33186/1027-3689-2022-3-105-126. (In Russian).
7. *Redkina N.S.* 2019. Modern trends in research data management. Scientific and technical information. Series 1: Organization and methodology of information work 4:1–7. (In Russian).
8. *Afonin S.A., and others.* Ed. Academician V.A. Sadovnichiy. 2014. Intellectual system of thematic research of scientific and technical information. M.: Moscow University Press. 262p. (In Russian).
9. JINR Publications Server, Available at: https://publications.jinr.ru (accessed November 23, 2022).
10. *Filozova Irina,Tatiana Zaikina, Galina Shestakova, Roman Semenov, Martin Köhler, Alexander Wagner, Laura Baracci on behalf of the JOIN² project.* 2020. JINR Open Access Repository based on the JOIN² Platform. Proceedings of the Data Analytics and Management in Data Intensive Domains 2020. CEUR Workshop Proceedings, 2790:142-155. Available at: http://ceur-ws.org/Vol-2790/paper14.pdf (accessed November 23, 2022).
11. Materials of the JINR official website. Available at: http://jinr.ru (accessed November 23, 2022).
12. Official web-site of the Higher Attestation Commission under the Ministry of Science and Higher Education of the Russian Federation. Available at: https://vak.minobrnauki.gov.ru/ (accessed November 23, 2022).

I.A. Filozova. Joint Institute for Nuclear Research, 6 Joliot-Curie St, 141980 Dubna, Moscow Region, Russia, e-mail: fia@jinr.ru

T.N. Zaikina. Joint Institute for Nuclear Research, 6 Joliot-Curie St, 141980 Dubna, Moscow Region, Russia, e-mail: ztanya@jinr.ru (correspondent author)

G.V. Shestakova. Joint Institute for Nuclear Research, 6 Joliot-Curie St, 141980 Dubna, Moscow Region, Russia, e-mail: shestakova@jinr.ru

R.N. Semenov. Joint Institute for Nuclear Research, 6 Joliot-Curie St, 141980 Dubna, Moscow Region, Russia, e-mail: roman@jinr.ru

Development of an Information-analytical System for the Support and Maintenance of Licenses at MLIT JINR

E.I. Alexandrov, I.N. Alexandrov, D.V. Belyakov, N.A. Davydova, L.A. Kalmykova, M.A. Lyubimova, T.F. Sapozhnikova, T.S. Syresina, A.V. Yakovlev, P.V. Zrelov

Joint Institute for Nuclear Research, Dubna, Russia

Abstract. The License Management System (LMS) was developed at the JINR Information Technology Laboratory. The purpose of creating an LMS is to automate the management, acquisition, maintenance and use of licensed software products. The system consists of a network licensing system (NLS), databases and a web interface. NLS provides network license management, collects and transmits to the time series database information about which network license was used by the user and at what time. The features of collecting this type of data are given. This data is used in monitoring implemented on the basis of the Grafana platform. The main LMS database stores data related to corporate, private, and other types of licenses. It also stores the necessary data regarding license users. The database is implemented in PostgreSQL. The system provides the ability to process workflows such as ordering new licenses that users need, ordering to add to the catalog of purchased licenses, and other functions. The LMS web interface is implemented in the development environment of the Electronic Document Management System "EDMS Dubna" using the LegoToolkit web application. A website has been developed and created for LMS users.

Keywords: *license management system, database, web interface.*
DOI: 10.14357/20790279230103

Introduction

The Meshcheryakov Laboratory of Information Technologies (MLIT) of the Joint Institute for Nuclear Research (JINR) is actively working to streamline the licensed software used. There are various aspects of work related to optimizing the use of licenses. For example, choosing the optimal consistent set of licenses for complex open source projects [1]. In our work, we primarily consider the issues of optimizing the use of licensed products in terms of the cost of their acquisition, and for the same purposes, the collection and accounting of information on the use of licenses. The variety of types of purchased licenses creates a potential difficulty in considering and choosing the optimal licensing conditions for the specific needs of the Laboratory. There are many types of licenses, such as personal, local, pre-installed software, full package product, volume licensing, group, network, etc. To select the required types of licenses, it is necessary to have information on both existing licenses and the needs of users. Additional difficulties arise when solving the problem of collecting dynamic information on the use of network licenses. Licenses of this type are purchased for a certain number of users who have the right to simultaneously use the license. The cost of network licenses used in the Laboratory is quite high and varies greatly depending on the number of users it is purchased for. Therefore, detecting the maximum number of concurrent users is an important task for future license renewal. To solve this problem, it is necessary to collect data on the beginning and end of the use of the license by the user. The corresponding license usage and release requests are sent from different user machines, and license managers are third-party products from which it is not always possible to obtain license usage information. Based on this, and also taking into account the high costs of acquiring licensed software products, the obligation to comply with the rules of the licensing policy and the need to plan and optimize the purchase of licensed software, it was proposed to automate the processes of license management. It resulted in the development and most of the implementation of the LMS. The architecture of the system, the database object model schema, the Web Interface design and implementation are presented in this paper below.

1. Architecture of the LMS

The LMS consists of three parts: Network Licensing System (NLS), databases and Web Services.

The general architecture of the LMS is shown in fig. 1.

Important parts of the system are Web services and the associated database implemented in PostgreSQL. The database contains all information on purchased licenses, licensed software product (LSP), LSP owners, main and auxiliary catalogs, LSP development companies, vendors and other necessary information. Based on this information, personal accounts of the user, operator and auditor, the corresponding forms and documents are formed. The monitoring of Network Licenses Usage information is based on data stored in the Influx database (InfluxDB). This data is used directly by the Grafana Server [2] and its plugin implemented within the LMS.

The external, public part of the Web services is a Web site [3] built on the Drupal 9 platform. The site contains up-to-date data on software licenses available in the Laboratory, the terms and conditions of their use, and also provides access to the distributions of server software products. The private part of the Web services is a personal account that allows various categories of users to manage the purchase and distribution of licenses, as well as to make changes to the database.

The NLS is a separate subsystem of the LMS. It consists of a separate network segment with private network addresses and a system of access to them based on the general entry into the pool of virtual machines and the transfer of a request to use a license to the corresponding virtual machine via the ports specified inside the request. This is done in order to protect (secure) the pool of virtual machines with license managers. Each virtual machine contains the manager of the corresponding LSP and the manager's log analyzer for obtaining information on monitoring the use of the LSP. When installing a software package, the user specifies a single entry point as the name of the license server, and, depending on the selected package, the required port. When the software package is launched on the user's computer, the license server is contacted, a connection is established with a proxy server, which, depending on the incoming ports, forwards the connection to the appropriate virtual machine with the required license manager, and the license is granted. Accordingly, at the end of using the software package, the license is returned to the pool of available licenses. Thus, the developed structure of the licensing system allows organizing a single entry point for users, easily updating versions of the license manager, and also provides a sufficient level of protection against external intrusions. The manager usually writes into a log file all activities concerning granting and releasing licenses, including the time of license occupation and release, as well as the user's name who requested the license. The log file can be used to collect information about license usage. MLIT has network licenses, all of which use the FlexNet Publisher license manager [4]. In any log, one can select keywords to search for the username and time when the license was granted, as well as another entry with the username and time when the license was returned. An example log file for Maple can be seen in fig. 2. The above keywords are “license granted” and “license accepted”. The output of such a parse is the same for all license managers. The structure of the InfluxDB time series database is in the center of fig. 2.

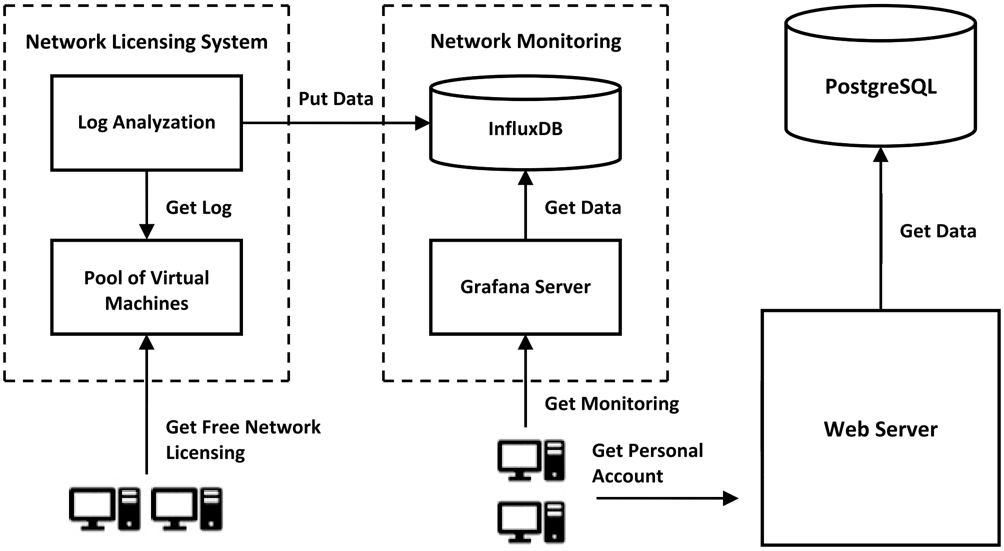
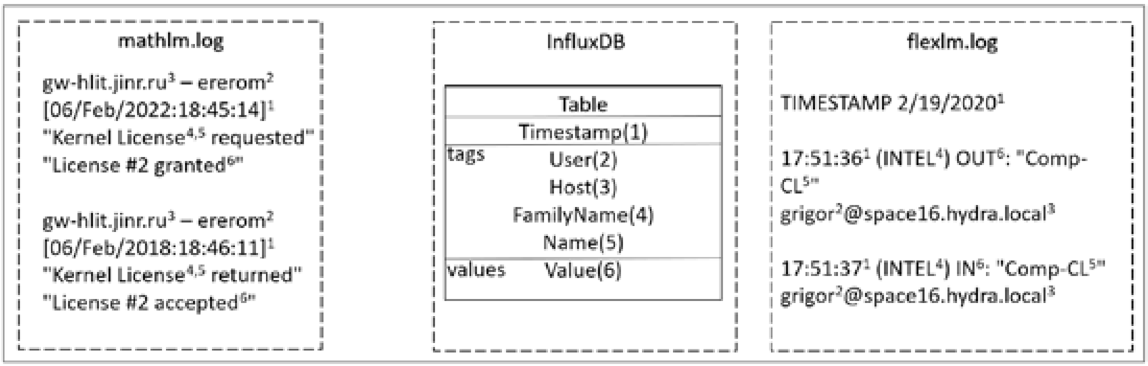


Fig. 1. General architecture of the LMS



mathlm.log

```
gw-hlit.jinr.ru[3] – ererom[2]
[06/Feb/2022:18:45:14][1]
"Kernel License[4,5] requested"
"License #2 granted[6]"

gw-hlit.jinr.ru[3] – ererom[2]
[06/Feb/2018:18:46:11][1]
"Kernel License[4,5] returned"
"License #2 accepted[6]"
```

InfluxDB

Table	
	Timestamp(1)

tags	User(2)
	Host(3)
	FamilyName(4)
	Name(5)

values	Value(6)

flexlm.log

```
TIMESTAMP 2/19/2020[1]

17:51:36[1] (INTEL[4]) OUT[6]: "Comp-CL[5]"
grigor[2]@space16.hydra.local[3]

17:51:37[1] (INTEL[4]) IN[6]: "Comp-CL[5]"
grigor[2]@space16.hydra.local[3]
```

Fig. 2. Structure of an InfluxDB record

InfluxDB records have the following structure: TimeStamp (always required for a time series database), tags, and values. The tag means the string type, and the value means the numeric type of the corresponding parameter. The tag part has the following parameters: User, Host, FamilyName (license family), Name (license name). The value part has only one value: 1 for a license grant operation and 0 for a license release operation. The numbers show the correspondence between the log lines and database entry parameters. The corresponding Log Analyzer runs as a daemon on each virtual machine. It checks the size of the log file, and if the size is changed, it reads the log file, detects the unprocessed part of the log file and parses it finding log records that reflect the user's activity on request or release of licenses. The Log Analyzer writes the parsed data to the InfluxDB. It can parse 2 types of logs: Maple and common flexlm used by Matlab, Comsol, etc. This information is used by the Grafana server to generate network license dashboards at the request of users.

2. Database object model

The developed object model of the LMS is shown in fig. 3. The LMS database keeps permanent data concerning licenses and their usage, users and their roles, software products and companies that produce it, vendors that provide software licenses, as well as other information necessary for the implementation of the project. All information is structured and represented by the corresponding related entities (objects).

The 'license' object keeps parameters that describe a license. The 'license' object stores parameters that describe information about the license, such as the name of the license, vendors contact information, vendor code, license purchase date, who purchase the license, price, and license description. The object also stores important parameters describing the license validity period and other relevant information.

The 'license' object has associations with the 'license_type' object (parameter 'license_type' in use), the 'software' object (parameter 'software_id' in use) and the 'provider' object (parameter 'provider_id' in use). The 'license_type' object has 'license_type_name' parameter that describe the type of the license. There is a set of license types such as Public (information about the license is available to any user of the system), Closed License (license information is available only to those who use it and to special users), Personal License (a license intended for a single user), Local License (a license intended for local installation on a single computer), Group License (a license intended for multiple users or local installation on multiple computers, and the number of computers cannot be limited).

A highly important type is the Network License, which is obtained through a special license server. The network license entitles the user to install the product on several computers, but the number of concurrent users must not exceed the number of purchased licenses. Preinstalled Software (OEM) is also a type of licenses. The user of such software purchases the software together with a personal computer (PC) or a server, and it can only be used on the purchased PC. The last type of licenses used is the Boxes product or Full Package Product. The purchase in this case of one "box" gives permission to use the software product on one computer. Users are allowed to reinstall the product on another PC, but only a certain number of times.

The 'provider' object describes the license provider. The 'software' object represents the licensed software using its name and description. This object has an association with 'company' objects that describe the license owner company. The 'sys_users', 'license_use', 'computer' and 'computer_type' objects describe such essences that are connected with users who use the license and the computer on which it is installed. Only the 'license_use' objects are associated with the 'license' object ('license_id' parameter in use).

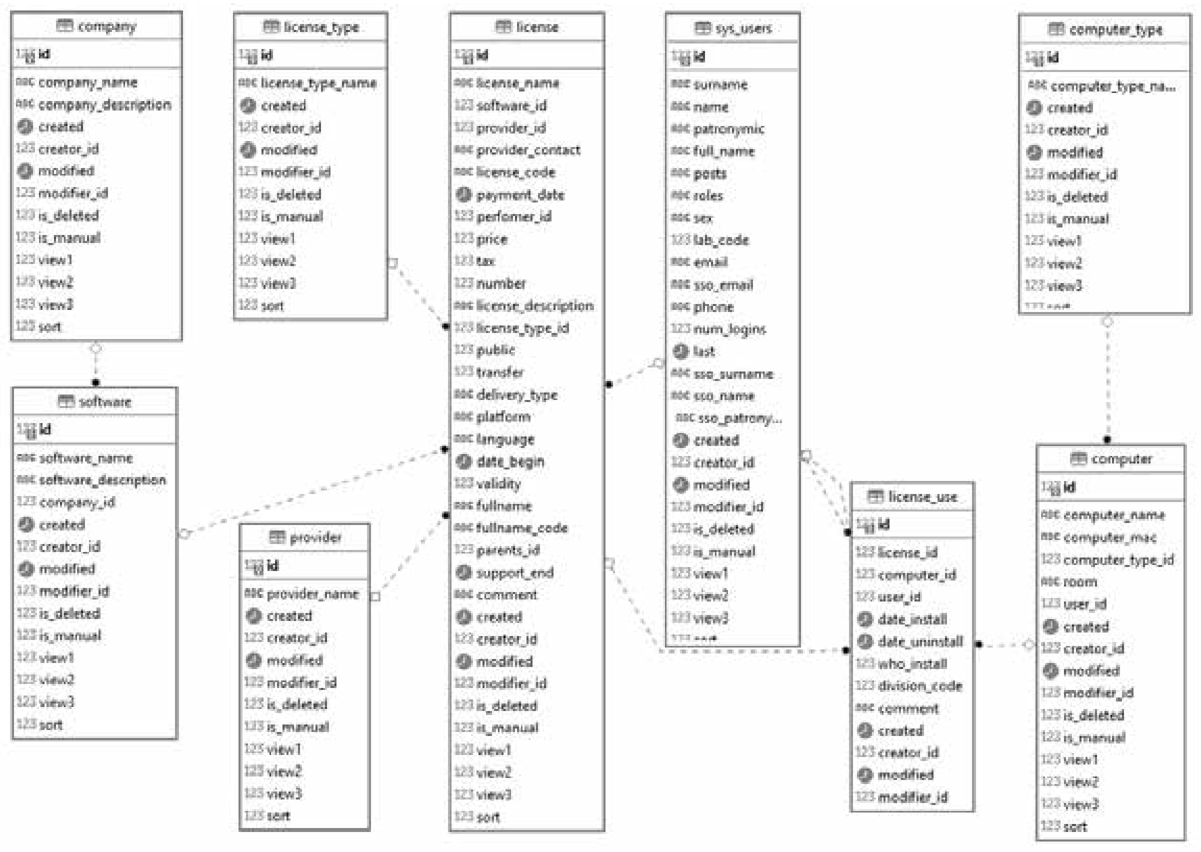


Fig. 3. Object model of the LMS database

The LMS database schema also contains other objects not shown in fig. 3. They provide the functioning of various subsystems of the LMS:

- storage of parameters describing information about users, their roles and their registration in the system;
- functioning of the subsystem “Reference books”;
- integration with other JINR information services, in particular, with “EDMS Dubna” [5].

The main database is implemented on the PostgreSQL platform.

The database that keeps monitoring data is a separate database. This is not an SQL database, and it keeps such data as time series with data in the InfluxDB for Grafana.

3. Web interface design and implementation

The Web interface is designed as a set of personal accounts (PA), which differ depending on user rights or roles. The JINR single sign-on (SSO) is used to authenticate the user and his role when entering the personal account of the LMS.

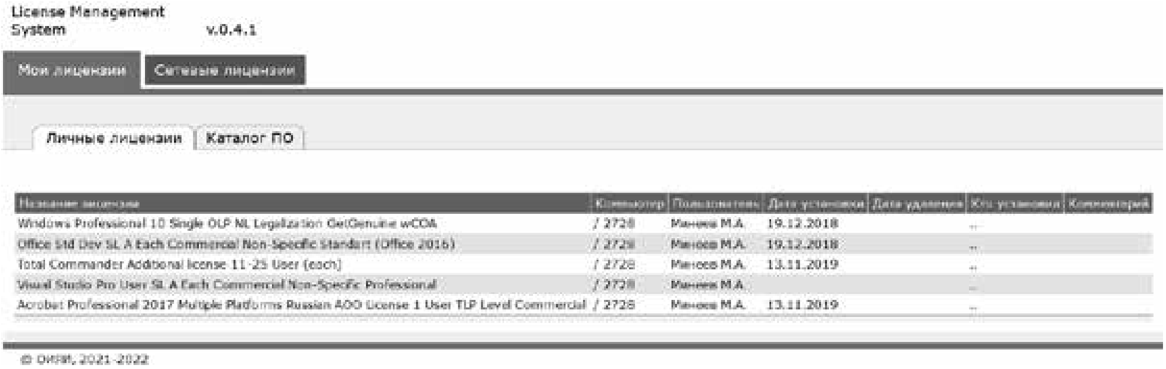
The LMS defines four roles: user, auditor, operator and administrator. The personal account view depends on the role of the client. The user with the user role can see information about all public licenses, any information about own licenses and their expiration period, as well as receive notifications about new licenses. The user can request a new license from the list of existing licenses or a renewal of the old one. The user may request that a new license be added to the catalog of available software for purchasing licenses. The operator has all user rights and can additionally enter into the DB all data concerning licenses, their owners and other information that should be presented in the LMS database. The auditor, additionally to the operator’s rights, can also see any information concerning any licenses, including their cost and any analytical information. The administrator performs any actions available in the LMS database. The administrator has the same rights as the auditor, but with additional purely administrative capabilities. The view of the auditor’s personal account is presented in fig. 4. The user account has a set of TABs at the top of the Web page, and any TAB can also have a set of SUBTABs below the corresponding TAB.

The implementation of the context of tabs and subtabs for different types of user roles is the same with rare exceptions (nested tabs may differ in the presence of some buttons), thus, the pages differ mainly only in the presence of certain tabs and subtabs.

The view of user personal licenses in the user's personal account page is shown in fig. 5 as an example of the subtab context.

Work to exchange information with other JINR information services, such as "EDMS Dubna", is also underway.

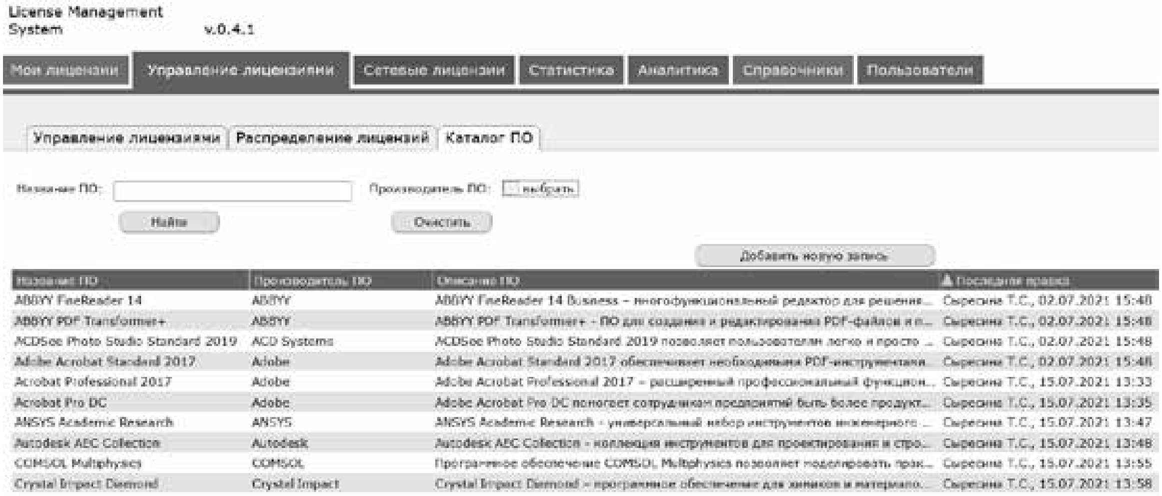
The LMS Web interface is implemented in the development environment of the electronic document management system "EDMS Dubna" using the entire Web Application LegoToolkit (WALT) [6]. This toolkit was developed at MLIT JINR and has been successfully used in other systems for a long time. WALT is a template-oriented platform designed for the development of Web applications of various degrees of complexity, and the main idea of WALT is to provide transparent, extensible, and modifiable tools for solving some specific problems that arise when developing Web applications. The use of WALT for the LMS proved to be highly effective.



Название лицензии	Компьютер	Пользователь	Дата установки	Дата удаления	Кто установил	Комментарий
Windows Professional 10 Single OLP NL Legalization GetGenuine wCOA	/ 2728	Минеев М.А.	19.12.2018		..	
Office Std Dev SL A Each Commercial Non-Specific Standart (Office 2016)	/ 2728	Минеев М.А.	19.12.2018		..	
Total Commander Additional license 11-25 User (each)	/ 2728	Минеев М.А.	13.11.2019		..	
Visual Studio Pro User SL A Each Commercial Non-Specific Professional	/ 2728	Минеев М.А.			..	
Acrobat Professional 2017 Multiple Platforms Russian AOO License 1 User TLP Level Commercial	/ 2728	Минеев М.А.	13.11.2019		..	

© ОИЯИ, 2021-2022

Fig. 4. View of the page of the auditor's personal account



Название ПО	Производитель ПО	Описание ПО	Последняя правка
ABBYY FineReader 14	ABBYY	ABBYY FineReader 14 Business – многофункциональный редактор для решения…	Сыресина Т.С., 02.07.2021 15:48
ABBYY PDF Transformer+	ABBYY	ABBYY PDF Transformer+ - ПО для создания и редактирования PDF-файлов и п…	Сыресина Т.С., 02.07.2021 15:48
ACDSee Photo Studio Standard 2019	ACD Systems	ACDSee Photo Studio Standard 2019 позволяет пользователям легко и просто …	Сыресина Т.С., 02.07.2021 15:48
Adobe Acrobat Standard 2017	Adobe	Adobe Acrobat Standard 2017 обеспечивает необходимыми PDF-инструментами…	Сыресина Т.С., 02.07.2021 15:48
Acrobat Professional 2017	Adobe	Adobe Acrobat Professional 2017 – расширенный профессиональный функцион…	Сыресина Т.С., 15.07.2021 13:33
Acrobat Pro DC	Adobe	Adobe Acrobat Pro DC помогает сотрудникам предприятий быть более продукт…	Сыресина Т.С., 15.07.2021 13:35
ANSYS Academic Research	ANSYS	ANSYS Academic Research - универсальный набор инструментов инженерного …	Сыресина Т.С., 15.07.2021 13:47
Autodesk AEC Collection	Autodesk	Autodesk AEC Collection - коллекция инструментов для проектирования и стро…	Сыресина Т.С., 15.07.2021 13:48
COMSOL Multiphysics	COMSOL	Программное обеспечение COMSOL Multiphysics позволяет моделировать прак…	Сыресина Т.С., 15.07.2021 13:55
Crystal Impact Diamond	Crystal Impact	Crystal Impact Diamond – программное обеспечение для химиков и материало…	Сыресина Т.С., 15.07.2021 13:58

Fig. 5. View of user personal licenses in the user's personal account

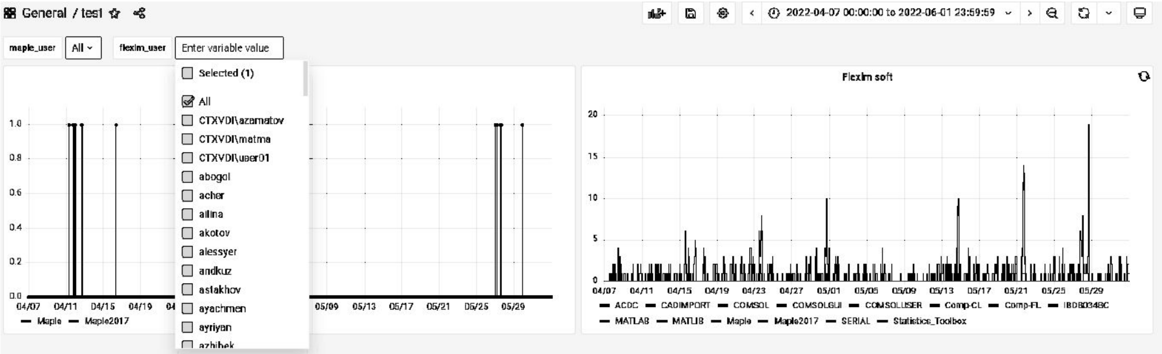


Fig. 6. View of the network license monitoring page

Monitoring the use of network licenses is the basis for analyzing the use of network licenses and the possibility of obtaining a cheaper set of network licenses in the future. The view of the monitoring page is illustrated in fig. 6. The left graph in fig. 6 shows the number of network licenses that users use for Maple. The use of Combobox allows the viewer to set a list of watched users. The right graph shows the number of network licenses that users use for Comsol and Matlab. These programs have several components with separate licenses for each.

The monitoring page can be used as a separate Web page and as the content of a tab of any user's personal account.

The initial version of the LMS release was implemented and is presented above. The current version implementation does not include some workflows, such as ordering for supplementing the catalog of purchased licenses or notification concerning license expiration. Work to develop the system according to the design is underway.

Conclusion

The architecture and object model of the LMS database were developed. The design of the Web interface of the system was completed. The initial release of the LMS was implemented. The system consists of the NLS, Web services and databases. The NLS is a separate LMS subsystem responsible for automatically issuing and releasing network licenses, as well as is a pool of virtual machines, each of which runs the appropriate license manager and log analyzer, supplying monitoring information to the Influx database. Monitoring data is visualized using a plugin for the Grafana server. The Web services consist of the external public part, which contains up-to-date data on software licenses and other corresponding data, and the private part of the services containing personal user accounts. The next version of the implementation is under construction.

References

1. *Pogrebnoy, Dmitry, Ivan Kuznetsov, Yaroslav Golubev, Vladislav Tankov, Timofey Bryksin*. Sorrel: an IDE plugin for managing licenses and detecting license incompatibilities // 2021 IEEE International Conference on Software Maintenance and Evolution (ICSME) [Rules for the citing of sources]. Available at: https://arxiv.org/pdf/2107.13315.pdf.
2. Grafana Home page [Rules for the citing of sources]. Available at: https://grafana.com/.
3. Licensed software in MLIT [Rules for the citing of sources]. Available at: http://soft-lit.jinr.ru.
4. FlexNET Publisher [Rules for the citing of sources]. Available at: https://docs.revenera.com/fnp/2021r2/pdf/fnp_LicAdmin.pdf.
5. *Alexandrov, I.N., O.V. Belyakova, V.V. Korenkov, S.V. Kuniaev, L.N. Pechnikova, M.S. Plyashkevich, S.V. Semashko, G.V. Trubnikov, P.V. Ustenko, S.N. Chikhalina, A.V. Yakovlev*. Development and implementation of electronic document management system "EDMS Dubna" at JINR // CEUR Workshop Proceedings. 2016. Vol. 1787. 97-102 [Rules for the citing of sources]. Available at: http://ceur-ws.org/Vol-1787/97-102-paper-15.pdf.
6. *Korenkov, V.V., S.V. Kuniaev, S.V. Semashko, I.A. Sokolov*. WALT platform for Web Application development // Proceedings of the 9th International Conference "Distributed Computing and Grid Technologies in Science and Education" (GRID'2021), Dubna, Russia, July 5-9, 2021. 2021. Vol. 3041. 387-392 [Rules for the citing of sources]. Available at: http://ceur-ws.org/Vol-3041/387-392-paper-71.pdf.

Alexandrov E.I. Joint Institute for Nuclear Research (JINR), 6 Joliot-Curie St, Dubna, Moscow Region, 141980, Russia, e-mail: aleksand@jinr.ru.

Alexandrov I.N. Phd, Joint Institute for Nuclear Research (JINR), 6 Joliot-Curie St, Dubna, Moscow Region, 141980, Russia, e-mail: alexand@jinr.ru.

Belyakov D.V. Joint Institute for Nuclear Research (JINR), 6 Joliot-Curie St, Dubna, Moscow Region, 141980, Russia, e-mail: dmitry@jinr.ru.

Davydova N.A. Joint Institute for Nuclear Research (JINR), 6 Joliot-Curie St, Dubna, Moscow Region, 141980, Russia, e-mail: luna@jinr.ru

Kalmykova L.A. Junior Researcher, Joint Institute for Nuclear Research (JINR), 6 Joliot-Curie St, Dubna, Moscow Region, 141980, Russia, e-mail: lidak@jinr.ru (correspondent author)

Lyubimova M.A. Joint Institute for Nuclear Research (JINR), 6 Joliot-Curie St, Dubna, Moscow Region, 141980, Russia, e-mail: lma@jinr.ru

Sapozhnikova T.F. PhD, Junior Researcher, Joint Institute for Nuclear Research (JINR), 6 Joliot-Curie St, Dubna, Moscow Region, 141980, Russia, e-mail: tsap@jinr.ru

Syresina T.S. Joint Institute for Nuclear Research (JINR), 6 Joliot-Curie St, Dubna, Moscow Region, 141980, Russia, e-mail: syresina@jinr.ru

Yakovlev A.V. Joint Institute for Nuclear Research (JINR), 6 Joliot-Curie St, Dubna, Moscow Region, 141980, Russia, e-mail: yakovleva@jinr.ru

Zrelov P.V. PhD, Joint Institute for Nuclear Research (JINR), 6 Joliot-Curie St, Dubna, Moscow Region, 141980, Russia, e-mail: zrelov@jinr.ru

On the analysis of individual data on transport usage*

M.V. Bulygin[I], D.E. Namiot[I], O.N. Pokusaev[II]

[I] Lomonosov Moscow State University, Moscow, Russia
[II] Russian Transport University (MIIT), Moscow, Russia

Abstract. The percentage of the world's urban population is currently more than 50\% and will increase according to UN forecasts. Urban infrastructure must develop along with population growth. This article provides an overview of methods for improving the city's transport infrastructure based on data analysis. The article presents methods for reducing harmful emissions, optimizing the operation of taxis and public transport, as well as recognizing transportation modes and some other tasks. These methods operate with data describing the transport behavior of individual users of the transport network. The sources of such data are smart card validators, GPS sensors, and smartphone accelerometers. The article reveals the advantages and disadvantages of using each of the data types, as well as presents alternative ways to obtain them. These methods, along with methods for aggregated data analysis, can become the main part of a single platform that will allow city authorities in the process of improving the transport infrastructure. We propose architecture of this platform which will allows developers to extend range of available algorithms and methods dynamically.

Keywords: *transport data analysis, Data on transport usage, Smart city, Digital urbanism, Smart card data analysis, GPS data analysis.*

DOI: 10.14357/20790279230104

Introduction

According to the Digital 2021 report, more than half of the world's population (56%) lives in cities [1]. According to UN forecasts, the city population will continue to increase further to 6.3 billion (about 65%) by 2050 [4]. According to the study [17], population density affects the quality of life. It is very important to compensate for the negative impact of urban population growth on the comfort of their inhabitants. One of the concepts that make the life of urban residents more comfortable is the concept of a smart city. A smart city is a concept for managing the resources and infrastructure of a city with the widespread use of information technology and Internet of Things technologies. In article [13], a smart city is defined as a well-defined geographical area, in which high technologies such as ICT, logistics, energy production, and so on, cooperate to create benefits for citizens in terms of well-being, inclusion, and participation, environmental quality, intelligent development. It is governed by a well-defined pool of subjects, able to state the rules and policy for the city government and development. Public transport is one of the main components of the urban infrastructure that provides a comfortable city life. In [15], the authors emphasize that smart projects in a technocity should be aimed at transport improvement. To build an easy-to-use public transport system, it is necessary to take into account the transport needs of citizens. In a time before information technology penetrated everyday life, surveys and censuses were used to study the needs of citizens. The collection of such data is expensive, requires the participation and time of citizens, and the data obtained quickly become outdated. In the modern world, new data sources describe the movements of citizens. They provide data at a lower cost and with accuracy and speed that was not possible with the methods of the past. One such data source is mobile phones. During their operation, mobile phones and smartphones exchange information about signal strength and delay with base stations. This information is stored in base stations and can be used to determine the location of devices. The main advantage of these cellular operators is mass character. According to [1], the penetration rate of mobile phones is more than 65% and is constantly increasing. In large cities

* This research has been supported by the Interdisciplinary Scientific and Educational School of Moscow University "Brain, Cognitive Systems, Artificial Intelligence".

and developed countries, according to [3], almost the entire population is subscribers of mobile networks. The important advantage of cellular operators' data is collecting at the operators' base stations, centrally and not visible to users. The disadvantages of such data include their low spatial accuracy, which is sufficient to determine, for example, the area of departure and arrival, but not a specific object. The data of cellular operators are convenient for analysis and mostly used in the form aggregated by city districts and time intervals. Example of aggregated transport data is presented in Table 1. Thus, researchers have access to data on the number of people in each of the city districts. Traffic flow data between each pair of districts for each time interval is also available. Due to the presence of base stations of cellular operators in the subways of some cities, for example, in Moscow, it is possible to measure the traffic flow of the subway separately. The development of data analysis algorithms also makes it possible to obtain information on the number of people traveling from home to work and from work to home.

Table 1

Aggregated transport data example

Timestamp	Source district id	Destination district id	Count of customers
20/01/22 10:00	1	2	1012
20/01/22 10:00	1	3	1258
20/01/22 10:30	1	2	521
20/01/22 10:30	1	3	620

An overview of cellular operators' data analysis methods is presented in the article [10]. Our paper [8] presents an anomaly detection method for aggregated data of cellular operators. Such anomalies correspond to important social events. These events do not need to be predicted. They are either known in advance (for example, large concerts, social events) or are unpredictable (for example, major accidents, fires). City authorities need to respond to such events on time and take measures to improve the traffic situation, so it is necessary to identify changes in traffic flows and measure them. The paper [9] proposes a method for clustering city districts based on aggregated data from cellular operators. As a result of clustering, the city areas were divided into five clusters (from residential to working areas). Urban infrastructure in districts of different clusters should have various development vectors to meet the needs of citizens living in these areas. This article is devoted to an overview of methods for analyzing individual transport data. Such data, in contrast to aggregated data, make it possible to obtain more accurate and granular information about the nature of movements in the city.

1. Main Part

Most modern mobile phones are smartphones. They have many sensors such as GPS [20] and accelerometers. Unlike location data received by cellular operators, data from GPS sensors has higher spatial accuracy. It is important to note that GPS data has a lower penetration rate than cellular data, as not all mobile phones have a GPS sensor. To transfer data, you need to use a special application. Crowdsensing can be used to motivate smartphone users to install applications. It is the provision of some bonuses to users, for example, additional free Internet traffic or free access to content, in exchange for data. The collection of trajectory information can also be embedded in applications that provide transportation services, such as a taxi/bus depot application. The data received by smartphone accelerometers are also collected on the user's side. It can also be obtained utilizing crowdsensing. These data are successfully used to predict the type of activity of a smartphone user and the type of transport they use [35] [29]. The solution to such problems helps to understand the distribution of the traffic flow between different transportation modes. It allows city governments to organize better the transport network of the city. It is possible to use data from smart card validators to analyze the movements of public transport passengers. Smart cards (transport cards) are a popular means of payment in transport systems. Such cards are used to pay for travel in cities such as Moscow, Beijing, Melbourne, and many others. They record information about the balance or the number of remaining trips. Admission to public transport is carried out after passing the validation process. If the cost of the trip depends on the travel distance or the endpoint, then the smart card can also be validated at the end of the trip. This solution is less convenient for passengers. The data of the validators, in this case, allows researchers to accurately determine the trip destination. If destination data is not available, there are heuristics to retrieve it. The paper [33] is devoted to an overview of the data formats collected by the validators and some processing methods. The advantages and disadvantages of using data from various sources are presented in the Table 2.

Data type Advantages Disadvantages GPS data Ability to obtain trajectories with high spatial accuracy, the ability to determine the speed. Low penetration rate, the need to install special applications to obtain Accelerometer data High accuracy in solving transportation mode prediction problems Low penetration rate, the need to install special applications to obtain, do not contain information about the location, a narrow range of tasks to be solved Smart card data Collected centrally, high penetration degree, accurate spatio-temporal marks Reflect only data on the traffic flow in public transport, the endpoint of the trip and the exact trajectory are not always known

Table 2

Advantages and disadvantages of using data from various sources

Data type	Advantages	Disadvantages
GPS data	Ability to obtain trajectories with high spatial accuracy, the ability to determine the speed	Low penetration rate, the need to install special applications to obtain
Accelerometer data	High accuracy in solving transportation mode prediction problems	Low penetration rate, the need to install special applications to obtain, do not contain information about the location, a narrow range of tasks to be solved
Smart card data	Collected centrally, high penetration degree, accurate spatio-temporal marks	Reflect only data on the traffic flow in public transport, the endpoint of the trip and the exact trajectory are not always known

2. GPS data

GPS data are collected from smartphones or navigators using specialized applications. Such an application can be installed on the device by the user himself to exchange his data for some bonuses or useful functionality. Also, applications that collect GPS trajectory data are often installed on the devices of employees of transport companies.

Taxi fleet data. There are two main directions in the analysis of taxi fleet data on trajectories collected using GPS. The purpose of the first direction is to optimize the operation of taxis. Taxi companies are most interested in the development of this area. It includes a behavior analysis of drivers, an analysis of passenger demand, and other tasks. The solution of these tasks will reduce costs and increase the income of taxi services. The second direction is the analysis of the transport network of the city. Within the framework of this direction, the tasks of assessing the total transport flow and its characteristics on the city streets are being solved. The possibilities of reducing harmful emissions from transport services are being analyzed. The latter challenge is especially relevant in some metropolitan areas where high levels of CO_2 emissions harm the health of citizens.

In the study [22], the authors, based on the analysis of GPS data on the trajectories of taxi cars, proposed a method for optimizing taxi ranks, taking into account the spatio-temporal distribution of passenger demand and the waiting time for taxi customers. The proposed methodology is demonstrated on 204 hours of trajectory records in Shenzhen (China). The research materials in this paper are useful to city authorities when planning taxi stops, taxi companies to predict demand, as well as taxi drivers and users who can better estimate the waiting time for a taxi. In [24], GPS data on the trajectories of Beijing taxis are used to study the influence of network companies such as Uber and Didi on taxis in the city. It is concluded that with the advent of taxi network services, the number of trips per day per taxi decreased by 18.08%, and the average daily income by more than 19%. The authors also compare strategies for searching and delivering passengers by taxi drivers with high and low efficiency in different periods. Research [48] is devoted to the detection of anomalous trajectories. According to the authors, anomalous trajectories are trajectories chosen by a small number of drivers that differ from the normal trajectories chosen by other drivers. Such trajectories can serve as signals of incidents in urban traffic systems or fraudulent activities on the part of drivers and passengers. The researchers propose an anomaly detection algorithm based on hierarchical clustering of different trajectories with the same source-destination pairs. They identify four main patterns of abnormal behavior. Computational experiments show the effectiveness of the proposed method for detecting fraudulent routes and adverse road events. The study [12] examines the impact of precipitation on taxi demand. The authors introduce the concept of a high-income passenger to study the demand for a taxi and the driver's efficiency. The researchers show that increased rainfall intensity has the opposite effect on passenger demand for taxis during weekday peak hours and off-peak hours. It is noted that a deeper study of the activities of a taxi is possible using data from taxi calling applications, sensors, as well as financial reports of drivers. Data on the taxi trajectories obtained from GPS sensors also are used to estimate the congestion of city roads. The study [30] proposes a method for calculating the road congestion factor based on the average speed of taxi drivers. As a result of a computational experiment carried out using data on the trajectories of Beijing taxi drivers, the authors set the time limits for the morning and evening rush hours. They estimate the intensity of traffic on the city roads at various time intervals. It notes that the workload in the morning peak hours is higher than in the evening. There are no peak hours on weekends. In the study [28], GPS data on the trajectories of taxi drivers in the city are also used to estimate traffic intensity. Based on this assessment, a model for the analysis of urban transport emissions is proposed. The territory of the city is divided into traffic analysis zones (TAZ). Within them, instantaneous emissions of CO_2, NO_x are estimated. The authors show the relationship between the density of roads and the number of emissions in the TAZ. They conclude that the highest emissions are in TAZs with large business centers. The authors show that within Beijing's fifth ring road, emissions are higher in the north than in the south. The results of this study can be used by the city authorities for lowcarbon urban transport planning, the promotion of alternative energy vehicles, and the design of charging stations. The article [52] is also devoted to the study of emission levels in megacities. The authors analyze data from more than three million GPS trajectories of mobile subscribers obtained in Setagaya, Tokyo. They propose a method to reduce harmful emissions by changing the bike-sharing system. The proposed method showed a reduction in emissions by more than 6% compared to previous methods. The materials of this study are useful in the design and improvement of bicycle-sharing systems around the world. The study [40] is devoted to the comparison of harmful emissions from classic taxis and taxi drivers working with aggregators. In it, the authors compare the trajectories of Didi taxi drivers with ordinary taxi drivers working without an aggregator. The study found that DiDi drivers drive less in search of passengers, heading directly to the pick-up point. Fuel consumption and carbon monoxide emissions, nitrogen oxides, and hydrocarbons per passenger-kilometer are about 1.36 times higher in classic taxi rides.

Bus fleet data. GPS data on vehicle trajectories are massively collected not only in taxi services but also in bus networks. A model for reducing public transport emissions based on GPS trajectory data using the concept of individual buses is proposed in [50]. During a computational experiment conducted on trajectory data recorded by mobile phones in Tokyo, 29 potential individual bus routes are calculated. The researchers identify three types of routes: radial, circular and suburban. The estimated emissions reduction is 13%. The authors find convenient places for the proposed stops of individual buses. This study may be useful to city authorities in the implementation of the concept of individual buses. In [27], the relationship between critical driving events (long stop, hard acceleration, and hard deceleration) and crashes is investigated. The authors use Spearman's rank correlation coefficient based on data on the trajectories of 300 Orlando buses. They find that sudden acceleration and long stops are positively associated with traffic accidents involving pedestrians and cyclists. The authors of the article propose to use the materials of their research in the design and implementation of proactive traffic safety management systems. Paper [53] proposes a framework for evaluating the performance of bus routes based on GPS trajectory data collected in Jinan, China. Several important bus performance metrics are studied, including route times, stop times, idle times, and groupings of buses. The results show that the travel times of the routes follow a correct skewed distribution. In addition, the passage time of a section between two successive stops varies at different periods and is longer during the evening peak hours. The article [41] is devoted to the solution to the problem of planning dedicated lanes for buses. The authors formulate the problem of planning dedicated lanes for buses as a multiobjective optimization problem in which road conditions, traffic flow, bus lane connectivity, and construction cost act as constraints. The use of the road and the punctuality of the bus are taken as objective functions. A method based on an evolutionary algorithm is presented for solving the problem. The operation of the method is illustrated by a computational experiment conducted on GPS data from buses in Shenzhen, China. Often the trajectories of buses following a fixed route are stored in a compressed form. The data includes only relevant bus station arrival and departure records. The article [37] presents a BVI system for visual data analysis. This system contains four data analysis modules. The first module cleans and displays sparse trajectory data. The second module is responsible for analyzing the state of global traffic and traffic patterns of road sections. In the third module, an analysis is made of bus station congestion patterns, and in the fourth, an analysis is made of the importance of bus stations in a complex public transport network. The authors demonstrate the performance and efficiency of the proposed system in three experiments using a data set of real bus GPS trajectories.

GPS data from other sources. The study [7] is devoted to ensuring safety at construction sites based on the analysis of GPS trajectories of builders. The paper proposes a system that processes GPS data, calculates stopping points, trajectory intersections, and provides this information to safety managers. The main distinguishing feature of the proposed system is that it does not use the GPS data directly, but the processed spatio-temporal trajectory data. The materials of this article are useful to identify potentially unsafe behavior at construction and other facilities. In [16], a method for transportation mode prediction based on GPS trajectory data is presented. The main part of the proposed method is preparing data for classifiers (random forests, decision trees, nearest neighbors, Naive Bayes). It consists of five stages. At the first stage, points of GPS trajectories are grouped. Then points signs of trajectories are generated. Trajectory characteristics (percentiles, medians, etc.) are highlighted. Noise is removed from the obtained data. Normalization is carried out. Classifiers built on prepared data show higher accuracy, surpassing classifiers built using other data preparation methods. City authorities need to create a comfortable urban environment for pedestrians. Active walking helps to improve the health of the city's population and improve the quality of life in the city. The study [38] is devoted to the analysis of pedestrian behavior. The authors use GPS data on the trajectories of pedestrians collected using a special application for smartphones. The authors show the influence of various street attributes, which, as is known from previous studies, influence the choice of a walking route. Unlike most studies where the data is limited to a specific type of destination (such as public transit stops), this paper examines a set of trajectories from a wide range of destinations and geographic regions. To evaluate the choice of paths, a new method of alternative path creating is used. The proposed approach obtains information about the attributes of the route using Google Street View image analysis.

The paper [55] proposes a system for constructing a city pedestrian network. This system includes three key modules. The first module is filtering data on walking trajectories. The second is building a pedestrian network and the third is its evaluation. Data for analysis is obtained using crowd-sensing from the GPS sensors of phones. The authors conduct an experiment showing that the pedestrian network extracted using the proposed system is accurate and complete. The work [25] is devoted to modeling the trajectories of pedestrians in the city. An important problem in the analysis of pedestrian trajectory data is large errors in positioning caused by large buildings and frequent stops, and direction changes. The authors proposed a system for modeling pedestrian trajectories that solve such problems. The paper describes in detail the architecture of the system, and the tools necessary to implement such an architecture. The study [46] proposes a method for people crowds detection based on GPS movement trajectories data. The main feature of the proposed method is resistance to noise and missing data, which is typical for data collected in urban areas. The results of the computational experiment show that the method accuracy of detecting crowds and isolating their members is 91.3

3. Smart card data

Currently, smart cards are widely used to pay fares in the transport systems of many cities. Data on the validation of transport cards are collected by special systems and are utilized mainly for invoicing. These data can also be used to solve other applied problems. Example of individual transsport data is presented in Table 3 A large number of individual transport contributes to an increase in the number of harmful emissions into the atmosphere. The main alternative to individual transport is public transport. The use of it is more environmentally friendly. For increasing public transport use, it is necessary to identify the reasons why urban transport is not attractive to citizens. The study [36] is devoted to identifying factors contributing to the public transport use reduction. The authors built a Cox regression model on features obtained from smart card data. These features include for each passenger the share of weekdays using public transport, the number of pairs of places of departure and arrival, the share of tram use, the type of transport card, and others. Based on the analysis, the authors propose incentive measures to maintain and increase the use level of public transport for various population groups.

Table 3

Smart card data example

Timestamp	Validator id	Smart card id	Balance
20/01/22 09:30:55	0	125558	500
20/01/22 09:31:32	1	136472	362
20/01/22 09:32:05	0	123657	414
20/01/22 09:32:55	0	130058	580

In Brisbane, Queensland, Australia, smart cards are used to access CityCat ferry transport. The article [39] provides a detailed study of the data of more than 1.5 million smart card data transactions. The authors establish that, despite the presence of only one route, more than 2% of trips on public transport are made on these ferries. The article notes that the use of ferries is mainly for going to work and school during rush hours. The ferry use level at the weekend remained at the same level. A cluster of users using the ferries for one-time recreational purposes is identified. A high degree of integration of the CityCat system with other transportation modes is established. More than 15% of trips continued using other transportation modes. The paper [54] is devoted to the transfer identification problem. This paper investigates transfers between subways and bike-sharing systems. An important difficulty in conducting such studies is the use of various transport cards for metro travel and bicycle rental. The authors proposed a method for comparing data from smart cards of two different transport systems for one passenger. In the computational experiment, this method showed an accuracy of 100% for 573 passengers under study. The authors examine the identified transfers and conclude the movement of passengers. The authors note that 2/3 of bike-to metro transfers and vice versa take place during peak hours, and bike-to-bike and bike trip times account for an average of 27% of the total trip time. For the correct distribution of transport resources in the city and the prompt response to various incidents, the city administration needs to understand how traffic flows change during the day, and in emergencies. The article [51] is devoted to the subway daily traffic fluctuation study. The authors use data from more than a million trips over five working days with normal weather conditions in Nyanzhin city. The researchers identify the coefficient of increase in traffic during peak hours, analyze and compare passenger traffic on different metro lines and at different stations. The authors use thermodynamic diagrams to visualize the inflow and outflow of passengers at stations. This representation allows researchers to identify congestion in the metro in the city. Also, it helps to visualize the features of passenger traffic in the city. The article [34] is devoted to the accident impact study on the city transport system. The authors found that medium-term disruptions can have long-term consequences for the travel patterns of long-term users of the affected infrastructure. They note that their method is one of the first in this area, using passively collected data. Other studies in this area use data from questionnaires or surveys that require participation from passengers. One of the main problems of modern megacities is their monocentric. According to our study [9], the working districts of Moscow are concentrated in the center of the city. This leads to the fact that urban residents from peripheral areas are forced to spend a lot of time traveling to their places of work and also stimulates the use of personal transport. This leads to an increase in harmful emissions into the atmosphere in the city center. The study [31] is devoted to identifying spatiotemporal patterns of trips to work for public transport passengers in Beijing. Using one month's smart card validation data, the authors identify the places of work and residence of individual passengers, as well as the time of their departure. Visualization of the obtained data showed significant differences between workplaces and residences in Beijing. The study materials are useful to build a balanced transport system in a monocentric city. The validator data contains not only information on general traffic flows and individual traffic behavior, but also indirect features describing the transport network users. For example, information on the smart card validation can also be used to obtain socioeconomic features describing their users. Such data are useful for the rational allocation of social subsidies, the detection of potential fraud with social smart cards, and for solving other applied problems. Modern methods for assessing socioeconomic behavior are based on data on the behavior of users in cyberspace, for example, on data from social networks [6]. Article [14] proposes an approach to assess the socioeconomic status of a smart card user based on deep learning. This approach combines the mass nature of modern methods based on data, but at the same time works with data on real, not virtual, user behavior. The paper illustrates the application of the proposed method on the Shanghai SCD dataset containing data on more than a million smart card users. Data similar in nature to smart card validator data can be collected in other ways. The paper [19] proposes an approach for calculating travel time based on data from Wi-Fi scanners. With this approach, passenger devices with a Wi-Fi receiver, such as a smartphone or tablet, act as a smart card, the MAC addresses of their devices act as a passenger identifier, and a Wi-Fi scanner acts as a validator. This approach makes it possible to obtain data that can be used for analysis by the methods described above and expand their scope. It is important to note that modern mobile operating systems may prevent such tracking of device owners to increase privacy. When connected to Wi-Fi, such devices generate random MAC addresses, which limits the application of the method proposed above. Alternatively, an approach can be proposed that uses unique Bluetooth addresses instead of Wi-Fi MAC addresses. Quite a lot of popular wearable wireless devices use this technology (wireless headphones, fitness bracelets, and others), while the Bluetooth address of these devices does not change when connected. Currently, the development and implementation of alternative payment methods to smart cards are underway. An example is the introduction of a payment system using face scanning in the Moscow metro. When using such systems, data on the validation of the transport system user is also saved, so smart card data analysis methods will not lose their relevance. Let us summarize all the presented methods for analyzing individual transport data in the Table 4.

Table 4

Problems and bibliography

Problem	Bibliography
Taxi research and optimization	[22], [24], [45], [48], [12]
Public transport research and optimization	[27], [53], [41], [37], [39], [36], [51], [34], [31
Identification and research of transfers	[23],[54]
City road congestion assessment	[30], [28]
Reducing harmful emissions in megacities	[28], [52], [40], [50]
Transportation mode prediction	[7], [16], [35], [29]
Pedestrian network research	[38], [55], [46], [25]

4. Discussion

Our literature review shows that there are many methods for solving urban transport management problems. In addition, there are ready-made platforms that combine groups of such them. Typically, such platforms, for example [18], con tain a fixed set of algorithms that solve some previously known set of problems. Other platforms, such as [26], fix a specific problem-solving method. In [26], an architecture based on neural networks is fixed for solving regression problems. Such algorithms have low explainability and can be difficult to tune. We propose a new approach to the platform architecture for data analysis in digital urbanism. Amount of tasks that arise in urban infrastructure management reduces to the tasks of finding deviations from some given “normality”. Thus, to solve them, the platform must provide modules for working with the initial data, and modules for describing/obtaining normality and searching for deviations from it. There are two data processing modules. The first is for reading and loading data, storing it, and also checking the correctness of the data type. It is important to note that in urban studies, most of the analyzed data can be reduced either to correspondence matrices (cellular operator data, aggregated GPS, or smart card data) or to a form similar to that presented in Table 1 (individual GPS data, smart cards). This feature allows you to work with data from different sources in a unified way. The second module is responsible for the semantic validation of the submitted data according to heuristic rules. The third module of the platform provides some standard ways of describing normalities and methods for detecting deviations. As an example, the normality model presented in [8] can be taken. The API module will allow developers to define their own “normalities” and deviation-detecting methods. This module will allow the platform to overcome the limitations of existing platforms in the form of a fixed range of tasks to be solved and a fixed architecture for their solution since new normality and deviation-detecting methods can be defined on any data of one of two standard types. The results of the algorithms must be visualized in a form convenient for perception. The platform must contain a data visualization module to do this. One of the important parts of such a platform is methods for implementing visualization on city maps, and graphs. Ready-made solutions, such as [2], can be used as the basis for such methods. The platform architecture diagram is shown in Fig. 1

According to this architecture, it is possible to create a comprehensive platform for data analysis in urban studies. The main advantages of such a platform are the ability to work with different data sources, and the ability to dynamically expand the base of platform analysis methods using the API, which will eventually expand the range of tasks to be solved with low labor costs.

Conclusion

In this article, individual transport data analysis methods are considered. Many of these methods can help cities design transportation systems. Data sources, combined with new methods of analysis, help to better understand the transport needs of urban residents and improve the comfort of their movement, as well as reduce travel time. Modern data sources allow answering questions about when, where, and how citizens move. We propose a platform architecture that will allow us to combine many urban data analysis algorithms from different sources. This architecture allows developers to describe the concept of normality and deviations themselves, which makes it possible to dynamically expand the range of available algorithms. The next step in the development of transport systems in the cities of the future may be the development of self-driving vehicles, which will provide a higher quality of service and lower operating costs. In recent years, the development of robotic cars [11], as well as the concepts of their interaction with regular cars driven by people and pedestrians [43,47], are actively carried out. In many respects, this became possible due to the development of algorithms for computer vision [21,42], depth estimation [32,44],

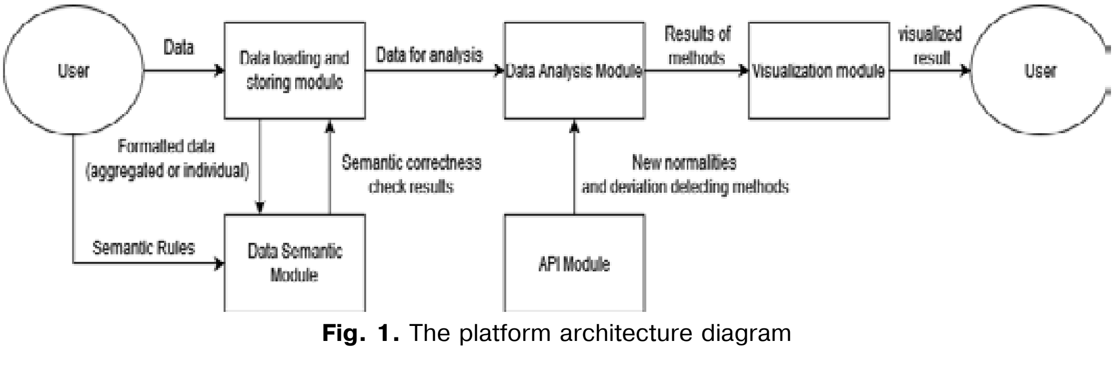

User
Data
Data loading and storing module
Data for analysis
Data Analysis Module
Results of methods
Visualization module
visualized result
User
Formatted data (aggregated or individual)
Semantic correctness check results
New normalities and deviation detecting methods
Semantic Rules
Data Semantic Module
API Module

Fig. 1. The platform architecture diagram

and three-dimensional object detection using LiDAR data (3D-object-detection) [49]. With the widespread use of self-driving vehicles, there will be opportunities for building automated logistics systems (autonomous logistics), where artificial intelligence solves not only the tasks of traffic flow planning but also the direct transportation of goods or passengers [5].

References

1. Digital 2021. Report by "We are social" agency, https://wearesocial.com/uk/blog/2021/01/digital-2021-uk/
2. Kepler.gl, https://kepler.gl
3. Measuring digital development Facts and figures 2021, https://www.itu.int/en/ITU-D/Statistics/Documents/facts/FactsFigures2021.pdf
4. New thinking urgently required to shift urbanisation onto a more sustainable path, https://fennerschool.anu.edu.au/news-events/news/new-thinking-urgentlyrequired-shift-urbanisation-more-sustainable-path
5. *Abosuliman, S.S., Almagrabi, A.O.*: Routing and scheduling of intelligent autonomous vehicles in industrial logistics systems. Soft Computing 25(18), 11975-11988 (2021)
6. *Ajrouch, K.J., Blandon, A.Y., Antonucci, T.C.*: Social networks among men and women: The effects of age and socioeconomic status. The Journals of Gerontology Series B: Psychological Sciences and Social Sciences 60(6), S311-S317 (2005)
7. *Arslan, M., Cruz, C., Roxin, A.M., Ginhac, D.*: Spatio-temporal analysis of trajectories for safer construction sites. Smart and Sustainable Built Environment (2018)
8. *Bulygin, M., Namiot, D.*: Anomaly detection method for aggregated cellular operator data. In: 2021 28th Conference of Open Innovations Association (FRUCT). pp. 42-48. IEEE (2021)
9. *Bulygin, M., Namiot, D.*: A new approach to clustering districts and connections between them based on cellular operator data. In: 2021 29th Conference of Open Innovations Association (FRUCT). pp. 71-80. IEEE (2021)
10. *Bulygin, M., Namiot, D.*: On the possibilities of using the data of cellular operators to solve the problems of digital urbanism. International Journal of Open Information Technologies 9(1), 48-57 (2021)
11. *Chan, T.K., Chin, C.S.*: Review of autonomous intelligent vehicles for urban driving and parking. Electronics 10(9), 1021 (2021) 12.
12. *Chen, D., Zhang, Y., Gao, L., Geng, N., Li, X.*: The impact of rainfall on the temporal and spatial distribution of taxi passengers. Plos one 12(9), e0183574 (2017)
13. *Dameri, R.*: Searching for smart city definition: a comprehensive proposal. International Journal of Computers & Technology 11, 2544 (10 2013). https://doi.org/10.24297/ijct.v11i5.1142
14. *Ding, S., Huang, H., Zhao, T., Fu, X.*: Estimating socioeconomic status via temporal-spatial mobility analysis-a case study of smart card data. In: 2019 28th international conference on computer communication and networks (ICCCN). pp. 1-9. IEEE (2019)
15. *Downey, J., McGuigan, J.*: Technocities: The Culture and Political Economy of the Digital Revolution. Sage (1999)
16. *Etemad, M., Soares Júnior, A., Matwin, S.*: Predicting transportation modes of gps trajectories using feature engineering and noise removal. In: Canadian conference on artificial intelligence. pp. 259-264. Springer (2018)
17. *Fassio, O., Rollero, C., De Piccoli, N.*: Health, quality of life and population density: A preliminary study on contextualized quality of life. Social indicators research 110(2), 479-488 (2013)
18. *Golubev, A., Chechetkin, I., Solnushkin, K.S., Sadovnikova, N., Parygin, D., Shcherbakov, M.*: Strategway: web solutions for building public transportation routes using big geodata analysis. In: Proceedings of the 17th international conference on information integration and web-based applications & services. pp. 1-4 (2015)
19. *Hidayata, A., Terabea, S., Yaginumaa, H.*: Time travel estimations using mac addresses of bus, passengers: A point to path-qgis analysis. Geoplanning: Journal of Geomatics and Planning 5(2), 259-268 (2018)
20. *Hofmann-Wellenhof, B., Lichtenegger, H., Collins, J.*: Global positioning system: theory and practice. Springer Science & Business Media (2012)
21. *Hong, Y., Pan, H., Sun, W., Jia, Y.*: Deep dual-resolution networks for real-time and accurate semantic segmentation of road scenes. arXiv preprint arXiv:2101.06085 (2021)
22. *Hu, X., An, S., Wang, J.*: Taxi driver's operation behavior and passengers' demand analysis based on gps data. Journal of advanced transportation 2018 (2018)
23. *Huang, Z., Xu, L., Lin, Y., Wu, P., Feng, B.*: Citywide metro-to-bus transfer behavior identification based on combined data from smart cards and gps. Applied Sciences 9(17), 3597 (2019)
24. *Jiang, W., Lian, J., Shen, M., Zhang, L.*: A multi-period analysis of taxi drivers' behaviors based on gps trajectories. In: 2017 IEEE 20th International

Conference on Intelligent Transportation Systems (ITSC). pp. 1-6. IEEE (2017)

25. *Karim, L., Boulmakoul, A., Zeitouni, K.*: From raw pedestrian trajectories to semantic graph structured model towards an end-to-end spatiotemporal analytics framework. Procedia Computer Science 184, 60-67 (2021)
26. *Ke, S., Song, L., Bao, K., Pan, Z., Zhang, J., Zheng, Y.*: East: An enhanced automated machine learning library for spatio-temporal forecasting
27. *Li, P., Abdel-Aty, M., Yuan, J.*: Using bus critical driving events as surrogate safety measures for pedestrian and bicycle crashes based on gps trajectory data. Accident Analysis & Prevention 150, 105924 (2021)
28. *Li, T., Wu, J., Dang, A., Liao, L., Xu, M.*: Emission pattern mining based on taxi trajectory data in beijing. Journal of cleaner production 206, 688-700 (2019)
29. *Liang, X., Zhang, Y., Wang, G., Xu, S.*: A deep learning model for transportation mode detection based on smartphone sensing data. IEEE Transactions on Intelligent Transportation Systems 21(12), 5223-5235 (2019)
30. *Liu, C., Wang, S., Cuomo, S., Mei, G.*: Data analysis and mining of traffic features based on taxi gps trajectories: A case study in beijing. Concurrency and Computation: Practice and Experience 33(3), e5332 (2021)
31. *Ma, X., Liu, C., Wen, H., Wang, Y., Wu, Y.J.*: Understanding commuting patterns using transit smart card data. Journal of Transport Geography 58, 135-145 (2017)
32. *Murez, Z., As, T.v., Bartolozzi, J., Sinha, A., Badrinarayanan, V., Rabinovich, A.*: Atlas: End-to-end 3d scene reconstruction from posed images. In: European Conference on Computer Vision. pp. 414-431. Springer (2020)
33. *Namiot, D., Sneps-Sneppe, M.*: A survey of smart cards data mining. In: AIST (Supplement). pp. 314-325 (2017)
34. *Nazem, M., Lomone, A., Chu, A., Spurr, T.*: Analysis of travel pattern changes due to a medium-term disruption on public transit networks using smart card data. Transportation Research Procedia 32, 585-596 (2018)
35. *Ni, Q., Fan, Z., Zhang, L., Zhang, B., Zheng, X., Zhang, Y.*: Daily activity recognition and tremor quantification from accelerometer data for patients with essential tremor using stacked denoising autoencoders. International Journal of Computational Intelligence Systems 15(1), 1-13 (2022)
36. *Nishiuchi, H., Chikaraishi, M.*: Identifying passengers who are at risk of reducing public transport use: A survival time analysis using smart card data. Transportation research procedia 34, 291-298 (2018)
37. *Pei, W., Wu, Y., Wang, S., Xiao, L., Jiang, H., Qayoom, A.*: Bvis: urban traffic visual analysis based on bus sparse trajectories. Journal of Visualization 21(5), 873-883 (2018)
38. *Sevtsuk, A., Basu, R., Li, X., Kalvo, R.*: A big data approach to understanding pedestrian route choice preferences: Evidence from san francisco. Travel behaviour and society 25, 41-51 (2021)
39. *Soltani, A., Tanko, M., Burke, M.I., Farid, R.*: Travel patterns of urban linear ferry passengers: Analysis of smart card fare data for brisbane, queensland, australia. Transportation Research Record 2535(1), 79-87 (2015)
40. *Sui, Y., Zhang, H., Song, X., Shao, F., Yu, X., Shibasaki, R., Sun, R., Yuan, M., Wang, C., Li, S., et al.*: Gps data in urban online ride-hailing: A comparative analysis on fuel consumption and emissions. Journal of Cleaner Production 227, 495-505 (2019)
41. *Sun, Y., Wu, M., Li, H.*: Using gps trajectories to adaptively plan bus lanes. Applied Sciences 11(3), 1035 (2021)
42. *Tao, A., Sapra, K., Catanzaro, B.*: Hierarchical multi-scale attention for semantic segmentation. arXiv preprint arXiv:2005.10821 (2020)
43. *Tran, T.T.M., Parker, C., Tomitsch, M.*: A review of virtual reality studies on autonomous vehicle-pedestrian interaction. IEEE Transactions on Human-Machine Systems (2021)
44. *Wang, L., Zhang, J., Wang, O., Lin, Z., Lu, H.*: Sdc-depth: Semantic divideand-conquer network for monocular depth estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 541-550 (2020)
45. *Wang, Y., Qin, K., Chen, Y., Zhao, P.*: Detecting anomalous trajectories and behavior patterns using hierarchical clustering from taxi gps data. ISPRS International Journal of Geo-Information 7(1), 25 (2018)
46. *Wirz, M., Schläpfer, P., Kjærgaard, M.B., Roggen, D., Feese, S., Tröster, G.*: Towards an online detection of pedestrian flocks in urban canyons by smoothed spatio-temporal clustering of gps trajectories. In: Proceedings of the 3rd ACM SIGSPATIAL international workshop on location-based social networks. pp. 17-24 (2011)
47. *Xing, Y., Lv, C., Cao, D., Hang, P.*: Toward human-vehicle collaboration: Review and perspectives on human-centered collaborative automated driving. Transportation research part C: emerging technologies 128, 103199 (2021)

48. *Yang, S., Bi, S., Athanase, N., Huang, T., Wan, L.*: Spatial clustering method for taxi passenger trajectory. Computer Engineering and Applications 54(14), 249-255 (2018)
49. *Yin, T., Zhou, X., Krahenbuhl, P.*: Center-based 3d object detection and tracking. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11784-11793 (2021)
50. *Yu, Q., Zhang, H., Li, W., Song, X., Yang, D., Shibasaki, R.*: Mobile phone gps data in urban customized bus: Dynamic line design and emission reduction potentials analysis. Journal of Cleaner Production 272, 122471 (2020)
51. *Yu, W., Bai, H., Chen, J., Yan, X.*: Analysis of space-time variation of passenger flow and commuting characteristics of residents using smart card data of nanjing metro. Sustainability 11(18), 4989 (2019)
52. *Zhang, H., Song, X., Long, Y., Xia, T., Fang, K., Zheng, J., Huang, D., Shibasaki, R., Liang, Y.*: Mobile phone gps data in urban bicycle-sharing: Layout optimization and emissions reduction analysis. Applied Energy 242, 138-147 (2019)
53. *Zhang, H., Shi, B., Song, S., Zhao, Q., Yao, X., Wang, W.*: Statistical analysis of the stability of bus vehicles based on gps trajectory data. Modern Physics Letters B 33(03), 1950015 (2019)
54. *Zhao, D., Wang, W., Ong, G.P., Ji, Y.*: An association rule based method to integrate metro-public bicycle smart card data for trip chain analysis. Journal of Advanced Transportation 2018 (2018)
55. *Zhou, B., Zheng, T., Huang, J., Zhang, Y., Tu, W., Li, Q., Deng, M.*: A pedestrian network construction system based on crowdsourced walking trajectories. IEEE Internet of Things Journal 8(9), 7203-7213 (2020)

Bulygin M.V. PhD student, Lomonosov Moscow State University, MSU, Faculty of Computational Mathematics and Cybernetics, Russia, 119991, Moscow, GSP-1, 1-52, Leninskiye Gory, e-mail: messimm@yandex.ru (correspondent author)

Namiot D.E. Dr. of Sci., Lomonosov Moscow State University, MSU, Faculty of Computational Mathematics and Cybernetics, Russia, 119991, Moscow, GSP-1, 1-52, Leninskiye Gory, email: dnamiot@gmail.com

Pokusaev O.N. PhD, docent, Russian Transport University, MIIT, Higher Engineering School, Russia, 127055 Moscow, Novosushchevskaya st., 22, building 2, email: dnamiot@gmail.com

Хранилище сымитированных типовых сигналов как основа разработки быстрых алгоритмов*

И.И. Дейкин, В.В. Сюзев, Е. В.Смирнова, А.В. Пролетарский

Федеральное государственное бюджетное образовательное учреждение высшего образования «Московский государственный технический университет имени Н.Э. Баумана (национальный исследовательский университет)», г. Москва, Россия

Аннотация. Статья посвящена структуре хранилища многомерных типовых сигналов и процессов в рамках задачи цифровой фильтрации, основанной на быстрых алгоритмах имитации и операторе универсального адаптивного матричного преобразования. Хранилище содержит актуальные уравнения имитации; сымитированные сигналы в виде таблиц значений и изображений; оригинальный код на Python, C++ и других языках программирования; дополнительную лицензионную информацию. В статье рассмотрены различные подходы к созданию хранилища. Выбранный путь заключается в формировании набора данных, его публикации на ресурсе GitHub и построении онтологии на данном наборе. Принципы структурирования данных и организации такой онтологии представлены в статье. Представлен готовый опубликованный набор данных.

Ключевые слова: *хранилище данных, хранилище сигналов, многомерные сигналы, имитация сигналов, цифровая обработка сигналов, ЦОС, онтология, гармонические методы.*

DOI: 10.14357/20790279230105

Введение

Задача имитации сигналов, рассматриваемая в рамках цифровой обработки сигналов, становится особенно актуальной в условиях цифровой экономики. Имитация позволяет не только проводить экономически выгодное натурное моделирование, посылая реалистичные сигналы на проектируемое оборудование, но также моделировать состояния различных систем в задачах калибровки существующего оборудования, тренировки персонала и прогнозирования [1, 2]. Существуют различные методы имитации сигналов. Распространенные методы имитации сигналов предлагают выбирать между качеством имитации и экономией вычислительных ресурсов [3]. Одним из перспективных направлений исследований является исследование гармонических методов имитации сигналов, представляющих возможность совмещения качества и экономного расхода вычислительных ресурсов для определенных типов сигналов. Гармонические методы имитации разнообразные, и поэтому становится целесообразным рассмотрение различных гармонических базисов и различных подходов к имитации в рамках гармонической имитации [4].

Коллективная разработка новых методов и инструментов цифровой обработки сигналов нуждается в организации хранилища результатов: математических выкладок, результатов их экспериментальной проверки и алгоритмов реализации методов имитации сигналов и процессов, часть из которых защищена авторскими свидетельствами на Программы для ЭВМ [5 – 8]. Вследствие необходимости их анализа предлагается разработать хранилище сымитированных эталонных сигналов и сопутствующих материалов: теоретических выкладок, кодов алгоритмов, лицензионных документов.

Разработка современного хранилища опирается на совершенствование алгоритмов, структур данных и их взаимодействия. Тенденции вроде кэш-осознанности и энергетической осознанности, параллельные и облачные вычисления отражают требования, предъявляемые современным технологиям, и мотивируют поиск способов совершенствования базового канона алгоритмов

* Исследование выполнено за счет гранта Российского научного фонда №22-11-00049, https://rscf.ru/project/22-11-00049/ («Разработка корреляционной теории моделирования многомерных сигналов и процессов в гибридных системах искусственного интеллекта реального времени»).

и структур данных общего назначения [9]. Особенно важна организация устойчивого развития в фундаментальной для современного общества сети Интернет, работа которой требовательна к вычислительным и, значит, к энергетическим ресурсам [10]. Хотя стандартные алгоритмы и простые структуры данных полезны при разработке новых систем и составляют основу для будущего развития, существует потребность в новых умных способах организации взаимодействия данных и алгоритмов.

Проект #22-11-00049 при поддержке Российского Научного Фонда, включающий сравнительный анализ различных методов имитации сигналов, требует организации взаимного доступа исследователей к результатам исследования этих методов. Разрабатываемое хранилище призвано позволить исследователям, работающим над разными методами, проверять себя и коллег путем отработки методов на одинаковых экспериментальных настройках и сравнения полученных результатов. Также предполагается доступ третьих лиц к результатам научной деятельности команды.

1. Возможные подходы

Существует множество различных подходов к организации открытого доступа к научным результатам [11 – 20]. Используемые в них алгоритмы и структуры данных соответствуют базовому канону алгоритмов и структур данных общего назначения, что эффективно для ресурсов малого масштаба, но может представлять проблему в их дальнейшем развитии. Однако такой простой подход создает базу для дальнейшего совершенствования. Анализ существующих решений позволил систематизировать их по методу доступа к данным и выделить следующие виды хранилищ:

- хранилище как отдельное программное приложения;
- хранилище как интерактивное Интернет-приложение;
- хранилище как опубликованный в сети Интернет набор данных.

Все три вида нуждаются в рассмотрении по-отдельности, в выделении позитивных и негативных черт с целью дальнейшего сравнительного анализа.

Отдельные приложения позволяют получать имитационные результаты лицам, имеющим доступ к этим приложениям. При этом люди нередко не доверяют приложениям неизвестных разработчиков и не готовы запускать такие приложения на своем оборудовании. Такая проблема не возникает при обмене приложением в команде разработчиков-коллег, но может мешать распространению научных результатов за пределами команды разработчиков. Приложения не предоставляют открытого доступа к исходному коду методов имитации. Несмотря на то, что различные инструменты могут использоваться для извлечения этого кода, подобные риски малы, когда речь идет об узкоспециальном программном обеспечении.

Разработка приложения является отдельной проблемой. Исследование методов имитации сигналов в различных базисах сопровождается созданием их программной реализации. Так, участниками команды были разработаны программные системы имитации сигналов в базисе Фурье, в базисе Хартли и в других базисах [11, 12]. Разработка проводилась отдельными разработчиками под определенные задачи, что привело к гетерогенности результатов: использовались разные языки программирования, разные технологии, разные среды разработки. Объединение таких разрозненных продуктов в единую систему имитации является важной задачей и потребует анализа имеющихся результатов, перевода части программного кода на другие языки программирования, проектирования обновленного интерфейса и структуры системы.

Размещение веб-приложения в сети Интернет позволяет обеспечить доступ к результатам научной деятельности третьим лицам. Среди положительных черт можно обозначить поддержку интерактивности приложения, доступ из любой точки мира и с любого устройства, отсутствие необходимости настройки системы на оборудовании пользователя. Интерактивность находит применение в исследовании спектров атомов различных веществ и спектров космических тел в ресурсе университета Вашингтона и ресурсе NIST [13, 14]. Особенный интерес предоставляет внедрение тетрадей Jupyter Notebook в веб-сайты с возможностью как их запуска с последующей интерактивной работой, так и просмотра их программного кода [15].

Организация веб-приложений, тем не менее, также имеет негативные черты. Разработка веб-приложений отличается от разработки обычных приложений и требует от команды разработчиков иных компетенций. Применяются другие языки, системы и парадигмы разработки, необходимо выбрать между поддержкой приложения на стороне клиента или на сервере, что потребует настройки и поддержки такого сервера. Широта доступа подразумевает самодокументированность веб-приложения и его предсказуемое поведение, обработка возможных ошибок ввода и информационная безопасность ресурса. За счет задержки, существующей при передаче данных в сети Интернет, страдает скорость работы приложения.

Простейшим способом распространения результатов имитации сигналов может считаться организация публично доступных наборов данных или же обычная публикация файлов в сети Интернет. В таком случае доступ к исходному коду возможен, только если разработчики сами разместят его, но и доступ к результатам неинтерактивен. Только опубликованные данные оказываются доступными посетителям набора данных. Возникает задача определения важности тех или иных данных и их структуризация. При этом обмен файлами является одним из традиционных применений сети Интернет, поэтому существует множество различных простых способов реализации этой функции.

Отдельные разработчики публикуют свои наборы сигналов на ресурсе GitHub, который обычно предназначен для контроля версий разрабатываемого программного обеспечения, но также успешно справляется с задачей публикации наборов данных. Например, набор коммуникационных сигналов для автоматической классификации модуляции, представленный на GitHub, включает одиннадцать типов сигналов модуляции одной несущей с различными шумами [16]. Он собран из сигналов с различными отношениями сигнал-шум на основе реальной географической среды. Набор данных содержит 22 000 выборок, и каждая выборка состоит из сигнала во временной области и метки. Изначально информация о наборе была опубликована в журнале Sensors с открытым доступом в 2018 г. в статье исследователей из Китая «Автоматическая классификация модуляции на основе глубокого обучения для беспилотных летательных аппаратов» [17].

Следующий пример – сжатый набор радарных сигналов RadarCommDataset, размещенный на Github. В статье “Multi-task Learning Approach for Automatic Modulation and Wireless Signal Classification”, в частности, говорится, что набор данных о беспроводных сигналах выпущен для публичного использования в соответствии с лицензией некоммерческого использования [18, 19]. Отсутствие существующих многозадачных размеченных наборов данных для машинного обучения в области беспроводной связи и явилось основной причиной выпуска этого набора данных.

Публикация на ресурсе GitHub снимает с разработчиков задачу обеспечения бесперебойной работы файлового сервера и предоставляет ряд дополнительных преимуществ. Одним из преимуществ является контроль версий, который позволяет сохранять различные итерации разработки набора данных. При этом, конечно, GitHub, как универсальное решение, ограничен с точки зрения гибкости организации хранилища – не позволяет реализовать собственные способы структуризации и поиска данных. Другой негативной чертой является ограниченный набор доступных лицензий, который требует согласования со стандартами, принятыми в отдельных странах – в Российской Федерации в данном случае. Традиционным способом публикации данных является публикация на собственном сайте, где разработчики имеют большую творческую свободу по организации хранения и поиска данных. Примером является раздел публикации разнообразных наборов данных общества обработки сигналов IEEE, который показывает, что в большинстве ситуаций такая гибкость оказывается невостребованной, так как все ограничивается простой публикацией файлов с простой древовидной файловой иерархией и без специфических инструментов поиска, применяемых, например, в базах знаний [20].

Описанные рассуждения записаны таблице 1 в рамках сравнительного анализа рассмотренных способов организации доступа с целью выбора способа, который следует применить в проекте, выполняемом авторами статьи.

В табл. 1 рассмотрены следующие характеристики: интерактивность, закрытость исходного кода, гибкость, сложность разработки и доступа,

Табл. 1

Сравнительный анализ подходов к организации доступа к результатам имитации

Подход, пример	Преимущества	Недостатки

Отдельное приложение
1D Fourier basis signal simulation system [4]
	Интерактивность;
	Закрытый код;
	Гибкость.
		Сложность разработки;
		Сложность доступа.

Веб-приложение
UW cosmic spectra [10]; NIST atom spectra [11].
	Интерактивность;
	Закрытый код;
	Гибкость;
	Легкий доступ.
		Сложность разработки;
		Сетевые проблемы;
		Поддержка сервера;
		Безопасность.

Набор данных
RadarCommDataset [16]
	Закрытый код;
	Простота разработки;
	Простота доступа.
		Нет интерактивности;
		Мало гибкости.

безопасность и, в случае веб-приложений, специфика веб-разработки. Интерактивность подразумевает взаимодействие с пользователем, открытость пользовательскому вводу. Гибкость подразумевает возможность точного подбора результатов под конкретную задачу.

Реализация метода имитации сигналов в любом случае приведет к разработке отдельного приложения. Такое приложение отвечает критериям интерактивности и гибкости, защищает исходный код. В промышленных сферах применения имитации сигналов веб-приложения нецелесообразны по причине наличия задержки в сети Интернет. Даже при игнорировании задержки в исследовательских задачах сложности при разработке веб-приложения не оправдывают преимуществ по сравнению с отдельным приложением. Разработка отдельного приложения должна быть намечена на будущее, однако соответствующие трудозатраты могут быть чрезмерными на исследовательском этапе. Вариант набора данных позволяет обмениваться результатами имитации разными методами, элементами исходного кода, сравнивать эффективность разных подходов.

В результате сравнительного анализа было принято решение публиковать данные научных экспериментов в виде набора данных на ресурсе GitHub. Подготовка такого набора предполагает структурирование имеющихся данных, что также необходимо и при разработке объединяющего приложения в его отдельном или Интернет представлении.

2. Структура данных

Как указано выше, при построении набора данных важным становится структура набора – данные должны быть организованы в древовидную иерархию. Используются данные, необходимые для описания процесса гармонической имитации и получаемые в результате его выполнения:

- формулы, задающие функцию спектральной плотности энергии (ФСПЭ);
- исходные параметры: параметры дискретизации N1, N2; периоды сигнала T1, T2; граничные частоты ω1, ω2;
- графики и таблицы значений полученных сигналов, ФСПЭ, теоретической, алгоритмической и экспериментальной автокорреляционных характеристик.

Цель публикации данных – сравнение результатов имитации на одинаковых и сходных параметрах исследователями, обзор результатов имитации третьими лицами, подготовка основы для дальнейшего построения онтологии на наборе данных.

Структурирование данных проводится по экспериментам - папка эксперимента с конкретными начальными значениями содержит соответствующие данные. Параметры эксперимента решено вынести в имена папок для упрощения навигации. Фрагмент полученной иерархии показан на рис. 1.

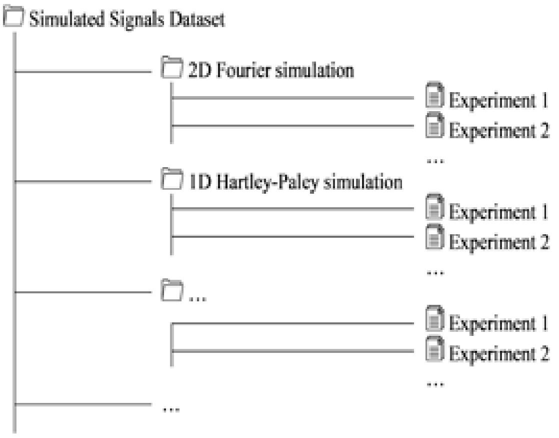


Рис. 1. Структура набора данных сымитированных сигналов

Описанная на рис. 1 структура достаточна для формирования и публикации набора данных и отвечает поставленной цели публикации данных. В целях гомогенизации данных результаты экспериментов решено представить в виде таблиц, содержащих значения, формата «.csv» и рисунков различных форматов. При публикации рисунков приоритет предоставляется векторным рисункам формата «.svg», что мотивировано перспективой применения рисунков в научной деятельности и их публикации, так как векторные рисунки выигрывают по качеству. Рисунки и таблицы располагаются в папках, описывающих эксперименты. Исходные параметры каждого эксперимента указываются в именах соответствующих папок с целью организации автоматизированного поиска. Параметры ω1, ω2 могут быть записаны как w1, w2 для упрощения дальнейшей автоматизации, так как английский алфавит доступнее в средах программирования. Папки экспериментов в свою очередь организуются в разделы, посвященные конкретным методам имитации.

Для организации онтологии на наборе данных предложенной структуры предлагается добавить сущность «задача», описывающую конкретную задачу реального мира [21]. Отношение между сущностями «метод», «эксперимент» и «задача» показано на рис. 2.

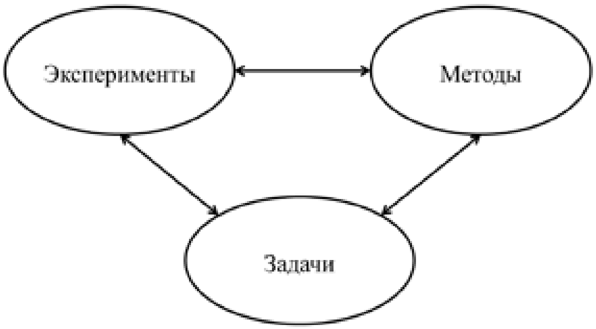


Рис. 2. Взаимоотношение сущностей в онтологии

Сущность «задача» связана с экспериментами, так как эксперименты принимают на вход параметры, соответствующие конкретным задачам реального мира. Также сущность «задача» связана с методами, так как конкретные методы оказываются наиболее полезными при решении определенных задач. Включение сущности «задача» выводит хранилище за пределы обычного набора данных, но позволяет автоматизировать подбор методов имитации под конкретные задачи. Автоматизированный подбор состоит из следующих этапов:

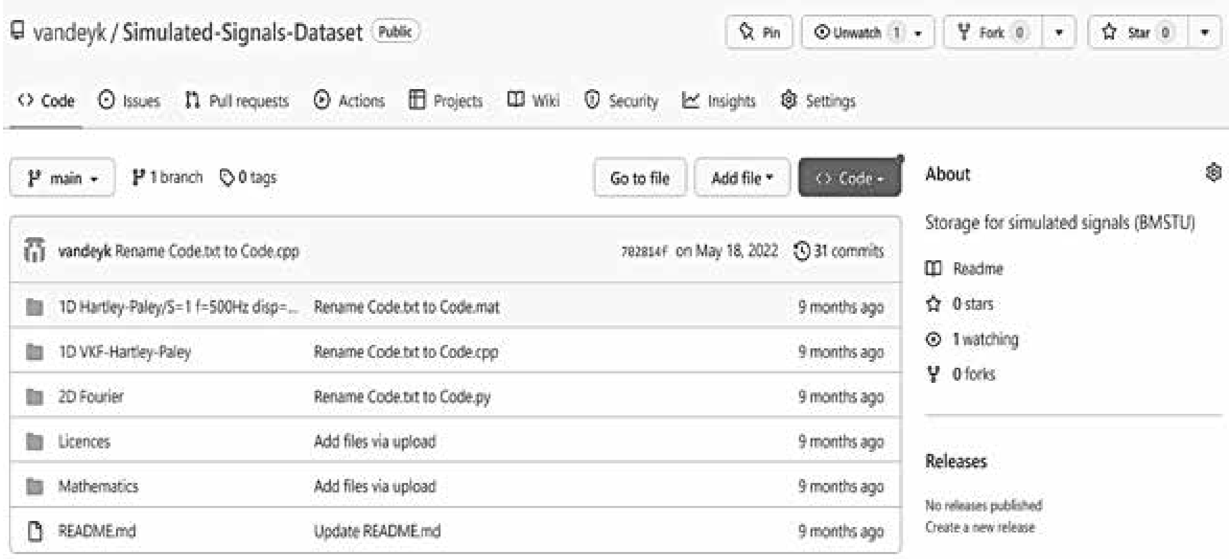
- сопоставление экспериментов и методов путем сравнения значений автокорреляционных функций, определяющего точность эксперимента имитации сигнала;
- сопоставление экспериментов и задач путем вычисления вероятностей значений входных параметров имитации для каждой задачи;
- формирование правила вывода или ассоциации, связывающей задачу и метод, основываясь на предыдущих сравнениях, в целях поддержки решения по применению отдельного метода для решения конкретной задачи.

Обработка больших объемов данных может приводить к необходимости применения методов работы с большими данными. Первые два шага направлены на создание экспертной подсистемы. Формирование правил вывода, характерных для онтологий, должно учитывать возможные неопределенности входных данных в случаях новых задач или ранее не проводимых экспериментов. Законченное хранилище должно представлять собой гибридную интеллектуальную систему, основанную на онтологии, объединяющей экспертные подсистемы.

3. Результаты

Формирование репозитория под названием Simulated-Signals-Dataset произведено на ресурсе GitHub [22]. Итоговый вид верхнего каталога репозитория показан на рис. 3.

Репозиторий содержит разделы, соответствующие трем методам имитации сигналов. Создана отдельная папка для хранения имеющихся в научной команде свидетельств государственной регистрации разработанных программных продуктов. Также имеется раздел, посвященный отдельным фрагментам математической составляющей метода имитации. На данный момент в репозитории наиболее полно представлены результаты экспериментов по имитации двумерных сигналов в базисе Фурье, расположенные в разделе “2D Fourier”. Содержимое раздела показано на рисунке 4. Название каждой папки эксперимента содержит значения всех входных параметров, описывающих этот эксперимент.



vandeyk / Simulated-Signals-Dataset Public

1 branch 0 tags

vandeyk Rename Code.txt to Code.cpp 782814f on May 18, 2022 31 commits

1D Hartley-Paley/S=1 f=500Hz disp=...	Rename Code.txt to Code.mat	9 months ago
1D VKF-Hartley-Paley	Rename Code.txt to Code.cpp	9 months ago
2D Fourier	Rename Code.txt to Code.py	9 months ago
Licences	Add files via upload	9 months ago
Mathematics	Add files via upload	9 months ago
README.md	Update README.md	9 months ago

About

Storage for simulated signals (BMSTU)

Readme

0 stars

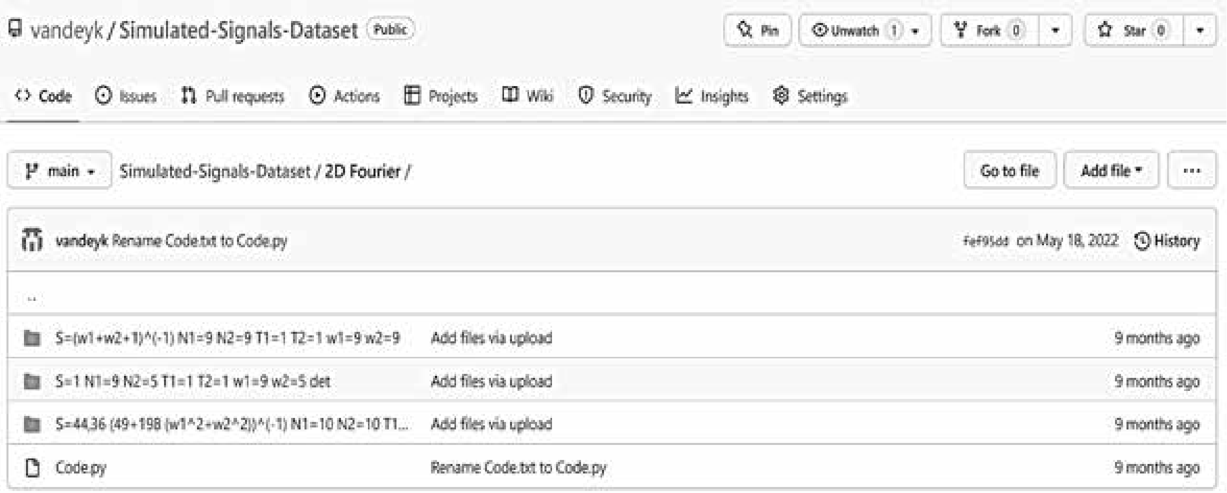
1 watching

0 forks

Releases

No releases published
Create a new release

Рис. 3. Набор сымитированных сигналов на GitHub



Simulated-Signals-Dataset / 2D Fourier /

vandeyk Rename Code.txt to Code.py fef95dd on May 18, 2022

..		
S=(w1+w2+1)^(-1) N1=9 N2=9 T1=1 T2=1 w1=9 w2=9	Add files via upload	9 months ago
S=1 N1=9 N2=5 T1=1 T2=1 w1=9 w2=5 det	Add files via upload	9 months ago
S=44,36 (49+198 (w1^2+w2^2))^(-1) N1=10 N2=10 T1...	Add files via upload	9 months ago
Code.py	Rename Code.txt to Code.py	9 months ago

Рис. 4. Эксперименты по имитации в двумерном базисе Фурье в наборе данных

В разделах, посвященных отдельным методам, хранятся кроме экспериментов фрагменты программного кода, реализующего эти методы. Планируется расширение имеющегося раздела репозитория, а также добавление других методов имитации и их результатов.

Заключение

В рамках проекта с помощью разработанного ПО были получены результаты имитации сигналов, сформированные наборы сигналов были опубликованы на GitHub [22]. В результате работы научная команда и третьи лица могут получить доступ к научным результатам. Статья описывает процесс выбора подхода к организации доступа, процесс построения структуры хранилища и сам набор данных. Приведены принципы построения онтологии на основе созданного набора данных. В будущем планируется расширить и усовершенствовать набор данных, дополнить его новыми результатами, разработать отдельное приложение для работы с различными методами гармонической имитации сигналов. Будет проводиться разработка онтологии на основе существующего набора данных.

Литература

1. *Smirnova, E., Syuzev, V., Gurenko, V., Alekhin, V.* Software system's usage for multidimensional signal's simulation as an engineering staff training tool. INTED2020, pp. 6270-6279. 2020.
2. *Суятинов, С. И.* Синергетическая модель ситуационной осведомленности человека-оператора в эргатических системах управления подвижными объектами / С. И. Суятинов, Т. И. Булдакова, Ю. А. Вишневская // Мехатроника, автоматизация, управление. – 2022. – Т. 23. – № 6. – С. 302-308. – DOI 10.17587/mau.23.302-308.
3. *Liu, Y., Li, J., Sun, S., Yu, B.* Advances in Gaussian random field generation: A review. Computational Geosciences. 2019.
4. *Сюзев В.В.* Алгоритмы многомерного имитационного моделирования случайных процессов / В. В. Сюзев, Е. В. Смирнова, А. В. Пролетарский // Компьютерная оптика. – 2021. – Т. 45. – № 4. – С. 627-637. – DOI 10.18287/2412-6179-CO-770.
5. База Федерального Института Промышленной Собственности [сайт]: свидетельство о регистрации программы для ЭВМ RU 2020610822. URL: https://new.fips.ru/registers-doc-view/fips_servlet?DB=EVM&DocNumber=2020610822&TypeFile=html (дата обращения: 01.06.2022).
6. База Федерального Института Промышленной Собственности [сайт]: свидетельство о регистрации программы для ЭВМ RU 2019610689. URL: https://new.fips.ru/registers-doc-view/fips_servlet?DB=EVM&DocNumber=2019610689&TypeFile=html (дата обращения: 01.06.2022).
7. База Федерального Института Промышленной Собственности [сайт]: свидетельство о регистрации программы для ЭВМ RU 2017619635. URL: https://new.fips.ru/registers-doc-view/fips_servlet?DB=EVM&DocNumber=2017619635&TypeFile=html (дата обращения: 01.06.2022).
8. База Федерального Института Промышленной Собственности [сайт]: свидетельство

о регистрации программы для ЭВМ RU 2017619554. URL: https://new.fips.ru/registers-doc-view/fips_servlet?DB=EVM&DocNumber=2017619554&TypeFile=html (дата обращения: 01.06.2022).

9. *Black, P., Flater, D., Bojanova, I.* Algorithms and Data Structures for New Models of Computation. IT Professional. 23. 9-15. 2021.
10. *Alan, I., Arslan, E., Kosar, T.* Energy-aware data transfer algorithms. The International Conference for High Performance Computing, Networking, Storage and Analysis. 1-12. 2015.
11. *Smirnova, E.V., Syuzev, V.V., Gurenko, V.V., Bychkov, B.* Spectral signal simulation as a scientific and practical task in the training of engineers. INTED2019, pp. 4511-4516. 2019.
12. *Smirnova, E., Syuzev, V., Samarev, R., Deykin, I., Proletarsky, A.* High-Dimensional Simulation Processes in New Energy Theory: Experimental Research (Extended Abstract). Data Analytics and Management in Data Intensive Domains : Extended Abstracts of the XXII International Conference DAMDID, pp. 160-163. Voronezh State University, Voronezh. 2020.
13. University of Washington [сайт]: spectral databases and tools. URL: http://depts.washington.edu/naivpl/content/spectral-databases-and-tools (дата обращения 31.05.2022).
14. National Institute of Standards and Technology [сайт]: atomic spectra database. URL: https://www.nist.gov/pml/atomic-spectra-database (дата обращения 31.05.2022).
15. *Beg, M., Belin, J., Kluyver, T., Konovalov, A., Ragan-Kelley, M., Thiéry, N., Fangohr, H.* Using Jupyter for Reproducible Scientific Workflows. Computing in Science & Engineering, 1-11. 2021.
16. Github [сайт]: набор данных «Communication Signal Dataset». URL: https://github.com/bczhangbczhang/Communication-Signal-Dataset (дата обращения 31.05.2022).
17. *Zhang, D., Ding, W., Zhang, B., Xie, C., Li, H., Liu, C., Han, J.* Automatic Modulation Classification Based on Deep Learning for Unmanned Aerial Vehicles. Sensors. 18, 924. 2018.
18. *Jagannath, A., Jagannath, J.* Multi-task Learning Approach for Modulation and Wireless Signal Classification. ICC2021, pp. 1–7. IEEE, New Jersey. 2021.
19. Github [сайт]: набор данных «Radar and communication signal dataset and modulation classification dataset». URL: https://github.com/ANDROComputationalSolutions/RadarCommDataset (дата обращения 31.05.2022).
20. Signal processing society [сайт]: dataset resources. URL: https://signalprocessingsociety.org/publications-resources/dataset-resources (дата обращения 31.05.2022).
21. *Skvortsov, N., Stupnikov, S.* Managing Data-Intensive Research Problem-Solving Lifecycle. In: Sychev, A., Makhortov, S., Thalheim, B. (eds) Data Analytics and Management in Data Intensive Domains. DAMDID/RCDL 2020. Communications in Computer and Information Science, vol 1427, pp. 29 – 33. Springer, Cham. 2021.
22. Github [сайт]: BMSTU simulated signals dataset. URL: https://github.com/vandeyk/Simulated-Signals-Dataset (дата обращения 31.05.2022).
23. Российский научный фонд [сайт]: карточка проекта № 22-11-00049 «Разработка корреляционной теории моделирования многомерных сигналов и процессов в гибридных системах искусственного интеллекта реального времени». URL: https://rscf.ru/project/22-11-00049/ (дата обращения 5.09.2022).

Дейкин Иван Игоревич. Федеральное государственное бюджетное образовательное учреждение высшего образования «Московский государственный технический университет имени Н.Э. Баумана (национальный исследовательский университет)». Аспирант. Количество печатных работ 10. Область научных интересов: цифровая обработка сигналов. E-mail: deykinii@student.bmstu.ru (ответственный за переписку).

Сюзев Владимир Васильевич. Федеральное государственное бюджетное образовательное учреждение высшего образования «Московский государственный технический университет имени Н.Э. Баумана (национальный исследовательский университет)». Доктор технических наук, профессор кафедры ИУ6. Количество печатных работ: более 120 (в т.ч. 5 монографий). Область научных интересов: цифровая обработка сигналов, многомерные структуры, разработка новой энергетической теории математического представления и преобразования моделей, сигналов и процессов в системах управления динамическими объектами на основе имитационного моделирования сигналов. E-mail: k_iu6@bmstu.ru

Смирнова Елена Валентиновна. Федеральное государственное бюджетное образовательное учреждение высшего образования «Московский государственный технический университет имени Н.Э. Баумана (национальный исследовательский университет)». Доктор технических наук, профессор кафедры ИУ6. Количество печатных работ: более 80 (в т.ч. 3 монографии). Область научных интересов: цифровая обработка биомедицинских сигналов, интернет-программирование, качество инженерного образования. E-mail: evsmirnova@bmstu.ru

Пролетарский Андрей Викторович. Федеральное государственное бюджетное образовательное учреждение высшего образования «Московский государственный технический университет имени Н.Э. Баумана (национальный исследовательский университет)». Доктор технических наук, профессор кафедры ИУ6, заведующий кафедрой ИУ6, декан факультета ИУ. Количество печатных работ: более 200 (в т.ч. 5 монографий). Область научных интересов: интеллектуальные системы и искусственный интеллект, мониторинг и управление динамическими системами, большие данные. E-mail: pav@bmstu.ru

Simulated reference signals storage: basis for fast algorithms creation

I.I. Deykin, V.V. Syuzev, E.V. Smirnova, A.V. Proletarsky

Bauman Moscow State Technical University, Moscow, Russia

Abstract. The paper is devoted to organizing a storage structure for multidimensional reference signals and processes within the framework of digital filtering based on fast simulation algorithms and a universal adaptive matrix transformation operator. The storage contains relevant simulation equations, simulated signals as value tables and images, and original code in Python, C++ and other languages, as well as a supplementary license data. Different approaches to storage creation are analyzed in the paper. The chosen route consists of structuring all the data into a dataset, publishing the dataset on GitHub and building an ontology based upon the dataset. The principles for structuring the data and organizing such an ontology are given. The resulting published dataset is presented. The work is supported by the Russian Scientific Fund (Project #22-11-00049).

Keywords: *data Storage, Signal Storage, Multidimensional Signals, Signal Simulation, Digital Signal Processing, DSP, Ontology, Harmonic Methods.*

DOI: 10.14357/20790279230105

References

1. *Smirnova, E., Syuzev, V., Gurenko, V., Alekhin, V.* Software system's usage for multidimensional signal's simulation as an engineering staff training tool. INTED2020, pp. 6270-6279. 2020.
2. *Suyatinov, S., Buldakova, T., Vishnevskaya, Y.* Building a model of situational awareness of a human operator based on the principles of synergetics. Mathematical Methods in Technologies and Technics 9, 92-96. 2021.
3. *Liu, Y., Li, J., Sun, S., Yu, B.* Advances in Gaussian random field generation: A review. Computational Geosciences. 2019.
4. *Syuzev, V.V., Smirnova, E.V., Proletarsky, A.V.* Algorithms of multidimensional random process simulation. Computer Optics 45(4), 627–637. 2021.
5. Computer program registration certificate RU 2020610822 at the Federal Institute of Industrial Property database. Available at: https://new.fips.ru/registers-doc-view/fips_servlet?DB=EVM&DocNumber=2020610822&TypeFile=html (accessed June 1, 2022)
6. Computer program registration certificate RU 2019610689 at the Federal Institute of Industrial Property database. Available at: https://new.fips.ru/registers-doc-view/fips_servlet?DB =EVM&DocNumber =2019610689&TypeFile=html (accessed June 1, 2022)
7. Computer program registration certificate RU 2017619635 at the Federal Institute of Industrial Property database. Available at: https://new.fips.ru/registers-doc-view/fips_servlet?DB =EVM &DocNumber=2017619635&TypeFile=html (accessed June 1, 2022)
8. Computer program registration certificate RU 2017619554 at the Federal Institute of Industrial Property database. Available at: https://new.fips.ru/registers-doc-view/fips_servlet?DB=EVM &DocNumber=2017619554&TypeFile=html (accessed June 1, 2022)
9. *Black, P., Flater, D., Bojanova, I.* Algorithms and Data Structures for New Models of Computation. IT Professional. 23. 9-15. 2021.

10. *Alan, I., Arslan, E., Kosar, T.* Energy-aware data transfer algorithms. The International Conference for High Performance Computing, Networking, Storage and Analysis. 1-12. 2015.
11. *Smirnova, E.V., Syuzev, V.V., Gurenko, V.V., Bychkov, B.* Spectral signal simulation as a scientific and practical task in the training of engineers. INTED2019, pp. 4511-4516. 2019.
12. *Smirnova, E., Syuzev, V., Samarev, R., Deykin, I., Proletarsky, A.* High-Dimensional Simulation Processes in New Energy Theory: Experimental Research (Extended Abstract). Data Analytics and Management in Data Intensive Domains : Extended Abstracts of the XXII International Conference DAMDID, pp. 160-163. Voronezh State University, Voronezh. 2020.
13. University of Washington spectral databases and tools. Available at: http://depts.washington.edu/naivpl/content/spectral-databases-and-tools (accessed May 31, 2022).
14. National Institute of Standards and Technology atomic spectra database. Available at: https://www.nist.gov/pml/atomic-spectra-database (accessed May 31, 2022).
15. *Beg, M., Belin, J., Kluyver, T., Konovalov, A., Ragan-Kelley, M., Thiéry, N., Fangohr, H.* Using Jupyter for Reproducible Scientific Workflows. Computing in Science & Engineering, 1-11 2021.
16. Communication Signal Dataset on Github. Available at: https://github.com/bczhangbczhang/Communication-Signal-Dataset (accessed May 31, 2022).
17. *Zhang, D., Ding, W., Zhang, B., Xie, C., Li, H., Liu, C., Han, J.* Automatic Modulation Classification Based on Deep Learning for Unmanned Aerial Vehicles. Sensors. 18, 924. 2018.
18. *Jagannath, A., Jagannath, J.* Multi-task Learning Approach for Modulation and Wireless Signal Classification. ICC2021, pp. 1–7. IEEE, New Jersey. 2021.
19. Radar and communication signal dataset and modulation classification dataset on GitHub. Available at: https://github.com/ANDROComputationalSolutions/RadarCommDataset (accessed May 31, 2022).
20. Signal processing society dataset resources. Available at: https://signalprocessingsociety.org/publications-resources/dataset-resources (accessed May 31, 2022).
21. *Skvortsov, N., Stupnikov, S.* Managing Data-Intensive Research Problem-Solving Lifecycle. In: Sychev, A., Makhortov, S., Thalheim, B. (eds) Data Analytics and Management in Data Intensive Domains. DAMDID/RCDL 2020. Communications in Computer and Information Science, vol 1427, pp. 29 – 33. Springer, Cham. 2021.
22. BMSTU simulated signals dataset on GitHub. Available at: https://github.com/vandeyk/Simulated-Signals-Dataset (accessed May 31, 2022).
23. Information about the project “Correlation theory creation for multidimensional signals and processes simulation in real-time hybrid artificial intelligence system” on Russian Science Foundation web-site. Available at: https://rscf.ru/project/22-11-00049/ (accessed September 5, 2022).

Deykin I.I. Post-graduate student, Bauman Moscow State Technical University, 2-nd Baumanskaya, 5, Moscow, 105005, Russia, e-mail: ideykin@mail.ru .

Syuzev V.V. Doctor of technical sciences, full professor, Bauman Moscow State Technical University, 2-nd Baumanskaya, 5, Moscow, 105005, Russia, e-mail: k_iu6@bmstu.ru

Smirnova E.V. Doctor of technical sciences, full professor, Bauman Moscow State Technical University, 2-nd Baumanskaya, 5, Moscow, 105005, Russia, e-mail: evsmirnova@bmstu.ru

Proletarsky A.V. Head of the Scientific and Educational Complex “Informatics and Control Systems”, head of Computer Systems and Networks department, doctor of technical sciences, full professor, Bauman Moscow State Technical University, 2-nd Baumanskaya, 5, Moscow, 105005, Russia, e-mail: pav@bmstu.ru

Интеллектуальный анализ данных

AutoML: исследование существующих программных реализаций и определение общей внутренней структуры решений

И.А. Попова Г.И. Ревунков, Ю.Е. Гапанюк

Федеральное государственное бюджетное образовательное учреждение высшего образования «Московский государственный технический университет имени Н.Э. Баумана (национальный исследовательский университет)», г. Москва, Россия

Аннотация. В статье рассматриваются различные программные реализации автоматизации процесса машинного обучения для решения задачи регрессии. Рассмотрено внутреннее устройство и возможности ряда существующих и обширно используемых инструментов автоматизированного машинного обучения таких, как LightAutoML (LAMA), TPOT, Auto-Sklearn, H2O AutoML, MLJAR. Возможности данных программных систем были исследованы для решения задачи регрессии на нескольких наборах данных. В результате исследования была выведена общая структура программного решения автоматизированного машинного обучения, которая может быть взята за основу при дальнейшем проектировании и разработке собственного программного продукта, а также проанализирована точность, с которой системы предсказывали значения целевого признака.

Ключевые слова: *автоматизированное машинное обучение, LAMA, TPOT, Auto-Sklearn, H2O AutoML, MLJAR, регрессия.*

DOI: 10.14357/20790279230106

Введение

AutoML является современной тенденцией в сфере машинного обучения. Данное направление активно исследуется научным сообществом, что подтверждается существованием и разработкой различных программных систем автоматизированного машинного обучения.

Специалисты в области машинного обучения регулярно сталкиваются с задачей выбора подходящего алгоритма с оптимальными гиперпараметрами для описания исследуемого набора данных. Для этого они обычно выполняют и оценивают множество конфигураций методом проб и ошибок. Однако для начинающих аналитиков данных это трудоемкая задача. Последние достижения в области исследований AutoML решают эту проблему путем автоматического поиска подходящего алгоритма с соответствующими гиперпараметрами. Основной задачей автоматизированного машинного обучения (AutoML) является автоматический поиск алгоритмов предварительной обработки входных данных и обучения выбранной системой модели с наилучшей производительностью обобщения на целевой (обрабатываемой) выборке.

Современные системы AutoML помогают автоматизировать практически весь процесс построения модели. На данный момент подлежат автоматизации следующие этапы моделирования: подготовка данных, обзор признаков, построение модели машинного обучения и оптимизация гиперпараметров, валидация построенной модели, построение отчетов, внедрение разработанной модели.

Алгоритмы машинного обучения для обработки и анализа данных комбинируются в конвейер, то есть применяются последовательно к обрабатываемой выборке. Далее происходит оптимизация этих алгоритмов путем подбора моделей и значений их гиперпараметров, причем большинство подходов выполняют оптимизацию всех параметров модели одновременно.

Однако стоит отметить, что системы автоматизированного машинного обучения широко применимы только для решения задач машинного обучения с учителем, то есть классификации и регрессии [1].

Каждая система отличается внутренней архитектурой, рядом возможностей, которые может использовать конечный пользователь и набором решаемых задач. Соответственно существующие решения по-разному могут справляться с одной и той же задачей машинного обучения.

На сегодняшний день направление AutoML развивается очень быстро, о чем свидетельствует большое количество публикаций. Примером соответствующих исследований являются [1, 3, 4, 7, 16, 17, 18]. Эти статьи не рассматривают производительность систем AutoML в отношении решения проблемы регрессии. Они не сравнивают окончательные модели, построенные системами на наборах данных, предназначенных для регрессии. Некоторые статьи [1, 3, 4, 7] описывают общие принципы концепции AutoML, но не учитывают конкретные системы AutoML либо авторы очень кратко приводят информацию в сравнительной таблице о некоторых системах AutoML. В работах [4, 7] даны очень краткие сведения о некоторых системах AutoML, которые более подробно рассмотрены в нашей статье (например, TPOT, H2O AutoML, Auto-Sklearn). В нашем исследовании даны рекомендации относительно применения рассмотренных систем AutoML при решении практических проблем. Вклад нашей работы заключается в том, что мы исследовали работу современных систем AutoML на примере задачи регрессии, описали подробно используемые системы, а также рассмотрели отечественное решение – LightAutoML в сравнении с существующими современными системами AutoML. Мы сравнили точность предсказания модели линейной регрессии с предсказаниями моделей, предложенных сложными системами AutoML. Чтобы оценить точность конечной модели, мы используем метрики MAE, RMSE, MedAE и R^2.

1. Цель исследования

В данной статье мы проведем сравнительный анализ ряда существующих программных систем автоматизированного машинного обучения: Light AutoML (LAMA), TPOT, Auto-Sklearn, H2O AutoML, MLJAR. Рассмотрим эффективность работы выбранных систем при решении задачи регрессии в нескольких предметных областях, используя ряд метрик точности (RMSE, MAE, MedAE, R^2).

2. Описание задачи

Основной задачей AutoML является автоматизация составления композиции моделей машинного обучения и дальнейшая параметризация подобранных алгоритмов, чтобы максимизировать значение выбранной метрики точности [4].

В данной статье сфокусируемся на использовании различных систем AutoML для решения задачи машинного обучения с учителем – линейной регрессии.

В процессе разработки модели машинного обучения с учителем необходимо предоставить на вход алгоритму набор признаков $\chi \subseteq \mathbb{R}^d$ и целевую переменную y. То есть модель обучается на образцах с известным значением целевой переменной y [3].

В сценариях машинного обучения используются подготовленные для последующего анализа и обработки статистические данные. Набор данных $D \subset \{(x, y) \mid x \in \chi, y \in Y\}$ является конечным отношением между пространством экземпляров и пространством меток, и мы обозначаем как D множество всех возможных наборов данных. Разработанная модель должна выдавать точные результаты для любых новых образцов данных. То есть, обобщающая способность является важнейшим свойством аналитической модели, приобретаемым в процессе обучения.

Зачастую процесс разработки модели машинного обучения представляет собой итеративный цикл обработки данных, обучения модели и ее оценки. Для того, чтобы получить на выходе удовлетворительную производительность модели, необходимо детально экспериментировать с различными комбинациями методов обработки данных, алгоритмов модели и гиперпараметров. Данный процесс достаточно затратен по времени и требует от исполнителя хороших знаний в области анализа данных.

Каждый этап приведенного на рис. 1 цикла разработки модели можно автоматизировать. При

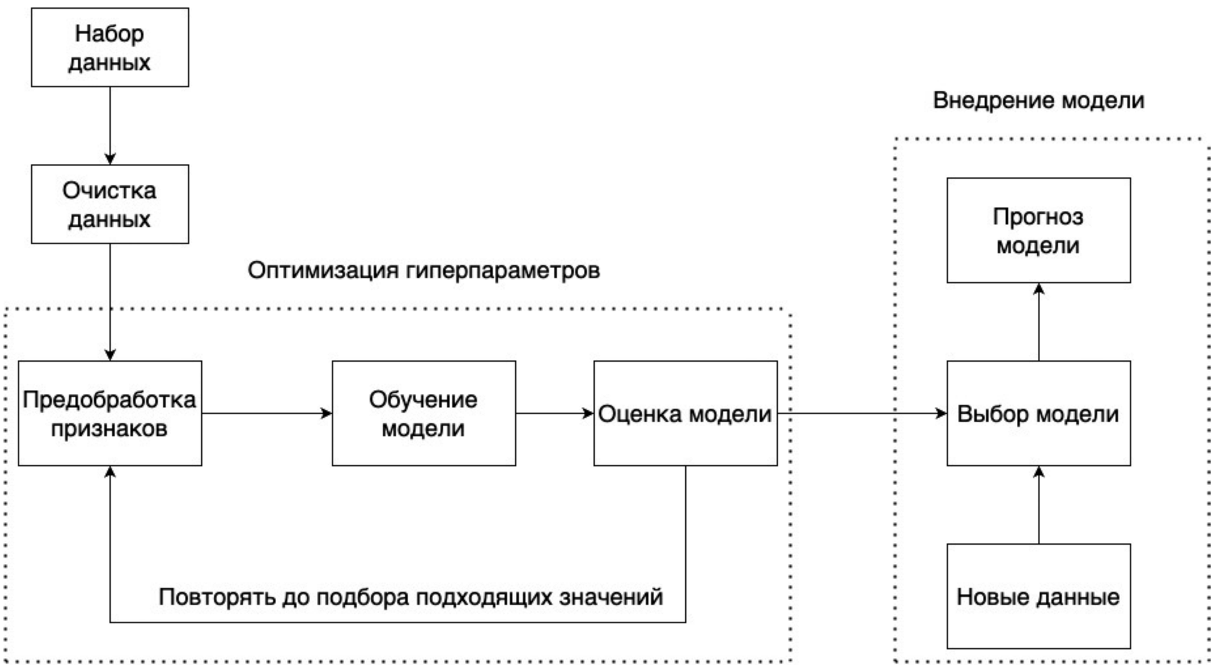


Рис. 1. Обобщенный цикл разработки модели машинного обучения

этом движение данных между модулями системы будет напоминать конвейер, состоящий из множества модулей: модуля обработки данных, выбора модели, построения отчетов, внедрения и т.д.

Основная цель машинного обучения с учителем состоит в том, чтобы найти конвейер, который минимизирует ошибку прогнозирования, усредненную по всем экземплярам выборки. Зачастую для оценки точности построенной модели используются такие метрики, как коэффициент детерминации (R^2), среднеквадратичная ошибка (RMSE), медианная абсолютная ошибка (MedAE), средняя абсолютная ошибка (MAE), AUC ROC, F1, log-loss и другие [7].

У каждой системы автоматизированного машинного обучения поиск и оптимизация алгоритмов осуществляется по-разному, о чем свидетельствуют различия во внутреннем устройстве систем. Исследование внутренней архитектуры систем автоматизированного машинного обучения позволит выявить общие закономерности в их построении, обнаружить места, которые могут быть усилены алгоритмами машинного обучения либо оптимизированы для обеспечения лучшей производительности решения.

3. Конвейер AutoML

Задачу AutoML можно сформулировать как процесс поиска *f*, который лучше обобщается в любом возможном *T* с минимальным вмешательством пользователя. Где *f* может быть композицией нескольких функций, которые могут преобразовывать входное пространство признаков, обрабатывать входные данные, строить модели и т. д. Такую композицию в контексте AutoML называют конвейером, который можно формализовать (формула 1).

$$f(x) = \upsilon_{\theta_\upsilon}(T_{\theta_T}(x)) \quad (1)$$

В данной формуле υ – модель машинного обучения, T – механизм преобразования признаков с гиперпараметрами θ_υ и θ_T соответственно [2]. Причем каждая из этих моделей может быть композицией других моделей.

Как показано на рис. 2 конвейер AutoML состоит из множества процессов обработки данных: предварительная обработка выборки, отбор признаков, генерация модели и ее оценка.

Первый этап включает в себя «добычу» информации, которая помогает повысить производительность сгенерированных моделей, создавая дополнительную информацию для изучения. На втором этапе к набору данных применяются различные методы работы с признаками: устранение пропусков в данных, кодирование категориальных признаков, масштабирование и нормализация признаков, обработка выбросов в данных. Далее подготовленные данные подаются на вход модулей, которые занимаются поиском наиболее эффективной архитектуры модели. Построение модели машинного обучения можно разделить на поиск пространства решений и оптимизацию гиперпараметров. Пространство поиска определяет принципы проектирования моделей машинного обучения, которые можно разделить на две категории: традиционные модели (Linear Regression, SVM) и нейронные сети.

Методы оптимизации подразделяются на гиперпараметрическую оптимизацию (HPO) и оптимизацию архитектуры модели (AO). HPO оптимизирует параметры, связанные с обучением (например, скорость обучения и размер пакета), а AO оптимизирует параметры, связанные с моделью (например, количество слоев для нейронных архитектур и количество соседей для модели KNN) [5]. Завершающим этапом является тестирование производительности и выбор лучшей модели.

Зачастую первая часть конвейера не является последовательностью операций обработки данных, она имеет древовидную структуру с несколькими параллельными препроцессорами, которые затем объединяются. Оптимальный конвейер может быть реализован как строгий порядок, в котором должны применяться различные алгоритмы обработки данных, а также использоваться не более одного препроцессора каждого типа.

4. Обзор современных систем AutoML

Прогресс в области AutoML привел к появлению множества систем, которые автоматизируют проектирование и разработку моделей машинного обучения с учителем на разных этапах.

Рассмотрим ряд систем AutoML, которые используются в современных проектах машинного обучения.

4.1. LightAutoML

LightAutoML (LAMA) представляет собой решение с открытым исходным кодом, разработанное Sber AI Lab. Данная система позволяет автоматизи-

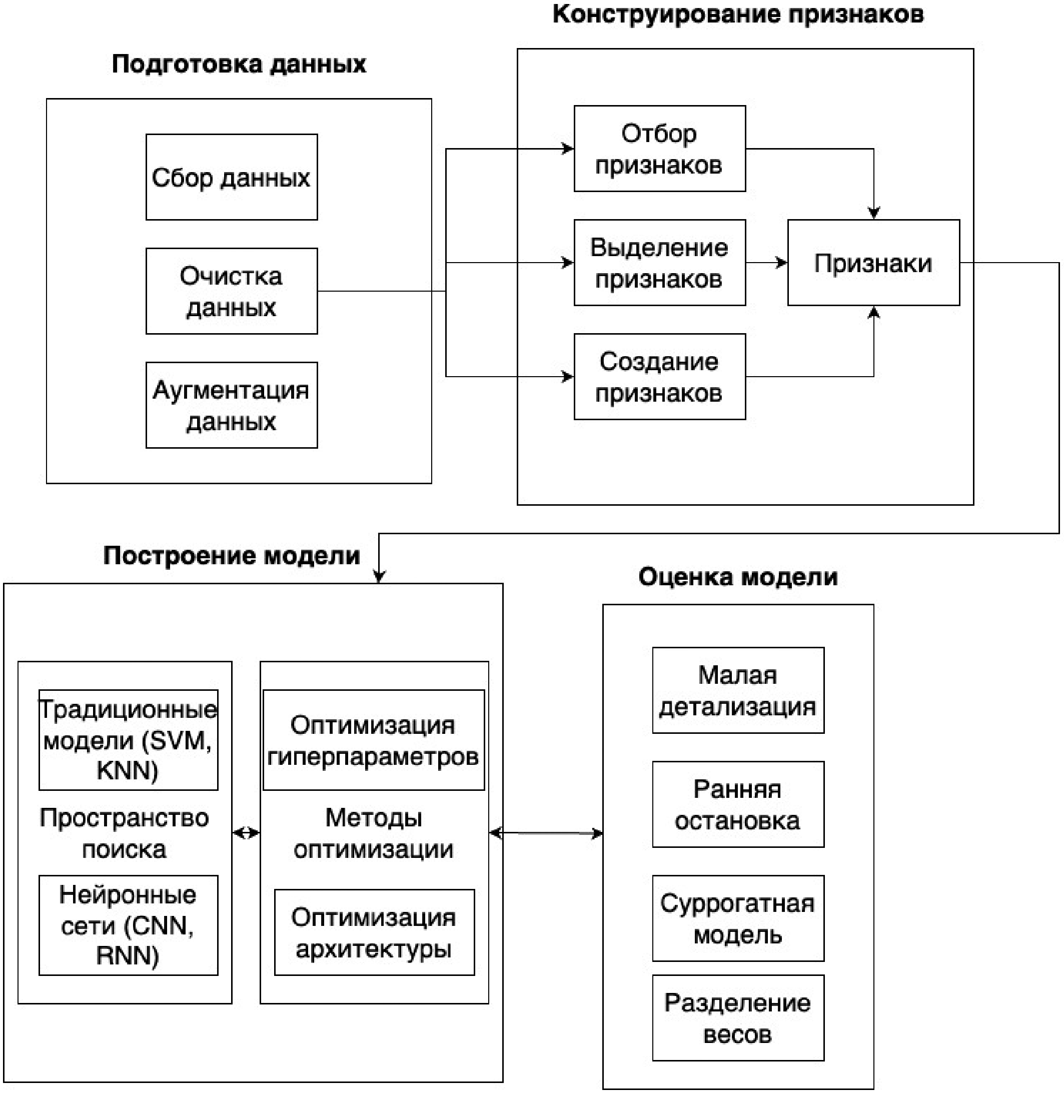


Рис. 2. Общая схема конвейера AutoML системы

ровать процесс построения модели для следующих задач: бинарная классификация, многоклассовая классификация, регрессия.

Разработанный конвейер дает возможность выполнять следующие операции: обрабатывать данные и автоматически настраивать гиперпараметры, строить отчеты, описывающие процесс разработки модели, конструировать собственные конвейеры из предоставляемых системой модулей, запускать модель в режиме предсказания.

Предлагаемая система LAMA работает только с двумя типами моделей - gradient boosted decision trees (GBMs) и линейными моделями, что значительно сокращает время без ущерба для производительности для решаемых типов задач и данных. LAMA состоит из модулей (пресетов), которые ориентированы на end-to-end разработку моделей для типичных задач ML. В настоящее время LightAutoML поддерживает следующие четыре предустановленных модуля:

1. TabularAutoML – фокусируется на классических задачах ML, работает с табличными наборами данных;
2. WhiteBox – решает задачу бинарной классификации с использованием простых интерпретируемых алгоритмов (логистическая регрессия);
3. NLP – способен комбинировать конвейер для обработки табличных данных с алгоритмами NLP (специальные средства извлечения признаков или предварительно обученные модели глубокого обучения);
4. CV – модуль для обработки изображений.

На рис. 3 представлен конвейер, который реализован в системе LightAutoML. Обязательными элементами конвейера являются:

- считыватель, который принимает на вход данные и выполняет их предварительную обработку, также на вход подается задача, которая подлежит решению с помощью системы;
- наборы данных, которые содержат метаданные и используются для валидации результатов; несколько конвейеров ML, которые складываются и / или усредняются с помощью Blender для получения единого прогноза.

Конвейеры ML могут быть вычислены независимо на одних и тех же наборах данных, а затем объединены вместе с использованием усреднения (или взвешенного усреднения).

4.2. TPOT

TPOT (Tree-based Pipeline Optimization Tool) является проектом на Python с открытым исходным кодом. TPOT автоматизирует конвейеры машинного обучения с помощью генетического программирования (GP), хорошо известного метода автоматического построения программ. В данной системе конвейер машинного обучения полностью автоматизирован и для определения оптимальной модели применяется генетический алгоритм.

Основное внимание в этом проекте уделяется обучению с учителем, а именно задаче классификации с поддержкой ста пятидесяти алгоритмов ScikitLearn [4], включая алгоритмы предварительной обработки. Система, как и Auto-Sklearn

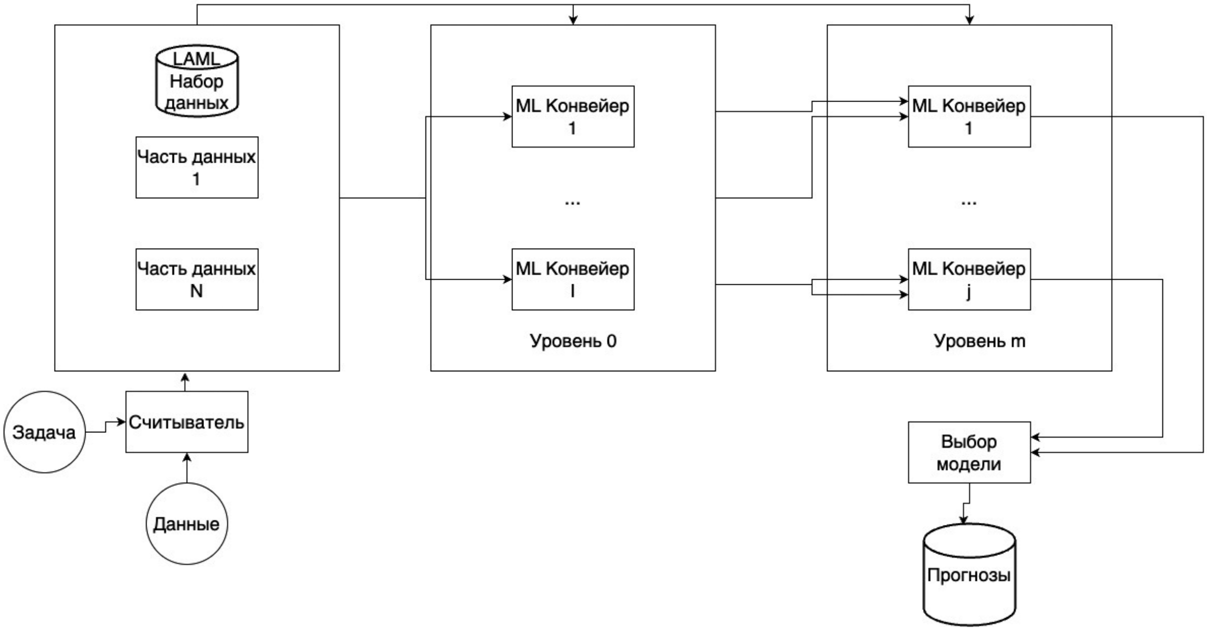


Рис. 3 Общая схема конвейера системы LAMA.

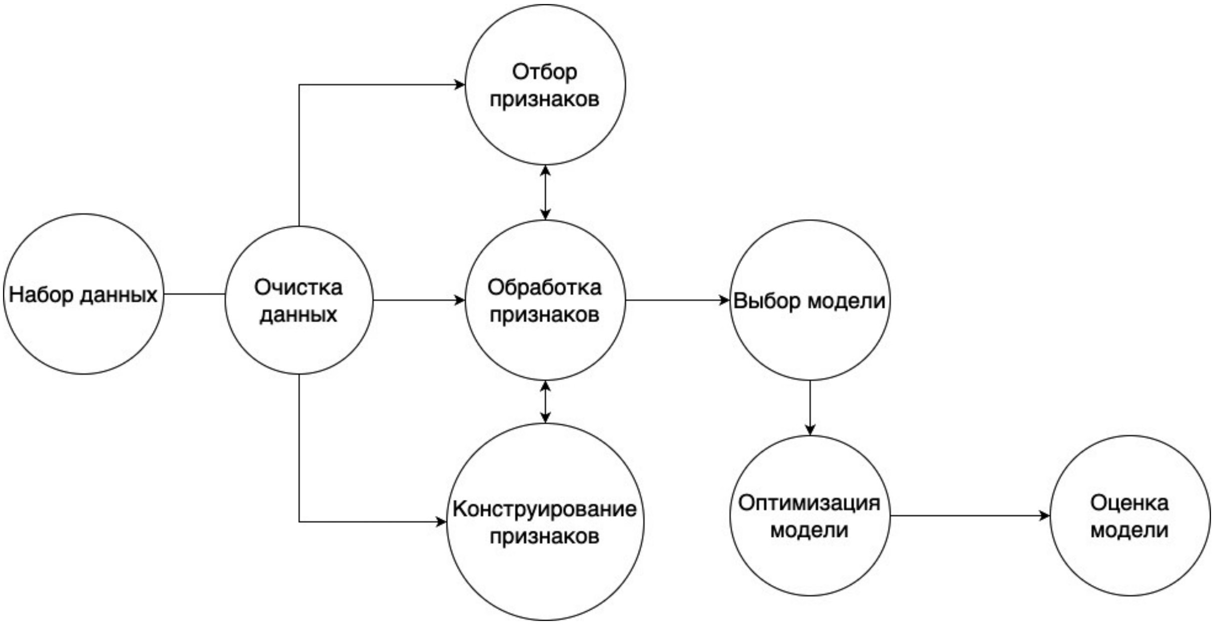


Рис. 4. Общая схема конвейера TPOT.

представляет собой надстройку над библиотекой Sklearn, однако использует собственные алгоритмы регрессии и классификации.

Отобраны двадцать лучших конвейеров с точки зрения максимальной точности перекрестной проверки и минимизации количества процессов, которые видоизменяются для создания конвейеров нового поколения. Каждый из конвейеров производит еще пять с помощью перекрестных методов или случайных вставок и усадок. Алгоритм выполняется для 100 поколений для каждого из обновляемых решений.

На рис. 4 представлен пример конвейера TPOT.

Система сохраняет копию набора данных и распараллеливает процесс обработки признаков и снижения размерности данных.

На данный момент TPOT не может работать с естественным языком и категориальными признаками.

4.3. Auto-Sklearn

Система написана на Python и использует алгоритмы и методы из программной библиотеки Scikit-Learn (15 алгоритмов классификации, 14 методов обработки признаков, 4 метода предварительной обработки данных) [10]. Auto-Sklearn реализует алгоритм SMAC для настройки гиперпараметров. Система умеет генерировать признаки, выбирать модель, настраивать гиперпараметры. Auto-Sklearn представляет два новшества: использование метаобучения для запуска процедуры байесовской оптимизации и шаг построения ансамбля, который использует более одной конфигурации, найденной в процедуре оптимизации.

Auto-Sklearn запускает процедуру байесовской оптимизации, предоставляя начальные экземпляры из конфигураций, которые дали наилучшие результаты для аналогичных наборов данных [6].

На рис. 5 представлен пример конвейера Auto-Sklearn. Настройка гиперпараметров и предварительная обработка данных частично выполняются с помощью метаобучения. Результаты метаобучения предопределяют пространство поиска. Конвейер AutoML использует это пространство поиска итеративно, начиная с процессора данных, за которым следует препроцессор обработки признаков, за которым следует обучение классификатора либо регрессора. Результаты оцениваются, и гиперпараметры оптимизируются с помощью байесовского оптимизатора. Зачастую лучшие значения производительности показывают ансамблевые модели.

4.4. H2O AutoML

H2O является системой машинного обучения с открытым исходным кодом, с распределенной памятью и линейной масштабируемостью. Текущая версия системы AutoML может обучать и выполнять кросс-валидацию для случайного леса, градиентного бустинга, глубоких нейронных сетей, а затем обучать составной ансамбль, используя все модели. H2O AutoML автоматизирует процесс построения большого количества моделей, чтобы выявить наиболее эффективную модель.

Ключевые особенности H2O AutoML: среда с открытым исходным кодом, которая предоставляет распределенные реализации методов машинного обучения; система реализована на Java, однако есть API для таких языков, как Python, R, Scala, а также

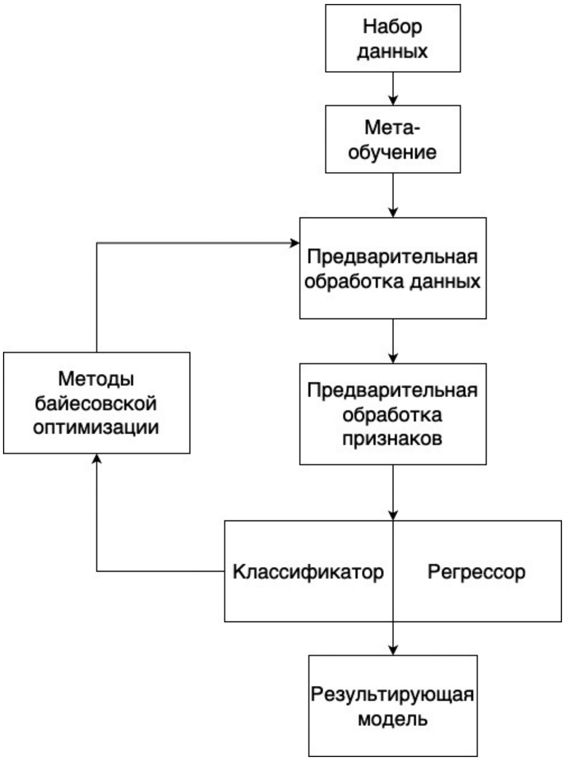

Набор данных
Мета-обучение
Предварительная обработка данных
Методы байесовской оптимизации
Предварительная обработка признаков
Классификатор
Регрессор
Результирующая модель

Рис. 5. Общая схема конвейера Auto-Sklearn

доступен веб-интерфейс; H2O AutoML работает на таких платформах, как Hadoop, Spark, AWS.

4.5. MLJAR

Систему AutoML MLJAR можно использовать для создания полного конвейера машинного обучения с конструированием признаков и настройкой гиперпараметров. MLJAR поддерживает следующие алгоритмы машинного обучения: нейронные сети, XGBoost, Catboost, LightGBM и другие. MLJAR строит несколько моделей в зависимости от выбранных алгоритмов и рассчитывает окончательный прогноз путем объединения в ансамбль или стекинга моделей.

На рис. 6 представлен пример конвейера MLJAR.

4.6. Сравнительный анализ AutoML систем

Системы AutoML различаются по своему внутреннему устройству и, соответственно, функциям, которые они предоставляют пользователю. Выделим отличительные особенности рассмотренных систем и дадим краткие рекомендации по их применению:

- Система LightAutoML эффективна для решения таких задач, как бинарная или многоклассовая классификация, а также регрессия, где входные данные могут содержать одновременно различные типы признаков: числа, тексты, категориальные данные, даты. Можно применять LightAutoML в качестве инструмента быстрой проверки гипотез, а также построения моделей машинного обучения, которые будут описаны с помощью линейных моделей и деревьев решений с градиентным бустингом. Также возможно расширять систему собственными модулями обработки данных, таким образом настраивая LightAutoML под решение новых задач.
- Система TPOT не может обрабатывать пропуски в наборе данных, а также не может работать с нечисловыми признаками. Пользователь должен предварительно обработать данные, только потом подавать их в систему. На предварительно обработанных данных TPOT позволяет строить модели классификации или регрессии, используя алгоритмы Sklearn. Поэтому данная система подойдет вам, если вы работаете с табличными данными, не содержащими пропущенные значения, а также все признаки объектов выборки являются числовыми. Иначе требуется предварительная экспертиза и обработка данных.
- Система H2O AutoML предоставляет метод импорта файлов, который позволяет загружать табличные данные, состоящие из категориальных и числовых признаков, а затем, используя внутреннюю эвристику, делит данные на подвыборки для обучения и тестирования целевой модели. Данная система автоматизирует такие этапы, как предварительная обработка данных, обучение и настройка модели, объединение различных моделей, чтобы выбрать модели с наилучшей производительностью (зачастую данные описываются ансамблями GBM, GLM, DNN моделей). H2O предоставляет удобный

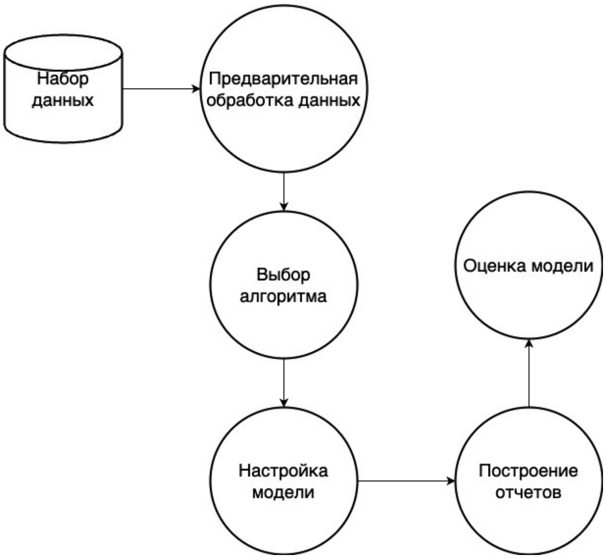


Рис. 6. Общая схема конвейера MLJAR

пользовательский интерфейс H2O Flow – интерактивную веб-среду, которая позволяет совмещать выполнение кода, математические вычисления, графики и мультимедиа в одном документе. Использовать данную систему рекомендуется, если выборка содержит пропуски, категориальные и числовые признаки, также если необходимо проанализировать предложенные системой модели, представленные в виде таблицы с метриками качества (RMSE, MSE, MAE и др.).

- Система MLJAR довольно быстро справляется с построением и обучением целевой модели на выборках разного размера. Она может создавать отчеты с помощью разметки Markdown, содержащие сведения о процессе обучения моделей, работая со множеством разных моделей машинного обучения, обрабатывать пропуски в данных и работать с большим количеством типов признаков. Если вам нужно быстро построить модель машинного обучения на данных, содержащих пропуски и большое количество признаков различных типов, то вы можете выбрать MLJAR.
- Auto-Sklearn не может обрабатывать пропущенные значения, но быстро справляется с задачей выбора самой эффективной модели машинного обучения. Эта система построена поверх алгоритмов машинного обучения из библиотеки Sklearn. Auto-Sklearn сочетает в себе методы, которые помогут создать модель с настроенными гиперпараметрами, но пользователю придется предварительно предобработать входные данные. Auto-Sklearn использует байесовские методы оптимизации для поиска наиболее производительного конвейера для заданного набора данных, поэтому вычисления даже на больших наборах данных будут производиться достаточно быстро.

5. Исследование систем AutoML для задачи регрессии

Мы сравнили рассмотренные выше системы AutoML, решая задачу регрессии на нескольких наборах данных, и оценили при помощи ряда метрик точности конечные результаты работы систем.

Для экспериментальной части исследования были выбраны два набора данных:

1) cars – выборка, содержащая 301 объект, в качестве целевого признака выступает цена автомобиля;
2) powerplant - набор данных, содержащий параметры 9568 предприятий, где в качестве целевого признака выступает почасовая выработка электроэнергии (МВт/ч).

Набор данных cars опубликован на ресурсе Kaggle [11], а набор данных plants опубликован в репозитории машинного обучения UCI [12].

Исходные данные были обработаны 5 системами автоматизированного обучения: LAMA, TPOT, Auto-Sklearn, H2O AutoML, MLJAR.

В итоге данными системами был построен конвейер для обработки данных, определены статистические модели и настроены их гиперпараметры. Затем нами была произведена оценка точности каждой итоговой модели.

Опишем метрики, которые были использованы для оценки точности конечной модели, построенной выбранными системами AutoML:

1. Средняя абсолютная ошибка (MAE). Эта метрика не чувствительна к выбросам в наборе данных, но она не нормирована.
2. Среднеквадратическая ошибка (RMSE). Данная метрика чувствительна к выбросам.
3. Средняя абсолютная ошибка (MedAE). Данная метрика не чувствительна к выбросам в наборе данных.
4. Коэффициент детерминации (R^2). Выбросы существенно влияют на коэффициент детерминации.

Соответственно, чем ближе значение метрик к нулю, тем точнее работает модель (кроме метрики R^2, которая принимает значения в диапазоне от 0 до 1, чем ближе значение R^2 к единице, тем точнее модель).

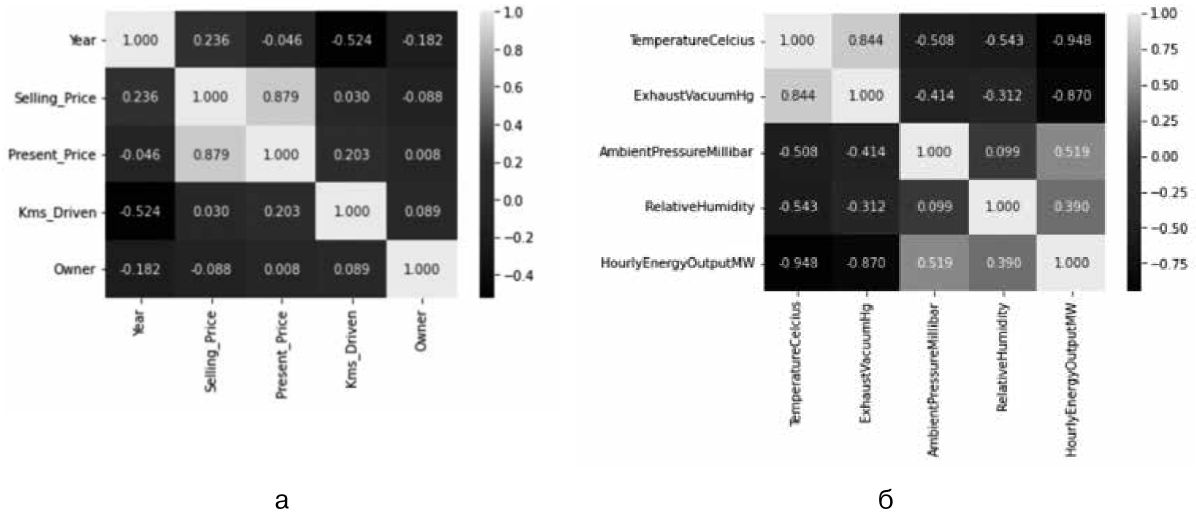
В работе используются библиотечные реализации описанных выше метрик, которые предоставляет модуль metrics пакета sklearn.

5.1. Анализ выборок

Для того, чтобы предварительно оценить внутреннюю организацию данных и построить гипотезы, визуализируем экспериментальные выборки.

Рис. 7а и 7б отображают корреляционную матрицу, которая содержит коэффициенты корреляции между парами признаков из набора данных. Матрица корреляции содержит на главной диагонали единицы, то есть является симметричной. В данном случае при расчете матрицы корреляции использовался коэффициент корреляции Пирсона. Для удобного восприятия корреляционной матрицы используем «тепловую карту», где при помощи цветов окрашены ячейки, содержащие коэффициенты корреляции.

Для набора данных cars на целевой признак «Selling Price» больше всего влияет признак «Present Price» (коэффициент корреляции равен 0,844). В наборе данных plants признак «ExhaustVacuumHg» коррелирует с целевым признаком «HourlyEnergyOutputMW» с коэффициен-



а б

Рис. 7. Корреляционные карты

том -0,87. В данном случае важен модуль коэффициента корреляции, который показывает силу зависимости между признаками, отрицательный знак указывает на обратную связь между признаками, что для рассматриваемой задачи не имеет значения.

Проанализировав матрицу корреляции, можно сделать вывод, что данные в выбранных датасетах можно аппроксимировать прямой линией, следовательно, можно предположить, что модели линейной регрессии будет достаточно, чтобы добиться высокой точности предсказаний.

5.2. Анализ полученных результатов

В результате эксперимента, с помощью описанных выше AutoML систем, были определены модели машинного обучения, наилучшим образом описывающие входные данные. При этом каждой из систем AutoML был построен конвейер, который включал в себя все этапы обработки данных, подбора модели и настройки ее гиперпараметров.

На рис. 8-9 представлены гистограммы, на которых можно увидеть значения погрешности предсказания целевого признака для каждой исследуемой в работе AutoML системы. На рисунке 8-9 представлены значения метрик качества для тестовой выборки набора данных cars. Можно заметить, что наиболее точной оказалась модель линейной регрессии по оценкам MAE, RMSE, MedAE, также хороших результатов позволила добиться модель, выбранная системой MLJAR (метрики MAE, RMSE, MedAE).

На рис. 8 показаны показатели качества для тестовой выборки набора данных об автомобилях. Видно, что модель линейной регрессии по оценкам MAE, RMSE и MedAE оказалась наиболее точной,

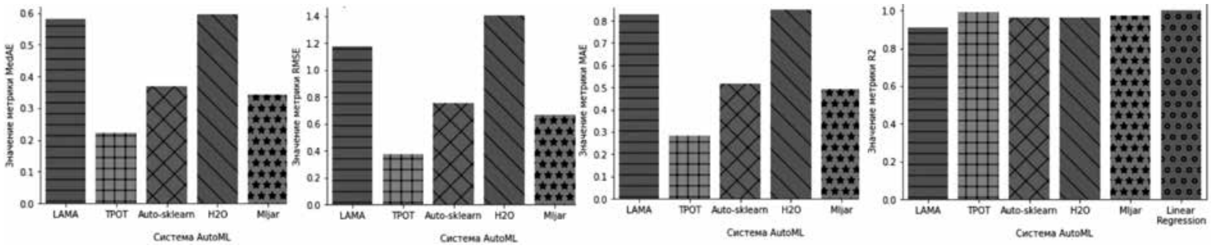


Рис. 8. Значения метрик для набора данных cars

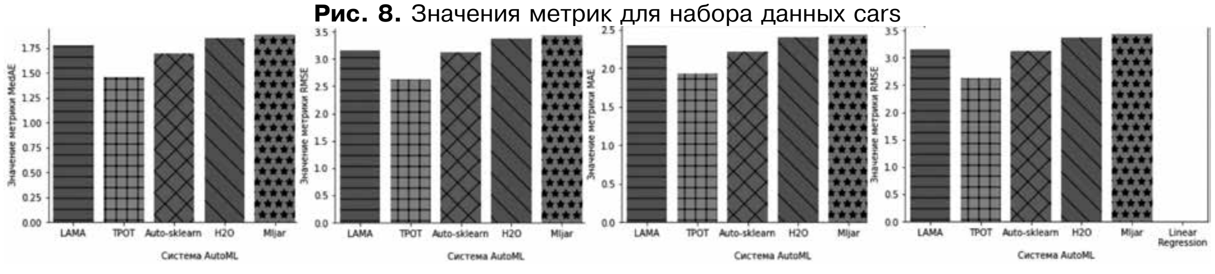


Рис. 9. Значения метрик для набора данных plants

а модель, выбранная по системе MLJAR (метрики MAE, RMSE, MedAE), также показала хорошие результаты. Поскольку мы используем 4 метрики для оценки точности окончательной модели машинного обучения, лучшая система AutoML определяется с помощью анализа Парето.

На рис. 9 показаны оценки качества для тестовой выборки набора данных об автомобилях. Видно, что модель линейной регрессии по оценкам MAE и RMSE оказалась наиболее точной, а модель, выбранная по системе TPOT (метрики MAE, RMSE и R^2), также показала хорошие результаты.

Выбранные системы показывают совершенно разные результаты при решении задачи регрессии на выборках разного размера. Набор данных автомобилей содержит 301 объект, а набор данных растений содержит 9568 объектов. Наилучшие результаты показывают системы TPOT и MLJAR (если учитывать значения метрик MAE, RMSE, MedAE).

Система LAMA неплохо справляется с задачей подбора и настройки модели, однако автоматизация всего процесса занимает длительное время (в среднем до 5 минут). Это связано с большим количеством различных расчетов, производимых системой из-за обилия предоставляемых ею возможностей.

В табл. 1 показано время, затрачиваемое каждой системой AutoML на выбор и обучение окончательной модели машинного обучения.

Табл. 1

Время работы систем AutoML

Название	Набор данных cars	Набор данных plants
Auto-Sklearn	2 мин 1 сек	1 мин 59 сек
TPOT	31.2 сек	1 мин 30 сек
MLJAR	29.7 сек	54.4 сек
H2O	1 мин 2 сек	1 мин 2 сек
LAMA	3 мин 18 сек	5 мин 40 сек
Linear Regression	19.8 мс	6.78 сек

На основании анализа Табл. 1 можно сделать вывод, что системе LightAutoML потребовалось больше всего времени для построения оптимального конвейера, выбора наилучшей модели и ее обучения. Модель линейной регрессии оказалась самой быстрой для построения, а система MLJAR выполняла вычисления быстрее, чем другие системы.

Заключение

Таким образом, в рамках данной работы формализован процесс оптимизации построения конвейеров данных и настройки алгоритмов машинного обучения. Рассмотрен ряд автоматизированных систем машинного обучения: Light AutoML (LAMA), TPOT, Auto-Sklearn, MLJAR, H2O AutoML. Исследован процесс построения модели регрессии с использованием перечисленных систем для нескольких наборов данных, содержащих категориальные и числовые признаки.

Литература

1. *Nagarajah, T., Poravi, G.* A Review on Automated Machine Learning (AutoML) Systems. In: 2019 IEEE 5th International Conference for Convergence in Technology (I2CT), pp. 1–6. Pune, India (2019). https://doi.org/10.1109/I2CT45611.2019.9033810
2. *Bahri, M., Salutari, F., Putina, A. et al.* AutoML: state of the art with a focus on anomaly detection, challenges, and research directions. Int J Data Sci Anal (2022). https://doi.org/10.1007/s41060-022-00309-0
3. *Karmaker, S., Hassan, M.M., Smith, M.J., Xu, L., Zhai, C., Veeramachaneni, K.* AutoML to Date and Beyond: Challenges and Opportunities. ACM Computing Surveys (CSUR) 54, 1–36 (2022)
4. *He X., Zhao K., Chu X.* AutoML: A survey of the state-of-the-art. Knowl. Based Syst., 212, 106622. https://doi.org/10.1016/j.knosys.2020.106622
5. *Escalante, H.J.* Automated Machine Learning – a brief review at the end of the early years. arXiv:2008.08516. https://doi.org/10.48550/arXiv.2008.08516
6. *Bahri, M., Salutari, F., Putina, A., Sozio, M.* AutoML: state of the art with a focus on anomaly detection, challenges, and research directions. International Journal of Data Science and Analytics, Springer Verlag, 2022. https://doi.org/10.1007/s41060-022-00309-0
7. *Koroteev, M.V.* Review of some modern trends in machine learning technology. E-Management 1(1), 26–35 (2018)
8. *Feurer, M., Eggensperger, K., Falkner, S., Lindauer, M.T., Hutter, F.* Practical Automated Machine Learning for the AutoML Challenge 2018. In: International Workshop on Automatic Machine Learning at ICML, pp. 1189-1232 (2018)
9. Car Dekho Data, https://www.kaggle.com/datasets/shindenikhil/car-dekho-data. Last accessed 12 December 2022
10. Combined Cycle Power Plant Dataset, https://archive.ics.uci.edu/ml/datasets/Combined+Cycle+Power+Plant. Last accessed 12 December 2022
11. LightAutoML – Automatic model creation framework, https://github.com/sb-ailab/

LightAutoML. Last accessed 12 December 2022

12. TPOT – A Python Automated Machine Learning tool, https://github.com/EpistasisLab/tpot. Last accessed 12 December 2022
13. Auto-Sklearn – An automated machine learning toolkit, https://github.com/automl/auto-sklearn. Last accessed 12 December 2022
14. H2O AutoML – Open-Source Automated Machine Learning, https://h2o.ai/platform/h2o-automl/. Last accessed 12 December 2022
15. MLJAR – Automate your Machine Learning pipeline, https://mljar.com/. Last accessed 12 December 2022
16. Chen, Yi-Wei, Qingquan Song, and Xia Hu.: Techniques for automated machine learning. ACM SIGKDD Explorations Newsletter, 35-50 (2021).
17. Elshawi, Radwa, Mohamed Maher, and Sherif Sakr: Automated machine learning: State-of-the-art and open challenges. arXiv preprint arXiv:1906.02287 (2019).
18. V*akhrushev, Anton, et al.* LightAutoML: AutoML Solution for a Large Financial Services Ecosystem. arXiv preprint arXiv:2109.01528 (2021).

Попова Инна Андреевна. Федеральное государственное бюджетное образовательное учреждение высшего образования «Московский государственный технический университет имени Н.Э. Баумана (национальный исследовательский университет)», г. Москва, Россия. Аспирант. Количество печатных работ: 5. Область научных интересов: информационные технологии, машинное обучение. E-mail: popovai1@student.bmstu.ru

Гапанюк Юрий Евгеньевич. Федеральное государственное бюджетное образовательное учреждение высшего образования «Московский государственный технический университет имени Н.Э. Баумана (национальный исследовательский университет)», г. Москва, Россия. Доцент. Количество печатных работ: более 100. Область научных интересов: проектирование автоматизированных систем, проектирование гибридных интеллектуальных информационных систем, сложные графовые модели. E-mail: gapyu@bmstu.ru (ответственный за переписку)

Ревунков Георгий Иванович. Федеральное государственное бюджетное образовательное учреждение высшего образования «Московский государственный технический университет имени Н.Э. Баумана (национальный исследовательский университет)», г. Москва, Россия. Доцент. Количество печатных работ: более 100 (3 монографии). Область научных интересов: проектирование автоматизированных систем, проектирование гибридных интеллектуальных информационных систем, базы данных. E-mail: revunkov@bmstu.ru

AutoML: Examining Existing Software Implementations and Determining the Overall Internal Structure of Solutions

I.A. Popova , G.I. Revunkov, Y.E. Gapanyuk

Bauman Moscow State Technical University, Moscow, Russia

Abstract. The article discusses various software implementations of the process of automating the task of using machine learning to solve the linear regression problem. The internal structure and capabilities of a number of existing and widely used automated machine learning tools such as LightAutoML (LAMA), TPOT, Auto-Sklearn, H2O AutoML, MLJAR are considered. The capabilities of these software systems have been explored to solve the regression problem on multiple datasets.

Keywords: *automated machine learning (AutoML), LAMA, TPOT, Auto-Sklearn, H2O AutoML, MLJAR, Regression.*

DOI: 10.14357/20790279230106

References

1. *Nagarajah, T., Poravi, G.* A Review on Automated Machine Learning (AutoML) Systems. In: 2019 IEEE 5th International Conference for Convergence in Technology (I2CT), pp. 1–6. Pune, India (2019). https://doi.org/10.1109/I2CT45611.2019.9033810
2. *Bahri, M., Salutari, F., Putina, A. et al.* AutoML: state of the art with a focus on anomaly detection, challenges, and research directions. Int J Data Sci Anal (2022). https://doi.org/10.1007/s41060-022-00309-0

3. *Karmaker, S., Hassan, M.M., Smith, M.J., Xu, L., Zhai, C., Veeramachaneni, K.* AutoML to Date and Beyond: Challenges and Opportunities. ACM Computing Surveys (CSUR) 54, 1–36 (2022)
4. *He X., Zhao K., Chu X.* AutoML: A survey of the state-of-the-art. Knowl. Based Syst., 212, 106622. https://doi.org/10.1016/j.knosys.2020.106622
5. *Escalante, H.J.* Automated Machine Learning – a brief review at the end of the early years. arXiv:2008.08516. https://doi.org/10.48550/arXiv.2008.08516
6. *Bahri, M., Salutari, F., Putina, A., Sozio, M.* AutoML: state of the art with a focus on anomaly detection, challenges, and research directions. International Journal of Data Science and Analytics, Springer Verlag, 2022. https://doi.org/10.1007/s41060-022-00309-0
7. *Koroteev, M.V.* Review of some modern trends in machine learning technology. E-Management 1(1), 26–35 (2018)
8. *Feurer, M., Eggensperger, K., Falkner, S., Lindauer, M.T., Hutter, F.* Practical Automated Machine Learning for the AutoML Challenge 2018. In: International Workshop on Automatic Machine Learning at ICML, pp. 1189-1232 (2018)
9. Car Dekho Data, https://www.kaggle.com/datasets/shindenikhil/car-dekho-data. Last accessed 12 December 2022
10. Combined Cycle Power Plant Dataset, https://archive.ics.uci.edu/ml/datasets/Combined+Cycle+Power+Plant. Last accessed 12 December 2022
11. LightAutoML – Automatic model creation framework, https://github.com/sb-ailab/LightAutoML. Last accessed 12 December 2022
12. TPOT – A Python Automated Machine Learning tool, https://github.com/EpistasisLab/tpot. Last accessed 12 December 2022
13. Auto-Sklearn – An automated machine learning toolkit, https://github.com/automl/auto-sklearn. Last accessed 12 December 2022
14. H2O AutoML – Open-Source Automated Machine Learning, https://h2o.ai/platform/h2o-automl/. Last accessed 12 December 2022
15. MLJAR – Automate your Machine Learning pipeline, https://mljar.com/. Last accessed 12 December 2022
16. Chen, Yi-Wei, Qingquan Song, and Xia Hu.: Techniques for automated machine learning. ACM SIGKDD Explorations Newsletter, 35-50 (2021).
17. Elshawi, Radwa, Mohamed Maher, and Sherif Sakr: Automated machine learning: State-of-the-art and open challenges. arXiv preprint arXiv:1906.02287 (2019).
18. V*akhrushev, Anton, et al.* LightAutoML: AutoML Solution for a Large Financial Services Ecosystem. arXiv preprint arXiv:2109.01528 (2021).

Popova Inna Andreevna. Graduate student, Federal state budgetary institution of higher professional education «Bauman Moscow State Technical University», Moscow, Russia. Number of publications: 5. Research interests: information technology, machine learning. E-mail: popovai1@student.bmstu.ru

Gapanyuk Yuriy Evgenievich. Associate professor, Federal state budgetary institution of higher professional education «Bauman Moscow State Technical University», Moscow, Russia. Number of publications: about 100. Research interests: designing of automated systems, designing of hybrid intelligent information systems, complex graph models. E-mail: gapyu@bmstu.ru

Revunkov Georgiy Ivanovich. Associate professor, Federal state budgetary institution of higher professional education «Bauman Moscow State Technical University», Moscow, Russia. Number of publications: about 100 (3 monographs). Research interests: designing of automated systems, designing of hybrid intelligent information systems, database systems. E-mail: revunkov@bmstu.ru

Web Application with GUI for Data Analysis Automation*

M.S. Manakhova[I], V.A. Dudarev[II]

[I] National Research University Higher School of Economics, Moscow, Russia
[II] A.A. Baikov Institute of Metallurgy and Materials Science of RAS, Moscow, Russia

Abstract. In the current digital age, the world has a huge amount of data. Therefore, people are more and more confronted with the use of such methods as data analysis and machine learning. Moreover, many people are considering using machine learning algorithms for their own purposes. However, data analysis is a complex process that can hardly be carried out by people who do not have sufficient knowledge both in this field and in programming. This paper presents an approach to give non-expert users the ability to apply machine learning algorithms to their datasets using an application with a graphical interface. There are a lot of challenges involved in creating ML-solutions, even if we take advantage of existing ML-algorithms: feature engineering, outliers' detection, filling the missing values, ML-method's hyperparameters optimization and so on. The main point of the research is to find a balance in solving these complex tasks and to provide a Web-based user interface for unexperienced people to enable them to utilize the power of ML-methods in automatic or semi-automatic way. The practical outcome is an information system development, that consists of three interrelated parts: a web application, an API and several microservices that implement ML-algorithms from Scikit-learn library.

Keywords: *web Application, Graphical User Interface, data analysis, ML automation.*
DOI: 10.14357/20790279230107

Introduction

Over the last decade, the level of application of machine learning has grown considerably [1]. Nowadays, it is widely used in the areas such as medicine, marketing, finance, retail, logistics, robotics and so on [2, 3]. Today, almost all companies that collect and store large amounts of data have entire departments involved in data analysis and machine learning. In addition, small teams or even individuals are also interested in using machine learning, because they find data analysis techniques useful for developing their projects and research. For example, scientists who could use the existing data to predict possible values of certain variables or coefficients for conducting scientific experiments.

Simple data analysis such as graph plotting, creating charts and calculating some statistical coefficients can be done using the standard tools such as Microsoft Excel. But often more complex methods involving the use of machine learning models are required. In this case, not only a comprehensive knowledge of linear algebra, mathematical analysis and statistics is required, but also understanding of machine learning concepts and coding experience, particularly in languages such as Python or R, as well as special frameworks and libraries are needed.

Recently, there is a growing number of automated machine learning solutions that allow training and using models for making predictions without writing code in a programming language but using visualization or other methods of user interaction. Nevertheless, some of these solutions, such as Google AutoML have been developed for users with significant experience and are mostly intended to simplify and speed up the work of machine learning engineers and data scientists rather than to lower the entry threshold into this area for non-expert users. Most of the other solutions that can be successfully used by people with basic domain knowledge are often designed to solve only some specific classes of problems, such as computer vision or text analysis.

This paper presents an approach that allows users with basic machine learning knowledge to train and

* This work was performed under partial financial support from the Russian Foundation for Basic Research, projects 20-01-00609 and 21-51-53019, and according to the State Order no. 075-00715-22-00.

use different models for obtaining predictions on the tabular datasets using a graphical interface. The problems encountered in creating systems for automated machine learning and approaches to solve them are also discussed. This work aims to develop a client-server web application with a graphical user interface for data analysis, especially for the supervised learning problems such as regression and classification.

The rest of this paper is organized as follows. Section 2 discusses related work and common challenges included in machine learning process. Then, Section 3 describes the algorithms and methods used in the implementation. Section 4 contains details about the implementation of the web application. Finally, Section 5 provides the conclusion of the work.

1. Literature review

As previously stated, the popularity of solutions for automated machine learning is growing. The paper [4] discusses the possible challenges involved in creating these kinds of solutions.

One of the most challenging problems is the process of extracting features from raw data, called feature engineering. This process often involves handling missing values, encoding of categorical variables, feature scaling and selection. It is notable that feature engineering is not always a simple task even for experts in machine learning and it is usually performed manually by empirical methods.

Missing values are one of the most common problems when it comes to preparing data for use with machine learning models. Human errors, privacy issues, and other factors can lead to the absence of values in the dataset. However, missing values require a correct handling, since most of the models in the existing machine learning libraries are not capable to work with the data that contain gaps.

In fact, there is no clear answer regarding how to handle missing values in datasets. Although, quite several papers discussing the problem of filling gaps in a tabular data have been published, most of the proposed solutions were developed for specific tasks and may not be as useful in other cases. The standard way to solve this problem is to remove features that contain many missing values or to fill in the missing values with some estimate based on other values of the same feature [5].

Most implementations of machine learning algorithms also require all values in the dataset to be represented in a numerical form. Because of this, categorical data must be converted to numerical values before being used for model training or prediction. The simplest way to encode categorical features is label encoding (ordinal encoding). The main idea behind this method is to associate each category with some integer number. This method of encoding is rarely used in practice because of its big disadvantage: it can add additional dependencies that did not exist in the original data, which is critical for linear models, and in general can lead to wrong interpretations of feature values.

One-hot encoding or dummy encoding is the modification that solves the problem of the previously discussed approach. In this algorithm, a new variable, sometimes called a dummy variable, is created for each category of a feature, where a binary value (0 or 1) denotes that a particular feature belongs to a certain category. The main problem with this method and its existing modifications is that a new attribute is created for each unique value of a category variable. Thus, the number of attributes grows quadratically, so naive encoding is only applicable when categorical variables contain a small number of unique values. Another problem of naive encoding is that it produces large number of binary features in the dataset, which can significantly reduce the quality of models when using tree-based algorithms (e.g., decision trees or random forest). In this regard, this algorithm is not suitable for use in systems for automated machine learning, because the number of unique values in the categorical variables in datasets can be quite large.

Another approach to encode categorical variables is target encoding. The idea of target encoding is to use the statistics of a target variable to encode a categorical value. According to the theoretical justification for this approach given in the paper discussing categorical feature preprocessing scheme [6], the key transformation used in this method is a transformation that maps each value of a categorical feature to an estimate of the probability of occurrence of the target variable.

When categorical features are encoded in the training sample, the numerical representation of the categorical feature corresponds to the posterior probability of occurrence of the target feature, provided that the categorical variable takes particular value. In the case of categorical feature encoding in the prediction sample, the numerical representation corresponds to the expected value of the categorical variable. Essentially, this means that for each category, the average value of the target variable is calculated, with which the category is subsequently encoded. This method works for both binary classification and regression. For multiclass classification a similar technique is used, where a categorical variable is encoded with $m - 1$ new variables, where m is the number of classes. It should be noted, that although the author of the article claims that the statistics of the target variable are used to encode the categorical variable, in fact only its mean value is used. Although the mean value is a sufficient statistic for binary classification, it is not suitable for regression because it disregards the intra-category variance of the target variable. In this regard, this algorithm in its pure form is prone to overfitting, so some modifications are often added to software implementations to reduce the probability of overfitting.

In most cases, not all the variables in the raw datasets are useful when building machine learning models. Using many redundant features may reduce the accuracy of the predictions and the generalization ability of the model, as well as dramatically increase its overall complexity.

The paper [7] contains a detailed review of the existing methods for feature selection. As stated, there are three main types of feature selection techniques: filtering, wrapper, and embedded models. Filtering methods are a general set of methods that do not involve the use of a specific machine learning algorithm. They are based on probability theory and statistical approaches and include visual analysis (e.g., construction of a correlation matrix to identify the features that have a weak correlation with the target variable), evaluation of features using some statistical criterion (variance, correlation, χ^2, etc.), and feature ranking by significance. In filtering methods, each feature is considered separately, so it is not possible to identify more complex dependencies in the data, and the resulting subset of features that are most correlated with the target will not always be the subset on which the prediction accuracy will be the highest.

In addition, the existing implementations of these methods often require a choice of a certain threshold value to filter out the redundant features, which is quite difficult to determine automatically, while ensuring equally good quality for datasets that differ in structure (as in the case of automated machine learning systems). In general, these methods are more suitable for a machine learning process fully controlled by user. The main advantage of this class of methods over other feature selection algorithms is a low computational complexity that linearly depends on the total number of features and, consequently, high computation speed. Moreover, filters can be used when the dimensionality of the feature space is larger than the number of observations in the sample, which is not always possible with other methods.

In wrapper methods, the process of feature selection is based on applying some classifier to different subsets of features in the training sample. After selecting the optimal subset, the algorithm is tested on the dataset that was not involved into selection process. This class of methods is divided into two main approaches: forward and backward selection. In the first case, the algorithm starts with an empty subset of features to which, at each iteration, the feature that has the greatest influence on the quality of the model is added. In the second case, the initial subset contains all the attributes of the training sample, from which the least significant attributes are removed at each iteration. In both cases, the process continues until a statistically significant improvement in the quality of the model is obtained (the stopping criterion is reached). Wrapper methods use a greedy search approach to evaluate all possible combinations of features using some evaluation criterion (e.g., p-value and determination coefficient (R^2) for regression; accuracy, precision, recall or F-score for classification), thus having a rather high computational complexity. Another problem with this approach is that the backward selection method cannot be used when the number of features exceeds the number of observations in the training sample.

Embedded methods combine the advantages of filters and wrappers, integrating feature selection into the learning process. The most common embedded methods are based on tree-based algorithms. At each recursion step, some feature is selected, and the sample is divided into smaller subsets. The more child nodes in a subset belong to the same class, the more informative the feature is considered. In classification problems, the partitioning is usually performed either according to the Gini coefficient (index) or using the information gain, which is based on the concepts of entropy and the volume of information. In regression problems, the partitioning is performed by a dispersion value. In addition to tree-based algorithms, regularization approaches are also common. The idea of regularization approaches to construct an algorithm that minimizes not only the model error but also the number of variables used. In such cases, both L1-regularization or L2-regularization and their combinations are used. These regularization methods reduce some model coefficients to zero, which allows removing such features from the model. Embedded methods allow to identify more complex dependencies in datasets and are less prone to overfitting and computationally complex than wrapper methods. Even though embedded methods are still more computationally complex than filtering methods, this class of methods is best suited for automating feature selection.

Another challenge is related to the hyperparameter optimization. Machine learning models often include hyperparameters whose values are very important for achieving high quality models [4]. The hyperparameter optimization algorithms work with the model as with a black box: only the value of the model loss function obtained by training with the considered set of hyperparameters is important, not the algorithm itself. In formalized form the problem of hyperparameter optimization can be written in the following way: let A be the model of the algorithm characterized by hyperparameters $\lambda = \{\lambda_1, \ldots, \lambda_n\}, \lambda_1 \in \Lambda_1, \lambda_n \in \Lambda_n$. Then, the space of hyperparameters associated with it is $\Lambda = \Lambda_1 \cdot \Lambda_2 \cdot \ldots \cdot \Lambda_n$. The goal is to find such set of hyperparameters $\lambda^* \in A$ with which the given model of algorithm A is the most efficient.

Several methods of automatic hyperparameter selection have been proposed by researchers in the field of computer science. As stated in one of the papers discussing the use of automated machine learning [8], the simplest ways to optimize hyperparameters are grid and random search. Grid and random search are uninformed methods, which means that they do not learn any information from previous iterations.

Grid search is a brute-force algorithm in which model is trained and evaluated for a complete set of hyperparameter combinations. Because of this, increase in the size of the hyperparameter search space leads to an exponential rise in computational complexity. Therefore, this algorithm is often an unsuitable choice as it could be inefficient in terms of performance.

In random search a complete set of hyperparameter optimization is replaced by a subset of a randomly chosen length. Since length of a hyperparameter set is less than in grid search, this algorithm requires less computational time, but here comes a risk that the best combination of hyperparameters would not be included in the tested set.

Recently, such method as Bayesian optimization is increasingly used for the hyperparameter optimization. Its major difference from the previously presented approaches is that it is an informed method, so the tuning algorithm optimizes the choice of parameters at each step according to the evaluation of the previous step. In summary, this method creates a probabilistic model which maps hyperparameters to their corresponding estimation probability. Instead of trying complete set or subset of hyperparameters, the Bayesian optimization method can converge to the optimal hyperparameters. Thus, the best hyperparameters can be obtained without examining the entire sample space. However, additional time is required to determine the next hyperparameters to estimate based on the results of previous iterations, so this method could be slower than random search.

2. Chosen algorithms implementations

The algorithm for preparing a dataset for further use in model training consists of four steps:

1. Imputation of missing values.
2. Encoding of categorical features.
3. Feature selection.
4. Data scaling.

Non-categorical features containing more than 50 percent of missing values are deleted. Categorical variables with more than half of the missing values are filled with a special mark. For imputing continuous variables, the k nearest neighbors (kNN) method based implementation from Scikit-learn library [9] called *KNNImputer* is used. Each sample's missing values are imputed using the mean value from 5 nearest neighbors found in the training set. The gaps in the remaining features are filled with the most popular value using *SimpleImputer* from Scikit-learn library.

The rows of the dataset are shuffled randomly before encoding categorical features, as some datasets may be sorted according to the value of the target variable, which can lead to problems when using target encoding algorithm. After random shuffling the text processing algorithm to determine whether the values of the feature are textual representations of integer numbers is applied (e.g., “seven” is converted to 7). Features containing only one unique value are removed since they have low effect on the target variable. If a feature consists of only two different values, then ordinal encoding is applied. In other cases, the target encoding algorithm from CatBoost library called *CatBoostEncoder* is applied.

The features are encoded according to the following formula:

$$\frac{targetSum + prior}{featureCount + 1} \tag{1}$$

where $targetSum$ is a sum of the target value for that particular categorical feature (before the current one), $prior$ is the constant value defined as the ratio of the sum of all values of the target variable in the dataset to the total number of observations, $featureCount$ is the total number of categorical features observed before the current one and having the same value as the current one. With this approach, the first few observations in the dataset always have the statistics of the target feature with much higher variance than the subsequent ones. To reduce this effect, many random permutations of the same data are used to calculate the statistics of the target variable, and the final encoding is calculated by averaging across these permutations.

As previously discussed, the best approach for selecting the most significant features in machine learning systems is the embedded methods, so the Scikit-learn implementation called *SelectFromModel* with *ElasticNet* estimator for regression problems and *DecisionTreeClassifier* estimator for classification problems was chosen. At the first step of the feature selection algorithm, a model based on a training sample is constructed. Then an approach based on feature importance calculation is used. Features are considered unimportant if the corresponding feature importance values are below a given threshold parameter. The threshold is calculated programmatically using the median value of importance of all features multiplied by a constant as a heuristic. At the last step of the algorithm, the features that the algorithm has marked as unsignificant are removed.

Feature scaling is based on Scikit-learn *StandardScaler* which standardize features by removing the mean and scaling to unit variance. The standard score of a sample *x* is calculated as:

$$z = \frac{x - u}{s} \tag{2}$$

where *u* is the mean of the training samples and *s* is the standard deviation of the training samples.

For hyperparameter optimization an implementation of the Bayesian optimization method from Scikit-optimize library called *BayesSearchCV* is used. The choice of Bayesian optimization method was made for reasons of reducing training time and increasing the models' quality. As mentioned earlier, grid search is not a suitable choice for automated machine learning systems because of its high computational complexity, as it can lead to an excessive load on the system when the system is used by a sufficient number of users at the same time. Random search, as stated before, may not find the best hyperparameter combination in a given number of iterations.

To prove the above statements, a couple of experiments with different number of hyperparameter combinations for random search classifier was conducted on the breast cancer dataset (https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data). The results of hyperparameter optimization on 2160 hyperparameter combinations is presented in the Table 1. Default algorithms parameters were not changed. As can be seen, the grid search is about three times slower than the Bayesian optimization and 62 times slower than the random search. However, the grid search algorithm gives the best model score, while the random search gives the worst. Therefore, to achieve a balance between computation time and model quality, *BayesSearchCV* is a suitable choice.

Table 1

Comparison of hyperparameter optimization algorithms from Scikit-learn library.

Algorithm	**F1 score**	**Elapsed time (seconds / s)**
GridSearchCV	0.971874	176.38
RandomizedSearchCV	0.959786	2.83
BayesianSearchCV	0.968036	63.71

For model evaluation the Scikit-learn implementation of Leave-One-Out cross-validation method is used. The advantage of this method is that each object of the sample participates in the control sample exactly once, and the length of the training subsamples is only one less than the length of the full sample. The main disadvantage of this method is high resource intensity, since the learning process is performed as many times as number of objects in the sample. Nevertheless, this method is the most accurate among all cross-validation methods because in most other cases (Hold-Out Validation, *k*-fold cross-validation) the training sample is divided into fewer parts.

The algorithm for preprocessing prediction data is similar to the dataset preparation before model training:

1. Imputation of missing values.
2. Removal of the features that are not presented in the training sample after preprocessing.
3. Encoding of categorical features.
4. Data scaling.

3. Web application implementation details

A web service (source code: https://github.com/sirenescx/fastml-web-application) is a system of five interrelated parts (Fig. 1):

1. Web application with a graphical user interface, which is responsible for processing user actions, data input and output, and forming and sending requests to the API.
2. A microservice (API) which processes incoming requests from a web application and distributes data according to the algorithms selected by user.
3. A microservice for data preprocessing.
4. Set of microservices with regression and classification algorithms for training and prediction.
5. A database used for storing user data and algorithms and microservices' settings.

The microservice architecture was chosen for the following reasons:

1. For implementation of a microservice for data processing and microservices for machine learning, the Python language was chosen because of significant number of tools and libraries for data analysis and machine learning. However, when it comes to creating of web services, Python is not the best choice due to its low performance compared to most other programming languages, as well as the difficulty of testing. The microservice architecture provides the ability to use different technology stacks for different tasks and allows to easily connect services written in different programming languages into a common system.

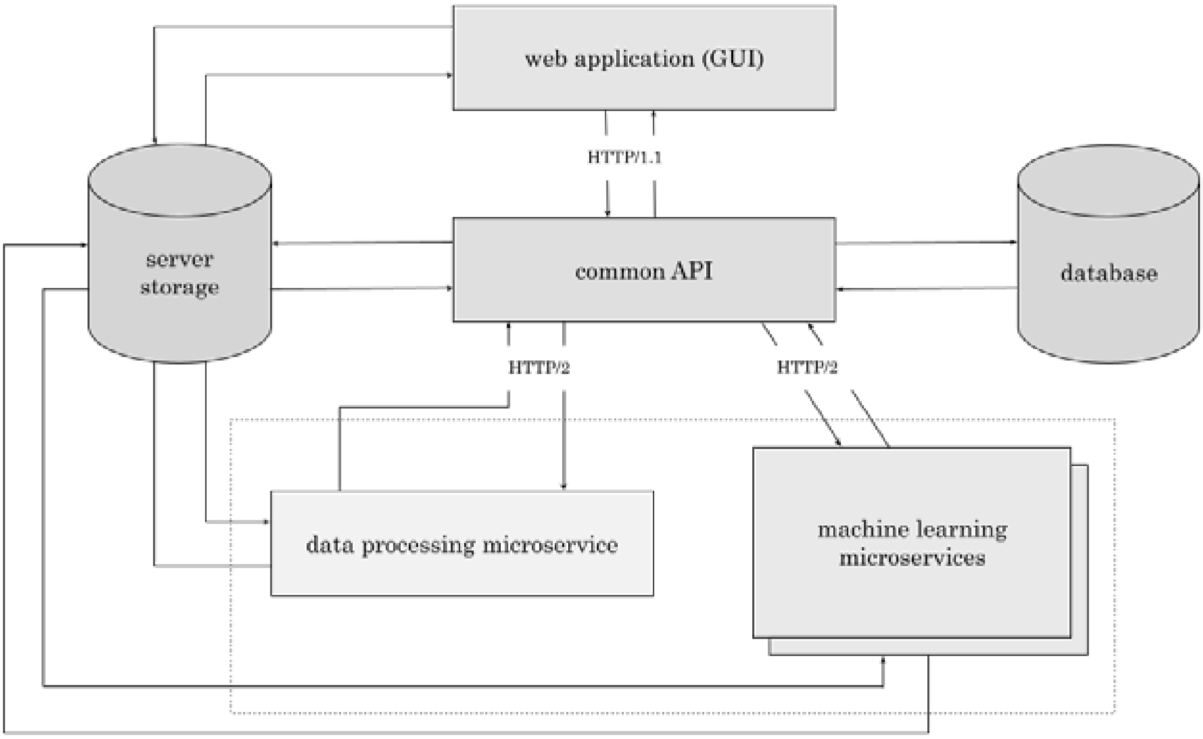


Fig. 1. Web service architecture scheme

2. The microservice architecture allows to extend the system functionality without rewriting the existing source code which makes future development easier and faster.
3. On certain datasets, usage of some machine learning algorithm implementations from libraries may lead to an infinite loop. Placing the algorithms in separate microservices prevents the entire application from crashing by using method execution timeouts. In the case when the specified time for method execution is exceeded, the task is terminated.

One of the most common tabular data formats for use in machine learning is CSV. In addition, Microsoft Excel is often used to create and edit tables, where files have the extension .xlsx. Therefore, the both XLSX and CSV file formats are supported.

The first line of the uploaded dataset should contain the names of the columns, the second and the next lines should contain the feature description of the objects (one object per line). Also, by default, it is considered that the first column of the dataset contains the object name. At the same time, the user can specify that the first column of the sample loaded also contains the feature description of the objects. The output files are in CSV format, with a comma as the separator. As with user-loaded datasets, the first line of the output file contains the column (feature) names. The column containing the result of the target variable prediction is marked as "target".

The graphical user interface is simple and consists of a set of HTML pages that requires minimum user interaction to create, configure, and use machine learning models for prediction. The training or prediction process involves a step-by-step navigation through the several web pages of the application.

First page of training process requires dataset upload and choice of delimiter character if the data is presented as a .csv file (screenshot: https://github.com/sirenescx/DAMDID-data/blob/master/gui/upload_train.png). Next page provides an ability to choose problem type, target variable and needed machine learning algorithms (screenshot: https://github.com/sirenescx/DAMDID-data/tree/master/gui/problem_settings). On this page user is also able to set custom model parameters for one or more selected models. After training process is set up, user is redirected to a web page on which he or she can track the progress of learning process (screenshot: https://github.com/sirenescx/DAMDID-data/blob/master/gui/log.png). Once training process is complete, the user will see a table containing the values of quality metrics for each of the selected algorithms which could be sorted by algorithms names or metrics values (screenshot: https://github.com/sirenescx/DAMDID-data/blob/master/gui/choice.png). At this stage, the user is prompted to select the best models for saving and further use.

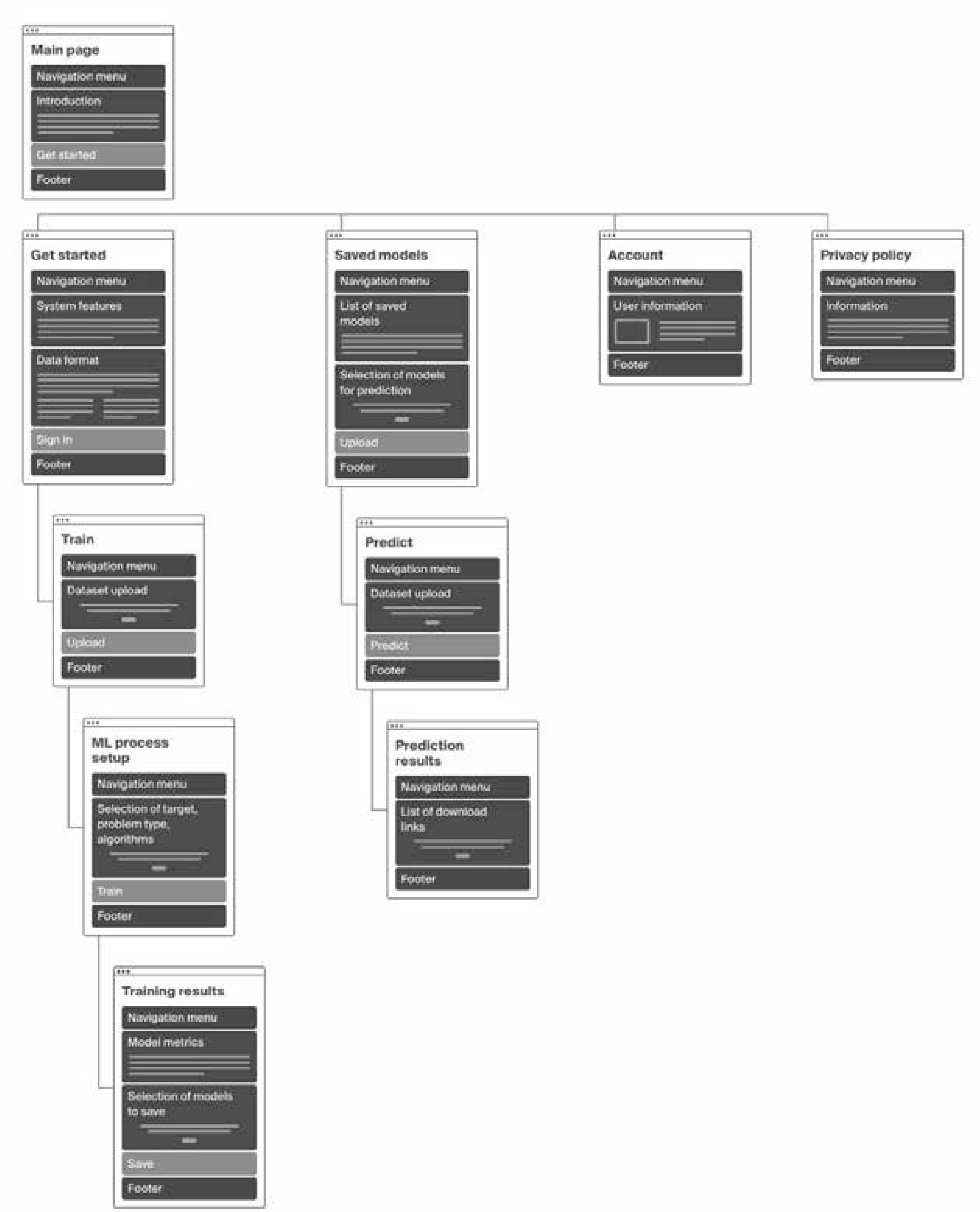


Fig. 2. Web pages navigation scheme

The interface of the prediction process is even simpler, the user just needs to select one of the saved pre-trained models (screenshot: https://github.com/sirenescx/DAMDID-data/blob/master/gui/results.png), upload a dataset (screenshot: https://github.com/sirenescx/DAMDID-data/blob/master/gui/predict.png), and wait for the prediction process to complete. Then, user can download prediction results to personal com-

Table 2

Model evaluation results

Algorithm	R^2	MSE	MAE
Ridge	0.987098625135458	0.0008175182548120166	0.01915270571705796
Lasso	0.987187519807067	0.0008118852879725164	0.021722537383973752
ElasticNet	0.986884345202036	0.0008310964787650111	0.02216231766619828

puter (screenshot: https://github.com/sirenescx/DAMDID-data/blob/master/gui/prediction_results.png). The page navigation scheme is presented below (Fig. 2).

To describe an example usage of the web application, a regression dataset containing 29 chemical objects – chalcospinels with $ABCX_4$ composition – with 108 continuous features was used for training (raw training dataset: https://github.com/sirenescx/DAMDID-data/blob/master/training_set.csv). This dataset contains data about chalcospinel compounds and their properties. The value of the target variable (crystal lattice parameter, *a,* ranges from 7.419 to 8.635 Å).

After data preprocessing one feature (E2-67) was dropped as non-informative because of a constant value of 1.8 (preprocessed training dataset: https://github.com/sirenescx/DAMDID-data/blob/master/training_set_processed.csv). Chosen algorithms set incuded three regularization methods (L2, L1 and L1/L2 regularization) and its implementations in Scikit-learn: *Ridge*, *Lasso* and *ElasticNet* models.

To evaluate the quality of the obtained models standard metrics for regression problems were used: the coefficient of determination (R^2), mean squared error (MSE) and mean absolute error (MAE). As can be noticed from the results (application output: https://github.com/sirenescx/DAMDID-data/blob/master/metrics.csv) given in Table 2, after applying hyperparameter optimization and cross-validation, all of the trained models had sufficient quality because R^2 score is close to 1. However, best algorithm is Lasso according to MSE value, and Ridge according to MAE value.

Conclusion

The popularity of machine learning is growing every year, so programs and web services for automated machine learning seem to be quite a promising area, as they make machine learning accessible not only to experts, but also to users with a basic understanding of the field. In addition, these systems can simplify and speed up development while analyzing data.

This paper presents one of the possible approaches to automate and simplify training, evaluation and obtaining predictions from machine learning algorithms for tabular datasets. The proposed approach is based on the development of the web application with graphical user interface.

Existing approaches for filling in missing data, encoding categorical variables, feature selection, and hyperparameter optimization were analyzed in this work. Chosen algorithms, methods, and implementations were also provided.

The current result of this work is a web application with GUI for the tabular data analysis which allows users to upload raw tabular dataset in one of the supported formats (.xlsx or .csv) and use Scikit-learn implementations of classification and regression machine learning algorithms to train the models or use them for making predictions on structurally identical data. Data preprocessing, hyperparameter optimization and model evaluation are done automatically by the web service. However, it is also possible for user to set custom model parameters if required.

As a next step, it is planned to add more complex solutions of value imputation, support of more regression and classification algorithms for tabular data from Scikit-learn library as well as to add such models from Keras library. Since the system can be easily extended due to the microservices architecture, it is also planned to provide users an ability to add handwritten models at a runtime.

In addition, at the time of writing this paper, the web application is being tested by real users, which allow us to collect a feedback and use it to improve the user interface and overall system performance.

References

1. *Sarker, I.H.* 2021. Machine Learning: Algorithms, real-world applications and research directions. SN Computer Science 2. Available at: https://doi.org/10.1007/s42979-021-00592-x (accessed November 15, 2021).
2. *Kumar Y., Kaur K., Singh G.* 2020. Machine learning aspects and its applications towards different research areas. 2020 International Conference on Computation, Automation and Knowledge Management (ICCAKM) Proceedings. Dubai. 150-156.
3. *Angra S., Ahuja S.* 2017. Machine learning and its applications: A review. 2017 International Conference on Big Data Analytics and Computational Intelligence (ICBDAC) Proceedings. Chirala. 57-60.
4. *Santu S. K. K., Hassan M. M., Smith M. J., Xu L., Zhai C., Veeramachaneni K.* 2022. AutoML to Date

and Beyond: Challenges and Opportunities. ACM Computing Surveys (CSUR) 54(8). Available at: https://doi.org/10.1145/3470918 (accessed November 17, 2022).

5. *Harrison M., eds.* 2019. Machine Learning Pocket Reference: Working with Structured Data in Python. 1st ed. Sebastopol, CA, USA: O'Reilly Media. 320 p.
6. *Micci-Barreca D.* 2001. A preprocessing scheme for high-cardinality categorical attributes in classification and prediction problems. SIGKDD Explorations 3(1). Available at: https://doi.org/10.1145/507533.507538 (accessed March 31, 2022).
7. *Jović A., Brkić K., Bogunović N.* 2015. A review of feature selection methods with applications. 2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO). Opatija. 1200-1205.
8. *Waring J., Lindvall C., Umeton R.* 2020. Automated machine learning: Review of the state-of-the-art and opportunities for healthcare. Artificial Intelligence in Medicine 104. Available at: https://doi.org/10.1016/j.artmed.2020.101822 (accessed November 21, 2021).
9. *Pedregosa F. et al.* 2011. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research 12. Available at: https://doi.org/10.48550/arXiv.1201.0490 (accessed May 12, 2022).

M.S. Manakhova. National Research University Higher School of Economics, 11 Pokrovsky boul., Moscow, 109028, Russia, e-mail: mmanakhova@hse.ru

V.A. Dudarev. PhD, A.A. Baikov Institute of Metallurgy and Materials Science of RAS, 49, Leninsky pr., Moscow, 119334, Russia, e-mail: vic@imet.ac.ru (correspondent author)

Проект System.AI: полностью управляемый стек машинного обучения и анализа данных для экосистемы .NET

Г.С. Брыкин

Московский Государственный Технический Университет им. Н.Э. Баумана,
г. Москва, Россия

Аннотация. В последние годы технологии машинного обучения становятся всё более распространёнными в таких известных задачах как стилизация изображений, окрашивание чёрно-белых изображений, супер-разрешение изображений, поиск поддельных данных, распознавание голоса и изображений. В связи с этим возникает необходимость в реализации набора инструментов для интеграции систем искусственного интеллекта в приложения для мобильных устройств, устройств умного дома и домашних ПК. Статья посвящена решению, позволяющему разработчикам интегрировать системы анализа данных и искусственного интеллекта непосредственно в приложение, что позволит получить легковесный, портативный, кроссплатформенный монолитный программный продукт, что зачастую невозможно с использованием существующих решений. Основными особенностями предлагаемого решения являются нацеленность на экосистему Microsoft .NET [1], а также использование только стандартных возможностей BCL и языка C#. Реализованный пакет инструментов является исключительно кроссплатформенным и аппаратнонезависимым. API во многом совпадает с аналогичными решениями для Python, что позволяет быстро перенести коды на Python в проект для .NET.

Ключевые слова: *машинное обучение, Анализ данных, .NET Framework, Mono [2], Xamarin [3], .NET Core [1], .NET Standard [1], Управляемый код.*

DOI: 10.14357/20790279230108

Введение

Платформа Microsoft .NET Framework и ее кроссплатформенные реализации, такие как Mono, .NET Core и Xamarin, предлагают разработчику широкий спектр инструментов для разработки мобильных, настольных и сетевых приложений; баз данных. Основной задачей проекта System.AI является внедрение в экосистему .NET стека технологий машинного обучения и анализа данных с сохранением при этом возможностей и синтаксиса аналогичных решений для Python или JavaScript, если это возможно. System.AI на данный момент включает в себя реализации библиотек *imageio* [4], *PyTorch* [5] и *convnet.js* [6], а также множество расширений стандартных классов .NET, которые обеспечивают гибкое взаимодействие между *System.AI* и .NET. Основными отличиями от предыдущих работ являются минимализм используемых инструментов из стандартного пакета .NET, язык C# 5, и полностью управляемый код, что делает *System.AI* по-настоящему кроссплатформенным программным обеспечением.

1. Предыдущие работы

Решения для **Python (PyTorch, TensorFlow, etc.).** Наиболее популярными фреймворками машинного обучения являются *PyTorch* от FaceBook и *TensorFlow* от Google [7]. Эти библиотеки для языка программирования Python обладают обширной функциональностью для реализации нейронных сетей для любой задачи, однако специфика языка Python крайне затрудняет его использование при разработке мобильных и настольных приложений. В частности, среда выполнения Python в настоящее время не имеет официальной поддержки для известной мобильной операционной системы Android. Кроме того, флагманские фреймворки машинного обучения не поддерживают 32-разрядные операционные системы, по-прежнему часто использующиеся на домашних ПК. Наконец, интеграция среды выполнения Python, базовых библиотек и библиотек машинного обучения в конечный продукт значительно увеличивает объём занимаемого дискового пространства, что критично для мобильных.

convnet.js и ConvNetCS. *convnet.js* – это библиотека для обучения свёрточных нейронных сетей непосредственно в браузере, написанная на JavaScript. Фактически, это первая широко известная библиотека машинного обучения, предназначенная для использования на широком спектре устройств. Преимущества включают кроссплатформенность и совместимость со многими браузерами и операционными системами. Однако эта библиотека предназначена для демонстрационных целей и обладает довольно узким функционалом, которого недостаточно для реализации современных нейронных сетей. Кроме того, JavaScript является скриптовым языком для веб-сайтов и редко используется для разработки приложений.

ConvNetCS [8] – это порт *convnet.js* на C#. Структура библиотеки позволяет использовать её в настольных приложениях, но ограниченная функциональность делает это бессмысленным.

Tensorflow.js (Deeplearn.js). Представляет собой библиотеку машинного обучения и линейной алгебры для запуска в браузере [9]. Она использует высокооптимизированные алгоритмы и поддерживает вычисления на графическом процессоре. Функционал достаточен для разработки и запуска практически всех современных нейронных сетей. Недостатками являются зависимость от версии браузера и невозможность использования при разработке приложений.

AlbiruniML. Является *Tensorflow.js* - подобной библиотекой машинного обучения и линейной алгебры с поддержкой автоматического дифференцирования [10]. Позиционируется автором как реализация *Tensorflow.js* для .NET и является кроссплатформенным программным обеспечением. Библиотека обладает достаточным функционалом для полноценной работы с нейронными сетями, но очень низким уровнем оптимизации алгоритмов. Последняя версия библиотеки была выпущена в 2018 году.

SciSharp Stack. Это готовый к использованию высокооптимизированный стек машинного обучения и анализа данных для экосистемы .NET [11]. Он написан в основном на C# с использованием кроссплатформенных возможностей BCL, но многие из наиболее важных библиотек в *SciSharp* используют неуправляемый код (неуправляемые .dll или файлы .so), например *TensorFlow.NET*, или вызывают коды на Python, например *Torch.NET* или *Numpy.NET*. Многие библиотеки в пакете фактически являются API, но не реализациями. Это не устраняет трудностей, связанных с использованием классических фреймворков машинного обучения.

Demos. Суть проекта заключается в реализации популярных нейронных алгоритмов для окрашивания, стилизации и супер-разрешения изображений на чистом C# [12]. Каждый алгоритм в предлагаемом проекте представлен в виде самостоятельного приложения для операционной системы Windows с удобным графическим интерфейсом. Цель проекта - предоставить возможность быстро запускать нейронные алгоритмы на настольных компьютерах. Проект считается предыдущим этапом работы над проектом *System.AI.*

MyCaffe. MyCaffe [13] – реализация библиотеки машинного обучения *Caffe* [14] на C#. Это кроссплатформенное и хорошо оптимизированное программное обеспечение, поддерживающее, однако, вычисления только на GPU от NVIDIA. Архитектура *MyCaffe* идеально подходит для использования в составе классических приложений на .NET, однако невозможность использования центрального процессора, а также ограничение на поддерживаемые графические процессоры существенно сужают круг поддерживаемых устройств.

2. Предлагаемое решение

Обзор структуры**.** *System.AI* - это набор взаимосвязанных библиотек. Все библиотеки из состава *System.AI* можно разделить на 3 глобальных типа: библиотеки ввода-вывода, предназначенные, соответственно, для обеспечения ввода и вывода данных различной природы; библиотеки расширений, реализующие отсутствующие в .NET типы данных (такие как, например, *Half*, *Quarter*, комплексные числа различной точности и т.д.), и обеспечивающие взаимодействие между типами *System.AI* и стандартными типами .NET, а также содержащие методы расширения для классов BCL. Например, библиотека *DotnetExtensions* реализует метод *tobytes()* для стандартных массивов. Это позволяет использовать подобный *NumPy* синтаксис для сохранения массивов в двоичной форме. Последний класс библиотек *System.AI* - это библиотеки машинного обучения, предназначенные непосредственно для создания, обучения и запуска нейронных сетей.

Для реализации всех компонентов *System.AI* был использован язык программирования C# 5. Это позволяет использовать консольный компиля-

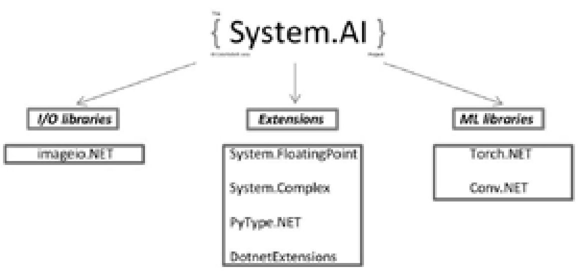


Рис. 1. Структура System.AI

тор C#, встроенный в Microsoft .NET Framework в ОС Windows, без необходимости установки стороннего программного обеспечения. Учитывая то, что требуемая минимальная версия (4.0) Microsoft .NET Framework является частью операционной системы, начиная с Windows 8, а также может быть установлена в операционных системах Windows XP, Windows Vista и Windows 7, мы можем предположить, что *System.AI* может быть скомпилирован из исходных кодов и использован с использованием стандартных инструментов самой распространённой настольной ОС – Windows. Кроме того, *System.AI* использует ограниченный набор исключительно кроссплатформенных возможностей BCL (Библиотеки базовых классов .NET), таких как многопоточность через *System.Threading.Tasks*, вывод на консоль через *System.Console* и работа с файлами и потоками через *System.IO*. Это позволяет запускать приложения с *System.AI* на любом устройстве, поддерживающем Microsoft .NET Framework или его реализацию, независимо от операционной системы или архитектуры процессора. Трассировки кода, выполняемого из *System.AI* и аналогичных решений представлены на рис. 2. Мы можем видеть, что код, основанный на *System.AI*, напрямую взаимодействует с операционной системой и аппаратным обеспечением, в то время как использование аналогичных решений предполагает один или несколько дополнительных уровней. Среда выполнения .NET содержит JIT-компилятор, который генерирует высокопроизводительный код для текущей архитектуры процессора. Таким образом, повышается производительность системы. IL-код *System.AI* и IL-код приложения почти так же эффективны, как машинный код. В настоящее время *System.AI* успешно собран и запущен на MIUI (Android) и облачной платформе Google Colab [15] (Ubuntu) с помощью Mono, а также на Windows (7, 10) с помощью .NET Framework.

В табл. 1 показано сравнение *System.AI* с аналогичными библиотеками.

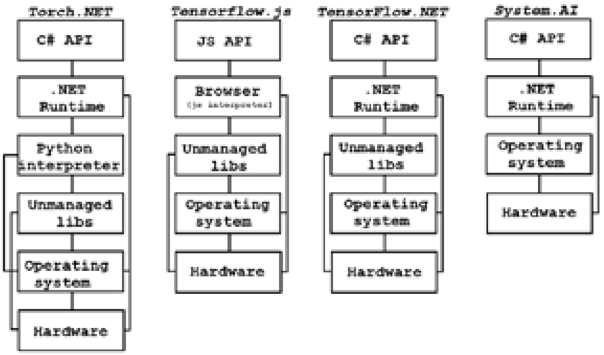


Рис. 2. Трассировка кода *System.AI* и предшествующих решений

imageio.NET. *imageio.NET* - это библиотека для ввода/вывода изображений. Она позволяет читать файлы в форматах bmp, gif, jpeg, png, psd и tga, возвращая изображение в виде трехмерного массива байт. Кроме того, библиотека позволяет сохранять изображение, представленное трехмерным массивом простого числового типа данных, в формах bmp, jpg, png, hdr и tga. В качестве ядра используются переработанные библиотеки *StbImageSharp* и *StbImageWriteSharp*. *imageio.NET* позволяет выполнять обработку изображений на любом устройстве, поддерживающем .NET или любую из его реализаций. Пример кода на PascalABC.NET для чтения изображения из файла в массив приведен ниже.

Табл. 1

Различия между *System.AI* и другими решениями. **Ведётся работа*.

	convnet.js	*ConvNetCS*	*tensorflow.js*	*AlbiruniML*	*SciSharp*	***System.AI***
Оптимизирован	Ч	Ч	√	Ч	√	√
Кросс-платформенность	√	√	~	√	~	√
Аппаратная независимость	√	√	√	√	~	√
Готов к использованию	Ч	Ч	√	Ч	√	~*
Для приложений	Ч	√	Ч	√	√	√

```
{$reference 'imageio.dll'}
uses System;
uses System.IO;
begin
   Console.WriteLine(imageio.__version__);  // current imageio.NET version
   var im1 := imageio.imread('Tuebingen_Neckarfront.jpg'); // read image as .NET byte array
   Console.ReadKey(true);
end.
```

System.FloatingPoint. Данная библиотека содержит типы данных *Quarter* (8-битный float), *Half* (16-битный float) и *BFloat16* (16-битный float от Google Brain). Реализованные типы показаны на рис. 3, 4 и 5.

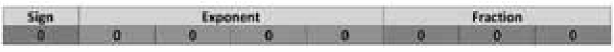


Рис. 3. Структура *Quarter*

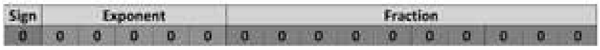


Рис. 4. Структура *Half*

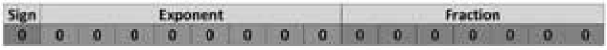


Рис. 5. Структура *BFloat16*

Типы данных *Half* и *BFloat16* активно используются в машинном обучении, и их поддержка должна быть доступна, по крайней мере, на уровне преобразований между встроенными типами .NET. Пример кода на VB.NET для работы с *System.FloatingPoint* предлагается ниже.

```
// Reference System.FloatingPoint.dll or System.AI.dll
Module Program
   Sub Main()
      Dim a As Quarter = Single.Parse(Console.ReadLine())
      Dim b As Half = Single.Parse(Console.ReadLine())
      Dim c As BFloat16 = Single.Parse(Console.ReadLine())
      Dim e As Object = a * b + c
      Console.WriteLine(String.Format("{0} * {1} + {2} = {3}, type = {4}", a, b, c, e, e.GetType))
      Console.ReadKey(True)
   End Sub
End Module
```

System.Complex. Библиотека реализует типы комплексных вещественных чисел, в которых действительная и мнимая части представлены значениями типа *Quarter*, *Half*, *BFloat16*, *Single* или *Double*. То есть каждому простому вещественному типу соответствует комплексный тип. Комплексные числа используются в машинном обучении и анализе сигналов.

PyType.NET. Эта библиотека реализует некоторые типы данных языка Python, которые позволяют сократить код. Одним из таких типов является шаблонный класс объединения (*Union*), который позволяет передавать значение одного из указанных типов данных. Он активно используется, когда необходимо передать одно число или пару чисел (например, в качестве параметров двумерного слоя свертки).

warnings.NET. Библиотека является аналогом библиотеки предупреждений стандартного пакета Python и предназначена для вывода предупреждений о некритических ситуациях, возникших во время выполнения программы.

DotnetExtensions. Библиотека содержит методы расширения для стандартных классов .NET. Задачи обработки данных часто требуют простых вспомогательных функций, применяемых, например, к числовым массивам. Такие функции могут быть использованы для отладки, ввода/вывода массивов данных и т.д. Подобный функционал также встроен в пакет *NumPy*, который является стандартным для работы с массивами в экосистеме Python. *DotnetExtensions* расширяет некоторые классы BCL функциями, аналогичными тем, которые содержатся в пакетах Python. В качестве примера приведём код на C# для сохранения числового массива в файл с использованием методов расширения *System.Array.tobytes()* и *System.IO.Stream.Write()*.

```
// string fname - file name
// arr - (multidimentional) array of basic .NET type (such as float or int)
using(var f = File.Create(fname))
{
 f.Write(arr.tobytes());
}
```

Torch.NET. *Torch.NET* - это библиотека машинного обучения и линейной алгебры с поддержкой автоматического дифференцирования. *Torch.NET* является прямым аналогом библиотеки *PyTorch* экосистемы Python. *Torch.NET*, как и другие библиотеки в составе *System.AI*, во многом копирует синтаксис своего аналога, однако, возможности *Torch.NET* всё же отличаются от возможностей *PyTorch*. Система типов *Torch.NET* значительно богаче, чем у *PyTorch*: дополнительно поддерживаются типы данных *Quarter*, *CQuarter*, *CBFloat16*, *UInt16*, *UInt32*, *UInt64*. Квантование на данный момент не поддерживается. Важное отличие *Torch.NET* от *PyTorch* заключается в том, что все операции со всеми типами данных реализованы для центрального процессора. На данный момент библиотека *Torch.NET* активно развивается.

Torchvision.NET. *Torchvision.NET* - это аналог модуля *torchvision*, входящего в состав *PyTorch*. *Torchvision.NET* содержит предобученные нейронные сети, методы чтения и дополнения данных, а также инструменты для работы с датасетами. Пример кода, который выполняет классификацию изображений с использованием *SqueezeNet*, показан ниже (Применены библиотеки *Torch.NET*, *Torchvision.NET*, *imageio.NET*).

```
using System;
using System.AI;
using System.Linq;
using System.Collections.Generic;
using models = System.AI.torchvision.models;
namespace Test
{
  public static class Program
   {
    public static torch.Tensor load(string uri)
     {
      return torch.tensor(imageio.imread(uri)).
      transpose(0,2).transpose(1,2).unsqueeze(0).@
      float() / 255f;
     }
     #region imagenet_classes
     public static string[] imagenet_classes = new
     string[]
     {
      <imagenet classes>
     };
     #endregion
     public static Dictionary<int, float> get_
     top5(float[] p)
     {
      <get top5 code>
     }
     public static void Main(string[] args)
     {
      var m = models.squeezenet1_1(true);
      var y = m.forward(load(args[0]));
      var pred = get_top5(y.squeeze(0).dotnet() as
      float[]);
      foreach(var p in pred)
      {
       Con-sole.WriteLine(imagenet_classes[p.
       Key]);
      }
      Console.ReadKey(true);
     }
   }
}
```

Conv.NET. *Conv.NET* - это реализация *convnet.js* для платформы .NET. Несмотря на то, что возможностей *Conv.NET* недостаточно для реализации сложных современных архитектур нейронных сетей, данная библиотека может быть использована в качестве учебной.

3. Методы

Ключевую роль в успехе и применимости фреймворка машинного обучения, как и любого другого программного продукта, играет эффективность кода. Реализация компонентов нейронной сети в управляемом коде представляет особый интерес, поскольку требует использования редко используемых специфических особенностей языка и платформы. Рассмотрим подход, используемый для реализации компонентов *System.AI*, на примере матричного умножения.

- Оптимизация алгоритма с учетом архитектуры вычислительной системы: минимизация непоследовательных обращений к оперативной памяти.
- Многопоточность.
- Прямой доступ к элементам матриц через указатели.
- Применение арифметики указателей.
- Векторизация.

Последовательное выполнение упомянутых выше оптимизаций позволяет получить прирост производительности более чем в 33 раза по сравнению с классическим (наивным) алгоритмом. Измерения проводились в 64-битном режиме для матриц 1000x1000.

Табл. 2

Влияние различных подходов к оптимизации кода на время его выполнения

Метод	Время, мс	Производительность
Наивный	3867	517,2 МФлопс
+ *Кэш*	2633	759,6 МФлопс
+ *Многопоточность*	903	2.2 ГФлопс
+ *Указатели*	**531**	**3.8 ГФлопс**
+ *Арифметика*	**261**	**7.6 ГФлопс**
+ *Векторизация*	**117**	**17 ГФлопс**

Существует несколько способов дальнейшего повышения производительности, одним из которых является использование .NET 5 или .NET Core, что позволит более эффективно, чем в .NET 4.8 с *System.Numerics.Vector4*, использовать векторные инструкции процессора.

При разработке *System.AI* используются компромиссные решения, позволяющие добиться максимальной эффективности по памяти и времени.

Зачастую разрабатываются новые алгоритмы и подходы. Например, в слоях свёртки используется модифицированный алгоритм im2col: вместо извлечения всех необходимых для свёртки участков входного изображения в матрицу с последующим умножением этой матрицы на матрицу ядер свёртки, как это происходит в известных фреймворках глубокого обучения, в *Torch.NET* каждый отдельный участок изображения преобразуется в вектор, который немедленно умножается на матрицу весов. Это решение позволяет радикально сократить объём занимаемой дополнительной памяти для вычисления свёртки, а также демонстрирует уровень производительности, сравнимый с уровнем классического im2col. Использование таких трюков возможно благодаря тому, что *System.AI* не полагается на готовые математические библиотеки, как это делают многие подобные продукты, вместо этого весь необходимый код пишется с нуля на C#.

4. Алгоритмы

Важным условием достижения высокой производительности программного обеспечения являются хорошие алгоритмы. При разработке *System.AI* выбираются компромиссные решения для достижения высокой эффективности по времени и памяти. Особое внимание уделяется алгоритмам свёрточных слоёв. В частности, был разработан новый алгоритм patch2vec, который представляет собой комбинацию наивного алгоритма свёртки и матричного im2col. Смысл большинства быстрых алгоритмов свёртки, таких как im2col или im2row, заключается в приведении свёртки к матричному умножению, что позволяет оптимизировать операции доступа к памяти за счёт использования кэша процессора. Однако такие методы требуют буфер для временных элементов, что крайне затратно. Предлагаемый метод (patch2vec) «на лету» разворачивает каждый участок входного изображения в вектор, а затем применяет к нему все фильтры свёртки. Этот алгоритм не уступает по эффективности классическим решениям вроде im2col, а на практике даже превосходит их в некоторых случаях. Буфер для этого алгоритма будет иметь размер , что много меньше, чем в случае аналогичных методов. Более того, patch2vec не накладывает ограничений на параметры свёртки, в отличие, например, от метода Шмуэля Винограда. Предлагаемый алгоритм трудно вписать в классические фреймворки машинного обучения из-за того, что они ориентированы на использование GEMM в качестве вычислительного ядра. Чистые реализации на C# или других языках программирования позволяют легко это сделать. В результате анализа существующих статей на тему быстрых свёрток [16] [17] было установлено, что предлагаемый алгоритм является новым и ранее не был представлен.

Patch2Vec.

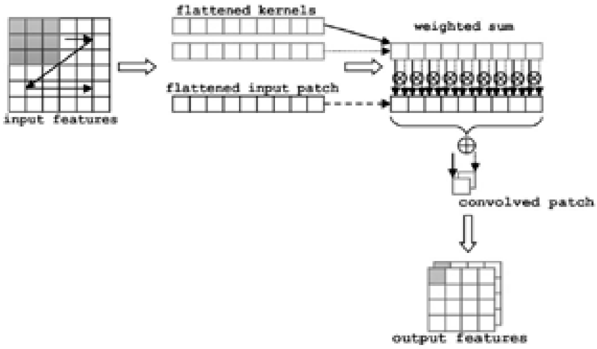


Рис. 6. Иллюстрация операции двумерной свёртки методом patch2vec.

На рис. 6 показана иллюстрация двумерной свёртки с помощью алгоритма patch2vec. Рассматривается случай, когда ядро свёртки квадратное и имеет размер 3x3, шаг симметричен и равен 1, расширение ядра отсутствует (=1), дополнение нулями отсутствует, количество групп равно единице. Тем не менее, алгоритм работает с любыми параметрами, разрешенными для операции свёртки. Кроме того, этот алгоритм также может быть применен для одномерных и трёхмерных свёрток.

Идея patch2vec состоит в том, чтобы избавиться от многократных (равных числу ядер свёртки) обращений к одним и тем же элементам входного тензора на каждом шаге свёртки. Вместо многократного доступа к одним и тем же данным, расположенным непоследовательно в памяти, предлагаемый алгоритм извлекает значения входного тензора, необходимые для применения ядра свёртки в данной точке, и записывает их в буфер (данные расположены последовательно в памяти), после чего ядра свёртки применяются не к области входного тензора, а к значениям, записанным в буфер. В этом случае все операции доступа к памяти при вычислении взвешенной суммы будут последовательными. Использование patch2vec уменьшает количество непоследовательных обращений к памяти в k раз (где k - количество ядер свёртки). Таким образом, алгоритм особенно эффективен при большом количестве ядер.

Vec2Patch. Vec2patch - это алгоритм, который является противоположностью patch2vec и предназначен для эффективного вычисления транспонированных свёрток и обратного распространения ошибки через обычную свёртку. Эффективная многопоточная реализация vec2patch сложна, поскольку требует синхронизации доступа к памяти при проецировании результирующего вектора на область результирующего тензора.

Заключение и дальнейшая работа

На данный момент основной задачей является реализация полного функционала *PyTorch* в библиотеке *Torch.NET*. Бета-версия *Torch.NET* будет опубликована в ближайшее время. Ещё одной важной задачей является добавление поддержки вычислений на графическом процессоре. Для этого планируется использовать технологию OpenCL - кроссплатформенный стандарт, поддерживаемый большинством современных видеокарт.

Исходные коды, двоичные файлы и документация, связанные с этим проектом, доступны на GitHub под лицензией Apache-2.0: https://github.com/ColorfulSoft/System.AI

Литература

1. .NET homepage. Available at: https://dotnet.microsoft.com (accessed November 22, 2022)
2. Mono homepage. Available at: https://www.mono-project.com (accessed November 22, 2022)
3. Xamarin homepage. Available at: https://dotnet.microsoft.com/apps/xamarin (accessed November 22, 2022)
4. Imageio homepage. Available at: https://github.com/imageio/imageio (accessed November 22, 2022)
5. *Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, Soumith Chintala.* 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. arXiv:1912.01703. Available at: https://arxiv.org/abs/1912.01703 (accessed November 22, 2022)
6. convent.js homepage. Available at: https://cs.stanford.edu/people/karpathy/convnetjs (accessed November 22, 2022)
7. *Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mane, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viegas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, Xiaoqiang Zheng.* 2016. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. arXiv:1603.04467. Available at: https://arxiv.org/abs/1603.04467 (accessed November 22, 2022)
8. ConvNetCS GitHub repository. Available at: https://github.com/mashmawy/ConvNetCS (accessed November 22, 2022)
9. *Daniel Smilkov, Nikhil Thorat, Yannick Assogba, Ann Yuan, Nick Kreeger, Ping Yu, Kangyi Zhang, Shanqing Cai, Eric Nielsen, David Soergel, Stan Bileschi, Michael Terry, Charles Nicholson, Sandeep N. Gupta, Sarah Sirajuddin, D. Sculley, Rajat Monga, Greg Corrado, Fernanda B. Viégas, Martin Wattenberg.* 2019. TensorFlow.js: Machine Learning for the Web and Beyond. arXiv:1901.05350. Available at: https://arxiv.org/abs/1901.05350 (accessed November 22, 2022)
10. AlbiruniML GitHub repository. Available at: https://github.com/mashmawy/AlbiruniML (accessed November 22, 2022)
11. SciSharp STACK homepage. Available at: https://scisharp.github.io/SciSharp/ (accessed November 22, 2022)
12. Demos GitHub repository. Available at: https://github.com/ColorfulSoft/StyleTransfer-Colorization-SuperResolution (accessed November 22, 2022)
13. *David W. Brown.* 2018. MyCaffe: A Complete C# Re-Write of Caffe with Reinforcement Learning. arXiv:1810.02272. Available at: https://arxiv.org/abs/1810.02272 (accessed November 22, 2022)
14. *Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama and T. Darrell.* 2014. Caffe: Convolution Architecture for Fast Feature Embedding. arXiv:1408.5093. Available at: https://arxiv.org/abs/1408.5093 (accessed November 22, 2022)
15. Google Colab homepage. Available at: https://colab.research.google.com (accessed November 22, 2022)
16. *Lavin, Scott Gray.* 2015. Fast Algorithms for Convolutional Neural Networks. arXiv:1509.09308. Available at: https://arxiv.org/abs/1509.09308 (accessed November 22, 2022)
17. *Anton V. Trusov, Elena E. Limonova, Dmitry P. Nikolaev and Vladimir V. Arlazarov.* 2021. p-im2col: Simple Yet Efficient Convolution Algorithm With Flexibly Controlled Memory Overhead. IEEE Access PP(99):1-1 (2021)

Брыкин Глеб Сергеевич. Московский Государственный Технический Университет им. Н.Э. Баумана, г. Москва, Россия. Количество печатных работ: 1. Область научных интересов: глубокое обучение, электроника. E-mail: glebbrykin@colorfulsoft.ru (ответственный за переписку).

The System.AI Project: Fully Managed Cross-Platform Machine Learning and Data Analysis Stack for .NET Ecosystem

G.S. Brykin

Bauman Moscow State Technical University, Moscow, Russia

Abstract. In recent years, machine learning technologies have become increasingly popular in widespread tasks such as image stylization, black-and-white image coloring, super-resolution of images, fake data searching, voice and image recognition. In this regard, there is a need to implement a set of tools for integrating artificial intelligence systems into applications for mobile devices, smart home devices, and home PCs. The paper describes a solution that allows developers to integrate data analysis and machine learning systems directly into a user application, which will allow to produce a lightweight, portable, and cross-platform monolithic application, which is often not possible with existing solutions. The main features of the proposed solution are the focus on the Microsoft .NET [1] ecosystem and the use of exclusively standard features of BCL and C# programming language. The implemented package of tools is completely cross-platform and hardware independent. The API is similar in many ways to its Python counterparts, which allows to quickly migrate Python codes into a .NET project.

Keywords: *machine Learning, Data Analysis, .NET Framework, Mono [2], Xamarin [3], .NET Core [1], .NET Standard [1], Managed Code.*

DOI: 10.14357/20790279230108

References

1. .NET homepage. Available at: https://dotnet.microsoft.com (accessed November 22, 2022)
2. Mono homepage. Available at: https://www.mono-project.com (accessed November 22, 2022)
3. Xamarin homepage. Available at: https://dotnet.microsoft.com/apps/xamarin (accessed November 22, 2022)
4. Imageio homepage. Available at: https://github.com/imageio/imageio (accessed November 22, 2022)
5. *Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, Soumith Chintala.* 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. arXiv:1912.01703. Available at: https://arxiv.org/abs/1912.01703 (accessed November 22, 2022)
6. convent.js homepage. Available at: https://cs.stanford.edu/people/karpathy/convnetjs (accessed November 22, 2022)
7. *Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mane, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viegas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, Xiaoqiang Zheng.* 2016. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. arXiv:1603.04467. Available at: https://arxiv.org/abs/1603.04467 (accessed November 22, 2022)
8. ConvNetCS GitHub repository. Available at: https://github.com/mashmawy/ConvNetCS (accessed November 22, 2022)
9. *Daniel Smilkov, Nikhil Thorat, Yannick Assogba, Ann Yuan, Nick Kreeger, Ping Yu, Kangyi Zhang, Shanqing Cai, Eric Nielsen, David Soergel, Stan Bileschi, Michael Terry, Charles Nicholson, Sandeep N. Gupta, Sarah Sirajuddin, D. Sculley, Rajat Monga, Greg Corrado, Fernanda B. Viégas, Martin Wattenberg.* 2019. TensorFlow.js: Machine Learning for the Web and Beyond. arXiv:1901.05350. Available at: https://arxiv.org/abs/1901.05350 (accessed November 22, 2022)
10. AlbiruniML GitHub repository. Available at: https://github.com/mashmawy/AlbiruniML (accessed November 22, 2022)

11. SciSharp STACK homepage. Available at: https://scisharp.github.io/SciSharp/ (accessed November 22, 2022)
12. Demos GitHub repository. Available at: https://github.com/ColorfulSoft/StyleTransfer-Colorization-SuperResolution (accessed November 22, 2022)
13. *David W. Brown.* 2018. MyCaffe: A Complete C# Re-Write of Caffe with Reinforcement Learning. arXiv:1810.02272. Available at: https://arxiv.org/abs/1810.02272 (accessed November 22, 2022)
14. *Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama and T. Darrell.* 2014. Caffe: Convolution Architecture for Fast Feature Embedding. arXiv:1408.5093. Available at: https://arxiv.org/abs/1408.5093 (accessed November 22, 2022)
15. Google Colab homepage. Available at: https://colab.research.google.com (accessed November 22, 2022)
16. *Lavin, Scott Gray.* 2015. Fast Algorithms for Convolutional Neural Networks. arXiv:1509.09308. Available at: https://arxiv.org/abs/1509.09308 (accessed November 22, 2022)
17. *Anton V. Trusov, Elena E. Limonova, Dmitry P. Nikolaev and Vladimir V. Arlazarov.* 2021. p-im2col: Simple Yet Efficient Convolution Algorithm With Flexibly Controlled Memory Overhead. IEEE Access PP(99):1-1 (2021)

Gleb S. Brykin. Bauman Moscow State Technical University, ul. Baumanskaya 2-ya, 5/1, Moscow, 105005, Russia, e-mail: glebbrykin@colorfulsoft.ru

On the Practical Generation of Counterfactual Examples*

D.E. Namiot, E.A. Ilyushin, I.V.Chizov

Lomonosov Moscow State University, Moscow, Russia

Abstract. One of the important elements in evaluating the stability of machine learning systems are the so-called adversarial examples. These are specially selected or artificially created input data for machine learning systems that interfere with their normal operation, are interpreted or processed incorrectly. Most often, such data are obtained through some formal modifications of the real source data. This article considers a different approach to creating such data, which takes into account the semantic significance (meaning) of the modified data - counterfactual examples. The purpose of the work is to present practical solutions for generating counterfeit examples. The consideration is based on the real use of counterfactual examples in assessing the robustness of machine learning systems.

Keywords: *machine learning, adversarial examples, counterfactual examples.*

DOI: 10.14357/20790279230109

Introduction

This article is a continuation of a series of publications devoted to robust machine learning models [1, 2]. It was prepared as part of the project of the Department of Information Security of the Faculty of Computer Science of Moscow State University named after M.V. Lomonosov on the creation and development of the master's program "Artificial Intelligence in Cybersecurity" [3].

The special impact on the elements of the pipeline of machine learning systems is called attacks on machine learning systems. Among such influences, the so-called adversarial attacks stand out. This is a special change in the input data, designed to change the operation of the machine learning system ("cheat" the system) or to achieve its desired behavior.

Adversarial attacks rely on the complexity of deep neural networks and their statistical nature to find ways to exploit them and change their behavior. There is no way to detect malicious activity with the classic tools used to protect software from cyber threats. Adversarial attacks manipulate the behavior of machine learning models. Most of the examples relate to working with images, but in reality, there are examples of attacks on text analysis systems, audio data classification (speech recognition), and time series analysis. In general, they can be considered as some universal risk for machine learning (deep learning) models [2]. There are various attempts to explain the nature of their existence. According to one hypothesis, adversarial attacks exist due to the non-linear nature of the systems, which leads to the existence of some data areas not covered by the generalization algorithm. According to others, it is, on the contrary, retraining of the system, when even small deviations from the training data set are processed incorrectly.

The term adversarial attack is often used loosely to refer to various types of malicious actions against machine learning models. And the term adversarial example describes the data used in an adversarial attack. The formal definition of a threat model for a classification problem (it is a typical use case for so-called critical applications) could be described via the following set of statements. Suppose we have an initial data set X and a finite set of class labels Y, it is necessary to find a mapping $f: X \rightarrow Y$. This mapping f is vulnerable to adversarial attacks when there is a mapping A such that for any $x \in X$ there exists $\widetilde{x} = A(x)$ for which $f(\widetilde{x}) \neq y$, given that $f(x) = y$.

The construction of adversarial examples, classically, is the search for minimal perturbations of the correct input data, which change the operation of the classifier. In this case, the search for such perturbations is performed purely formally (L-norms for images). In the article, we consider the issues of meaningful generation of test cases for machine learning systems in the form of so-called counterfactual examples. There

* This research has been supported by the Interdisciplinary Scientific and Educational School of Moscow University "Brain, Cognitive Systems, Artificial Intelligence".

are already a fairly large number of publications on this topic, but not all of the proposed approaches are really useful or even just applicable. The purpose of this work is to present exactly the approaches actually used in practice.

Common to all machine learning systems is, obviously, only the presence of a pipeline with standard stages: data selection (selection), model training, and testing (practical use) of the trained model. At each stage, developers have a sufficient selection of ready-made tools. On the one hand, this is undoubtedly a big plus. And technically, it's hard to imagine that there will be a single model, a single data cleansing tool, and so on. But, on the other hand, the practice of IT (and machine learning systems are, of course, IT systems) suggests that it is necessary to reuse models, architectures, etc. For economic success, IT projects cannot always be unique and must reuse some solutions (components of other solutions). For machine learning systems, an example of such reuse is AutoML solutions [36], where solutions are selected (fixed) for all individual elements of the machine learning pipeline. It is the requirements of reuse (in fact, these are economic requirements - the need for rapid implementation) that make it important to develop practical recommendations for reusable implementations of individual elements of the pipeline. And first of all, it is some standard (de facto standard) architectural solutions that are important. For many positions, we have many implementations that differ in quite specific aspects, and the question of choosing a specific implementation is not the most important one. Much more important is the overall solution architecture.

In this paper, we consider one of the elements of the machine learning pipeline - testing machine learning systems. Such testing is obviously different from traditional software testing. First of all, because the nature of the results of a machine learning system is probabilistic in nature, the basis for the results (conclusions) often cannot be not only verified, but simply obtained. The contribution of this article is architectural solutions for the so-called adversarial testing, that is, the search for data examples that cause machine learning systems to work incorrectly. In addition to justifying the architecture itself, we provide examples of the use of specific software products. It should be noted that, as examples, they illustrate the proposed approach, but, of course, are not exclusive.

More generally, the task we are solving is to build a trusted environment for the development of artificial intelligence systems [37]. Such environments, in the end, are sets of products (software systems) that affect different aspects of the machine learning pipeline, designed to increase confidence in the results of the developed machine learning models. These products include, for example, tools for adversarial attacks [38], formal verification of machine learning models [1], data cleansing, and so on. Naturally, testing and adversarial testing are necessarily part of such an environment. To avoid being tied to specific vendors, and to enable the development of critical applications, we consider only systems with open source, which can be modified if necessary. The article discusses the choice of products for testing machine learning systems.

The remainder of the article is structured as follows. Section II describes the actual approach to using counterfactuals. Section III focuses on the use of counterfactuals in machine learning. Section IV compares counterfactual and adversarial examples. Section V contains direct guidance on practical application.

1. On counterfactual examples

The use of counterfactuals has become a hot and popular topic in the machine learning community for many reasons such as explainability, interpretability, checking algorithmic fairness, etc. The general idea is clear enough. If we have a known output of the model, that is, for a given input, we know the output (the result of work) of the model, then we will be interested in the change in this output (result) when the input changes. Under what input data (what changes in input data) will the output change?

So, there are several models (and definitions) for the counterfactuals. For example, in the classification task, an example counterfactual explanation provides the following information: "for an example that belongs to class *A*, what changes do we need to make to the input so that the output will be classified as *B*". As per another definition, if we consider *A* and *B* to represent events or facts and *A* precedes *B* in time in the statement "*A* and *B* is true", then the counterfactual statement is 'If *A* had not occurred, *B* would not have occurred'. So, a counterfactual analysis can help to find whether *A* is a cause of *B* (it is by supposing the non-occurrence of *A* and seeking for the effect of this assumption on *B*.

In NLP a counterfactual example is defined as synthetically generated text which is treated differently by a condition model. For example, given the text "This program is written in Python", the counterfactual text becomes "This program is written in Java". If we know how the original sentence was classified, then how the counterfactuals will be classified?

In fact, we are talking about the conclusion that a small change in the data changes the output (result) of the work. This obviously coincides with the definition of the robustness of a machine learning model. Adversarial examples, in particular, are looking for precisely the minimum possible changes in the initial data that change the operation of the model. Accordingly, at least in theory, counterfactuals can be used to test the robustness of a machine learning model as well as for studying (checking) the fairness and transparency of algorithms. Counterfactual examples as tests should be more understandable (interpretable) as we tie them to some real conclusion. These examples will be created based on the interpretation (explanation) of the real conclusion (the result of the model).

In order to discuss counterfactuals, we have to turn to causality. The ability to understand causal relationships and to reason from them is one of the main human abilities [4]. Understanding physical causal relationships are fundamental to using any tool [5]. For example, people to people management (relations) is based on the understanding of psychological causal relationships. In many works, it is noted that it is convenient for human psychology to explain any conclusions by means of contrasting rather than direct explanations. We can explain, for example, a certain classification by giving reasons why only a certain class is chosen and why others are rejected. In other words, the explanation can be based on the choice and rejection of specific alternatives (results). This "discriminatory" explanation is counterfactual. For a machine learning system, counterfactual examples are input data that changes the result (classification, solution). The academic literature notes that this approach is more in line with emerging regulatory constraints, such as the General Data Protection Regulation (GDPR) [20]. The counterfactual approach helps to establish three important characteristics of the interpretability of models:

- determine how the interpretation of the model was made,
- provides opportunities for correcting unfavorable decisions
- provides hints for obtaining expected results in forecasting

As it is stated in [4], causal understanding thus maintains the kinds of cognition that have been proposed as part of the distinctively human cognitive toolbox. And in both the physical and psychological domains, causal knowledge is linked with sophisticated inferences about the counterfactual past. There are two distinctive features of causal knowledge, which are captured by causal models [4]:

1) causal knowledge supports a distinctive set of inferences involving interventions and counterfactuals. So, causal knowledge supports counterfactual claims;
2) causal knowledge involves not only specific relations between particular causes and effects but coherent networks of causal relations.

And one of the main statements from [4]: "counterfactual and intervention reasoning, and Bayesian learning all involve the same cognitive machinery: the ability to consider events that have not occurred". This is, in fact, a direct reference to the basic idea of machine learning. We train the network on a training dataset with the idea that a generalization will be built that will work correctly on the rest of the data, which, generally speaking, are unknown to us. We want to generalize the learning outcomes for the entire general population. That is, train the network to process data (events) that do not yet exist.

2. The usage of counterfactuals in machine learning

Counterfactual explanations are gaining attention as a way to explain the decisions of a machine learning model. There are several technical ways to generate and evaluate counterfactuals, such as feature-based explanations, prototype explanations, example-based explanations, or causal explanations [6].

We define a feature-highlighting explanation as an explanation that points to specific features in the model that matter to the individual decision. Of course, each type of feature-highlighting explanation may define this "matter" differently. There are two types of feature-highlighting explanations: counterfactual explanations and principal reason explanations. Principal reason explanation is defined in [8] as the reasons defined by law. Or more broadly, we could describe them as reasons based on some predefined set (some vocabulary).

The goal of counterfactual explanations is to explain how things could have been different, as well as provide a set of features changes for reaching a different output of the model in the future. Counterfactual explanations are generated by identifying such features that, if minimally changed, would alter the output of the model. For example, counterfactual explanations are trying to find the "nearest" hypothetical point that is classified differently from the point currently in question [7].

In other words, identifying the set of features results in the desired prediction while remaining at a minimum distance from the original set of features describing the individual [7]. It is illustrated in Fig. 1.

Suppose we are going to present counterfactual explanations for classification models, which are functions mapping input feature vectors $x \in X$ into label $c \in \{C_1, C_2, \dots C_n\}$. Actually, the most of research papers in this area have applied counterfactual explanations to classification tasks. Given a classification model [7]

$$f : X \rightarrow \{ C_1, C_2, \dots C_n \},$$

we can define the set of counterfactual explanations for a (factual) input $\hat{x} \in X$ *as* $CF_f(\hat{x}) = \{x \in X \,|\, f(x) \neq f(\hat{x})\}$. In other words, $CF_f(\hat{x})$ contains all the inputs x for which the model *f* returns a classification different from $f(\hat{x})$.

For prediction models, this is defined similarly, only the mapping will be carried out into the set {0, 1}

Based on the above-defined counterfactual space $CFf(\hat{x})$, we would like to produce counterfactual explanations for the output of a model f on a given input by trying to find a nearest counterfactual, which is defined as: $\hat{x}^* \in argmin\ d(x, \hat{x})\ for\ x \in CF_f(\hat{x})$

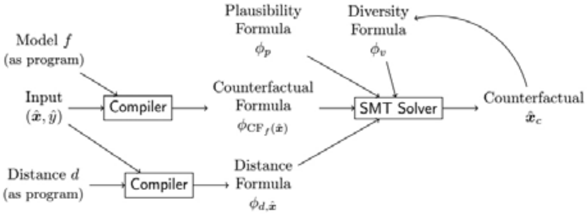

Model f (as program)
Input $(\hat{x}, \hat{y})$
Distance d (as program)
Compiler
Compiler
Plausibility Formula ϕ_p
Diversity Formula ϕ_v
Counterfactual Formula $\phi_{CF_f(\hat{x})}$
Distance Formula $\phi_{d,\hat{x}}$
SMT Solver
Counterfactual $\hat{x}_c$

Fig. 1. Architecture Overview for Model-Agnostic Counterfactual Explanations (MACE) [7]

Prototype-based counterfactual explanations are discussed in the paper [9]. Since we cannot interpret the black box, our first task is to create an interpreted view for it. This is exactly what prototyping-based methods do. The work of the black box is being prototyped. As a first step, we need to find a representative dataset. The maximum mean discrepancy (a distance on the space of probability measures) is used to calculate the representativeness. After that prototype-based explanations provide the nearest prototype as explanations for a given test instance [10]. In some papers, a similar approach is called an example-based explanation: example-based approaches seek to find data points in the vicinity of the explainee data point. They either offer explanations in the form of data points that have the same prediction as to the explainee data point or the data points whose prediction is different from the explainee datapoint [11]. In other words, example-based approaches are another kind of explainability technique used to explain a particular outcome.

In general, the explainability problem for machine learning systems can be presented as model explanation or outcome explanation problems. As per definition, a model explanation is about an interpretable and transparent explanation of the original model.

As the developed techniques for neural networks explanations, we could mention decision trees [12, 13] and rule sets [14, 15]. As per software tools, there are some model-agnostic packages. For example - Partition Aware Local Model (PALM) [16]. PALM allows you to study the structure of model conclusions using its approximation by surrogate models. This is to some extent a general approach, also called partial modeling in the literature. In such a model, we are trying to approximate the general black box with "understandable" models. This, in particular, should help in debugging models. PALM approximates a neural network using a two-part surrogate model, which includes a meta-model that partitions the training data, and a set of sub-models that approximate the patterns (solutions) within each partition. The paper [17] describes an approximation algorithm, GoldenEye, to select sets of attributes that influence the work of the classifier. In fact, this is combinatorics, when the possible combinations are sorted out.

Outcome explanation needs to provide an explanation for just a specific prediction from the model. This type of explanation does not affect the internal logic of the models, but only deals with inferences. There are model-specific approaches like Grad-CAM [18] and model agnostic approaches like LIME [19] have been proposed. All of them are provided either feature attribution or model simplification methods.

3. Counterfactual and adversarial examples

In general, the performance (the predictive of classification performance) for any machine learning model is based on the assumption of a statistical similarity of the distributions of training and production (testing) data.

Note that in the general case, the general set of data is unknown to us. Accordingly, it is unknown not only how the training and test data correlate with each other, but also how the test and training data separately correlate with the general population.

In the classic example [22], we have (an actually unknown) some sine curve, the test and training data for which just happened to be on different crests (Fig. 2). Both the training and test data are perfectly (very accurately) approximated by some straight line, but these are completely different straight lines (different angles of inclination).

And in the general case, we cannot assume similar distributions. The only way to somehow guarantee this, obviously, involves exhaustive knowledge of the population. In some tasks, this is really possible, but this is usually the point that is not discussed in the works on machine learning, although, of course, it deserves separate consideration.

Since, in fact, when creating a model, we only know the training data and, accordingly, their statistical characteristics, studying the operation of the model on data whose distribution differs from the distribution of the training data looks like a natural process. As noted in [23], evaluation of out-of-distribution data is a common practice in NLP and image proceedings.

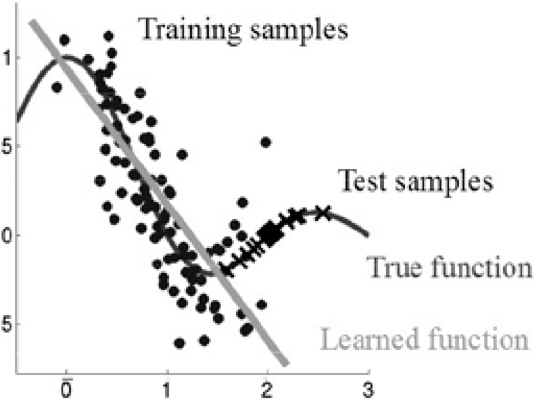


Fig. 2. Training and test samples [22]

And the reason for this is the different distributions of test and training data (Fig.3).

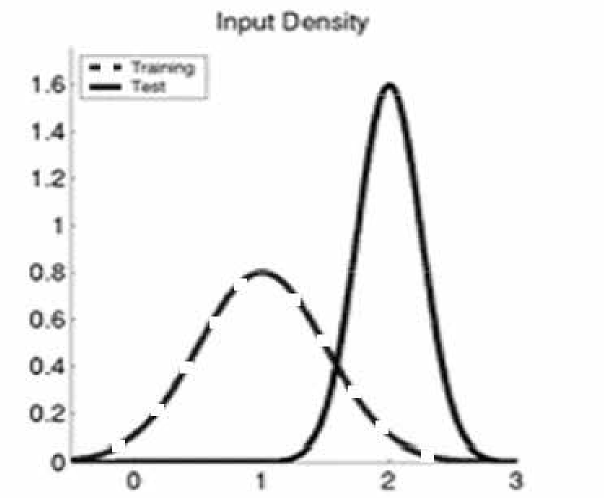


Fig. 3. Distribution shift [22]

One fact may be noted here. Model errors (poor generalization) are often associated precisely with what is called network overtraining [24]. This happens when the models rely on some training dataset-specific biases and artifacts rather than intrinsic properties of the data. When these biases do not exist in the production data, the performance of the models can drop dramatically. The keywords here are "intrinsic properties". And the way to identify them is just counterfactual examples. We change the data, the solution of the system is reversed, which allows us to assume that the changed (deleted) data are the main characteristics based on which the machine learning model makes a decision.

In ideology, this is similar to competitive examples, but their search is carried out not through a sequential selection of modifications that change the solution, but by one-time changes that change the "meaning" of the image. The emphasis is on pairs: the image and its counterfactual example(s).

In this connection, we could cite works that discussed generalization from a causal perspective [25, 26]. To provide generalization, the model must reflect the real causal mechanisms behind the data.

One practical example that can be mentioned (and actually used) in this situation.

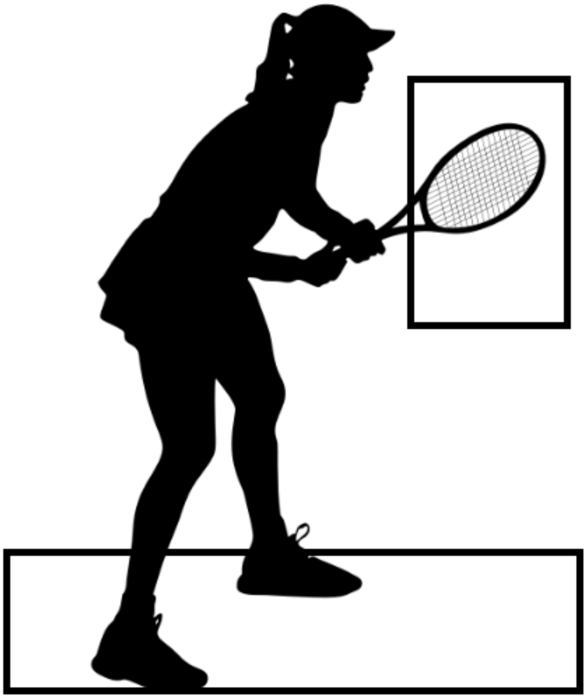


Fig.4. Attention areas [27]

Using the well-known approach for evaluating areas of the image that attract attention [27], one can try to build counterfactual examples by removing just these areas. In this work, this was illustrated by a tennis player, where attention was drawn to the racket and the surface (Fig. 4). Strictly speaking, it was on these parts of the image that a person evaluated the image. In our case, this was used on a traffic sign recognition system (Fig. 5), and images (Fig. 6)

The counterfactual examples constructed in this way were used for adversarial training and increasing robustness.

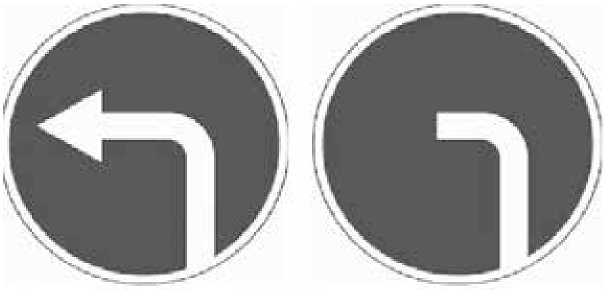


Fig.5. Road sign: original (left) and counterfactual (right)

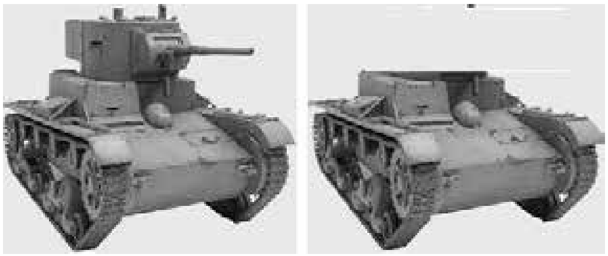


Fig. 6. Images: original (left) and counterfactual (right)

Summarizing these results, we can say that counterfactual examples are adversarial examples built with image semantics in mind. If we talk about the classical form of constructing adversarial examples, then we operate with pixels in the framework of L-norms, and not with image fragments. Of course, we should be talking about data in general here, but in practice, we are talking about images (as in the vast majority of works related to sustainable machine learning) and, as discussed in the next section, texts.

Note that the analysis of the content of images corresponds to the ideas outlined in the pioneering work [28], where the authors propose a new model for deep learning based precisely on enlarged fragments. The current state of research suggests that, within the framework of per-pixel processing, the robustness of machine learning systems cannot be ensured.

4. On practical examples

Which in the end can be used to prepare counterfactual examples in practical applications?

A. Counterfactual examples for texts

Counterfactual examples for text represent the simplest and most clearly interpretable model. Several approaches to generating such examples are described in the literature, in a simple form they can be presented:

If we have a sentence in the "Who did what" format, then building a counterfactual example is, in fact, a well-known exercise from foreign language textbooks - "make a new sentence that denies the original statement ("No one did it" format)

For example, such negation formats are given in [29] for the source text "I am very disappointed with the service":

There is practically no difference between the approaches, since technically they do pretty much the same thing. As an example of software (a toolkit that can be used in your own projects), you can cite, for example, Checklist [30].

It is a system built on the basis of templates. Here is a typical example (source data https://github.com/marcotcr/checklist)

```
import checklist
from checklist.editor import Editor
import numpy as np
editor = Editor()
ret = editor.template(‘{first_name} is {a:profession} from {country}.’,
profession=[‘lawyer’, ‘doctor’, ‘accountant’])
np.random.choice(ret.data, 3)
```

and here is the result:

```
[‘Mary is a doctor from Afghanistan.’,
‘Jordan is an accountant from Indonesia.’,
‘Kayla is a lawyer from Sierra Leone.’]
```

To generate the Checklist uses the set of predefined templates, lexicons, generic perturbations, and context-sensitive sentences. The main limitation of these pattern-based or rule-based approaches is that they cannot generate meaningful diversity [29].

Token-based Substitution in Table 1 uses either single word replacements or some templates to generate multiple test cases. Adversarial examples, as usual, do not evaluate the text at all (do not appreciate the meaning).

Other approaches are used, for example, to generate text GPT-2, BERT, or bag-of-words models [31]. In general, we can characterize this direction as quite developed from a practical point of view, with ready-to-use tools.

B. Counterfactual examples for images

Originally, Search for EviDence Counterfactual (SEDC) is the model-agnostic search algorithm

Table 1

TEXT COUNTEFACTUALS

Input Sentence Token-based Substitution	**Adversarial attack**	**Controlled Counterfactual Generation**
I am very pleased with the service. I am very happy with the service.	I am very impressed with the service. I am very witty with the service.	I am very happy with this service. I am very pleased with the service.

(SEDC) to find counterfactual explanations for document classifications. According to this algorithm, the explanation can be considered as an irreducible set of characteristics (for example, for text documents – words), which, in their absence, would change the classification of the document.

The modification presented in [32] as a set of characteristics uses segments into which the original image is divided. As noted above, for images, again, not pixel processing is used, but manipulations with significant elements of the image. Accordingly, the explanation for an image is an irreducible number of segments, the removal of which will change the classification of the image.

As per [32], consider an image *I* assigned to class *c* by a classifier *C* the objective is to find a counterfactual explanation *E* as an irreducible set of segments that leads to another classification after removal. Formally:

$E \subseteq I$ (segments in image)

$C(I \setminus E) \neq c$ (class change)

$\forall E' \subset E : C(I \setminus E') =$ c (irreducible)

The counterfactual explanation, in this case, is the classification of the image that is obtained after removing the segments.

The original work thus defined the "minimal" image, which was still classified as an "airplane" (fuselage without wings). Approaches similar to this have been illustrated above.

In [33], with the telling title "Explanations based on the missing", this is described as pertinent positive (PP) and pertinent negative (PN). The PP is a factor that is minimally required for the justification of the final decision and the PN is a factor whose absence is minimally required for justifying the decision.

The basic implementation of SEDC is represented by the resource [34]. In the basic case, the image is segmented, and then the segments are removed one by one until the classification of the remaining "image" changes. In a modified version of SEDC-T [32], segments are also removed one at a time, until the classification reaches the specified value. In other words, SEDC-T gives a more detailed explanation of why the image is not predicted as the correct class (removing which segments leads to a given misclassification).

As for the actual segmentation of images, many approaches can be used here, in addition to the attention map presented above. In fact, you can use a simple grid to divide the image into segments or use more complex approaches that are widely presented in open implementations [35]. The obvious advantages of semantic segmentation are the possible explainability of the results and the ability to use the results for physical attacks (for example, to hide part of the image with a patch). We also note that semantic image segmentation is also present in popular packages for machine learning, such as Keras and Tensorflow [39].

C. Counterfactual examples for sounds

Counterfactual examples for sound classification present a more exotic challenge. In practice, we can only give an example from [21]. Here, the validity of automatic speech recognition (ASR) models was studied. The same text was recorded in different voices (for different ethnic groups, different sexes, and ages). At the same time, the work of the recognition system should not be disturbed. The paper practically shows that the widely used automatic speech recognition systems are unfair, since some groups of users had a higher error rate than others. One way to define fairness in ASR is to require that changing any person's demographic group (for example, changing their gender, age, education, or race) does not change the probability distribution between the possible speech-to-text transformations. In the counterfactual justice paradigm, all variables that do not depend on group membership (for example, the text read by the speaker) remain unchanged, while variables that depend on group membership (for example, the speaker's voice) change counterfactually. Therefore, one can attempt to achieve a fair ASR performance by teaching the ASR to minimize the change in the probabilities of recognition outcomes despite the counterfactual change in human demographics.

Conclusion

In this article, we focused on generating adversarial tests for machine learning systems. As we noted in previous works, testing machine learning systems is robustness testing.

Traditional methods, considered as an optimization problem of finding the smallest modifications that change the results of the classification, give, in the end, very limited results in terms of increasing stability. And, most importantly, the proposed modifications are completely artificial by their nature, in no way connected with possible attacks. In this regard, in this paper, we justified the use of counterfactual examples for generating tests, since they are related to the semantical analysis of data.

The purpose of this paper was to present practical reusable solutions for generating counterfactual examples for various types of input data. The result of our research, based on the practical use of various products, is the presentation of a pipeline for constructing counterfactual examples in image recognition and text classification problems.

Creating counterfactual examples for text classification is currently a purely technical task. The question is only in choosing the most convenient software implementations. The algorithms are quite transparent and can be built into your own applications. We propose to use template-based systems, like the above-mentioned Checklist.

In terms of building counterfactual examples for images, the best choice, in our opinion, is the semantic segmentation of images. We propose to use the open source implementation of SEDC-T. Alternative methods to some extent reproduce approaches to constructing adversarial examples and are based on a formal assessment of the change in the quality of the system when modifying images.

Counterfactual examples for sound classification (important, for example, for biometric identification systems) are the least developed area. To date, we cannot offer practical solutions in this direction. One reason for this is the nature of existing classification systems, which rely on various artificially created characteristics. For example, wavelet transforms, etc. With their use, the reverse transition to modifications of the original sound characteristics becomes unclear.

We are grateful to the staff of the Department of Information Security of the Faculty of Computational Mathematics and Cybernetics, Lomonosov Moscow State University for valuable discussions of this work.

References

1. *Namiot, Dmitry, Eugene Ilyushin, and Ivan Chizhov*. “On a formal verification of machine learning systems.” International Journal of Open Information Technologies 10.5 (2022): 30-34.
2. *Li, Huayu, and Dmitry Namiot.* “A Survey of Adversarial Attacks and Defenses for image data on Deep Learning.” International Journal of Open Information Technologies 10.5 (2022): 9-16.
3. Artificial Intelligence in Cybersecurity. http://master.cmc.msu.ru/?q=ru/node/3496 (in Russian) Retrieved: May, 2022
4. *Buchsbaum, Daphna, et al.* “The power of possibility: Causal learning, counterfactual reasoning, and pretend play.” Philosophical Transactions of the Royal Society B: Biological Sciences 367.1599 (2012): 2202-2212.
5. *Sterelny, Kim.* “Language, gesture, skill: the co-evolutionary foundations of language.” Philosophical Transactions of the Royal Society B: Biological Sciences 367.1599 (2012): 2141-2151.
6. *Kasirzadeh, Atoosa and Andrew Smart.* “The use and misuse of counterfactuals in ethical machine learning.” Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. 2021.
7. *Amir-Hossein Karimi, Gilles Barthe, Borja Belle, and Isabel Valera.* 2019. Model-Agnostic Counterfactual Explanations for Consequential Decisions. arXiv preprint arXiv:1905.11190 (2019)
8. *Barocas, Solon, Andrew D. Selbst, and Manish Raghavan.* “The hidden assumptions behind counterfactual explanations and principal reasons.” Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. 2020.
9. *Duong, Tri Dung, Qian Li, and Guandong Xu.* “Prototype-based Counterfactual Explanation for Causal Classification.” arXiv preprint arXiv:2105.00703 (2021).
10. *Yadav, Chhavi, and Kamalika Chaudhuri.* “Behavior of k-NN as an Instance-Based Explanation Method.” arXiv preprint arXiv:2109.06999 (2021).
11. *Verma, Sahil, John Dickerson, and Keegan Hines.* “Counterfactual explanations for machine learning: A review.” arXiv preprint arXiv:2010.10596 (2020).
12. *Thiagarajan, Jayaraman J., et al.* “Treeview: Peeking into deep neural networks via feature-space partitioning.” arXiv preprint arXiv:1611.07429 (2016).
13. *Boz, Olcay.* “Extracting decision trees from trained neural networks.” Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining. 2002.
14. *Santos, Raul T., Júlio C. Nievola, and Alex A. Freitas*. “Extracting comprehensible rules from neural networks via genetic algorithms.” 2000 IEEE Symposium on Combinations of Evolutionary Computation and Neural Networks. Proceedings of the First IEEE Symposium on Combinations of Evolutionary Computation and Neural Networks (Cat. No. 00. IEEE, 2000.
15. *Andrews, Robert, Joachim Diederich, and Alan B. Tickle.* “Survey and critique of techniques for extracting rules from trained artificial neural networks.” Knowledge-based systems 8.6 (1995): 373-389.
16. *Krishnan, Sanjay, and Eugene Wu.* “Palm: Machine learning explanations for iterative debugging.” Proceedings of the 2nd Workshop on Human-In-the-Loop Data Analytics. 2017.
17. *Henelius, Andreas, et al.* “A peek into the black box: exploring classifiers by randomization.” Data mining and knowledge discovery 5 (2014): 1503-1529.

18. *Selvaraju, Ramprasaath R., et al.* “Grad-cam: Visual explanations from deep networks via gradient-based localization.” Proceedings of the IEEE international conference on computer vision. 2017.
19. *Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin.* “Model-agnostic interpretability of machine learning.” arXiv preprint arXiv:1606.05386 (2016).
20. *Gohel, Prashant, Priyanka Singh, and Manoranjan Mohanty.* “Explainable AI: current status and future directions.” arXiv preprint arXiv:2107.07045 (2021).
21. *Sari, Leda, Mark Hasegawa-Johnson, and Chang D. Yoo.* “Counterfactually Fair Automatic Speech Recognition.” IEEE/ACM Transactions on Audio, Speech, and Language Processing (2021).
22. Francisco Herrera Dataset Shift in Classification: Approaches and Problems http://iwann.ugr.es/2011/pdf/InvitedTalk-FHerrera-IWANN11.pdf Retrieved: Sep, 2021
23. *Teney, Damien, Ehsan Abbasnedjad, and Anton van den Hengel.* “Learning what makes a difference from counterfactual examples and gradient supervision.” Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part X 16. Springer International Publishing, 2020.
24. *Roelofs, Rebecca, et al.* “A meta-analysis of overfitting in machine learning.” Proceedings of the 33rd International Conference on Neural Information Processing Systems. 2019.
25. *Heinze-Deml, Christina, and Nicolai Meinshausen.* “Conditional variance penalties and domain shift robustness.” arXiv preprint arXiv:1710.11469 (2017).
26. *Meinshausen, Nicolai.* “Causality from a distributional robustness point of view.” 2018 IEEE Data Science Workshop (DSW). IEEE, 2018.
27. *Das, Abhishek, et al.* “Human attention in visual question answering: Do humans and deep networks look at the same regions?.” Computer Vision and Image Understanding 163 (2017): 90-100.
28. *Bengio, Yoshua, Yann Lecun, and Geoffrey Hinton.* “Deep learning for AI.” Communications of the ACM 64.7 (2021): 58-65.
29. *Madaan, Nishtha, et al.* “Generate your counterfactuals: Towards controlled counterfactual generation for text.” arXiv preprint arXiv:2012.04698 (2020).
30. *Ribeiro, M.T., Wu, T., Guestrin, C. and Singh, S.* 2020. Beyond Accuracy: Behavioral Testing of NLP Models with CheckList. arXiv preprint arXiv:2005.04118 .
31. *Dathathri, Sumanth, et al.* “Plug and play language models: A simple approach to controlled text generation.” arXiv preprint arXiv:1912.02164 (2019).
32. *Vermeire, Tom, and David Martens.* “Explainable image classification with evidence counterfactual.” arXiv preprint arXiv:2004.07511 (2020).
33. *Dhurandhar, Amit, et al.* “Explanations based on the missing: Towards contrastive explanations with pertinent negatives.” arXiv preprint arXiv:1802.07623 (2018).
34. SEDC implementation https://github.com/yramon/edc Retrieved: May, 2022
35. *Van der Walt, Stefan, et al.* “scikit-image: image processing in Python.” PeerJ 2 (2014): e453.
36. *He, Xin, Kaiyong Zhao, and Xiaowen Chu.* “AutoML: A survey of the state-of-the-art.” Knowledge-Based Systems 212 (2021): 106622.
37. *Namiot, Dmitry, Eugene Ilyushin, and Oleg Pilipenko.* “On Trusted AI Platforms.” International Journal of Open Information Technologies 10.7 (2022): 119-127. (in Russian)
38. *Ilyushin, Eugene, Dmitry Namiot, and Ivan Chizhov.* “Attacks on machine learning systems-common problems and methods.” International Journal of Open Information Technologies 10.3 (2022): 17-22. (in Russian)
39. *Dadhich, Abhinav*. Practical Computer Vision: Extract Insightful Information from Images Using TensorFlow, Keras, and OpenCV. Packt Publishing Ltd, 2018.

D.E. Namiot. Dr. of Sci., Lomonosov Moscow State University, MSU, Faculty of Computational Mathematics and Cybernetics, Russia, 119991, Moscow, GSP-1, 1-52, Leninskiye Gory, dnamiot@gmail.com (correspondent author)

E.A. Ilyushin. MSU, Faculty of Computational Mathematics and Cybernetics, Russia, 119991, Moscow, GSP-1, 1-52, Leninskiye Gory, john.ilyushin@gmail.com

I.V. Chizov. PhD, docent, Lomonosov Moscow State University; Federal Research Center “Computer Science and Control” of the Russian Academy of Sciences, MSU, Faculty of Computational Mathematics and Cybernetics, Russia, 119991, Moscow, GSP-1, 1-52, Leninskiye Gory, ichizhov@cs.msu.ru

A Survey of Model Inversion Attacks and Countermeasures*

Junzhe Song, D.E. Namiot

Lomonosov Moscow State University, Moscow, Russia

Abstract. This article provides a detailed overview of the so-called Model Inversion(MI) attacks. These attacks aim at Machine-Learning-as-a-Service (MLaaS) platforms, and the goal is to use some well-prepared adversarial samples to attack target models and gain sensitive information from ML models, such as items from the dataset on which ML model was trained or ML model's parameters. This kind of attack now becomes an enormous threat to ML models, therefore, it is necessary to research this attack, understand how it will affect ML models, and based on this knowledge, we can propose some strategies that may improve the robustness of ML models.

Keywords: *adversarial machine learning, model inversion attack, deep-learning, cyber security.*

DOI: 10.14357/20790279230110

Introduction

With the development of Machine Learning(ML) and the increase of data size, a large number of ML models have been created and utilized in many parts of human life, we define these models as *Machine-Learning as a service(MLaaS)*. Among these models, there is a kind of models that utilize personal privacy-sensitive information as training data and provide services, such as medical applications(analyze one's medical index and provide medical suggestions) and facial recognition applications(analyze given figure and return name & confidence-value).

Since there are many privacy-sensitive values stored in these models, there are also many malicious attackers who try to gain this privacy-sensitive information from ML datasets (usually we call these attackers *Adversaries*). Therefore, privacy risk becomes an important part of ML models, researchers and MLaaS providers start to try to utilize ML to preserve this privacy-sensitive information from adversary attacks, which creates a new section of ML called: Privacy-Preserving Machine Learning(PPML).

Model Inversion(MI) attack is one of the methods in PPML, which inverts the process of training data into an ML model. The threat, in this case, is potentially exposing data from the training set, which can contain private information, to the adversary. Although there are many types of MI attack methods that have been created, the main issue is: MI attack is unaware of the victims.

Unfortunately, the existing countermeasures can only defend against the corresponding attack, which means, to improve the ML model's robustness, we have to apply several countermeasures simultaneously, this is also a passive approach because we can't preserve our model from unrevealed attacks. Since to find out if our model is under attack as soon as possible, maybe we can create a MI attack detector.

For this detector, we hope we can develop one which can detect not only existing attack methods but also unrevealed attack methods or behavior that act like MI attacks. This will be hard work because attack methods can upgrade very quickly and their features are various. In section *Summary and Future Works*, we discuss some attributes that the detector should possess.

This research has been supported by the Interdisciplinary Scientific and Educational School of Moscow University "Brain, Cognitive Systems, Artificial Intelligence".

1. The Model Inversion Attacks and countermeasures

1.1. The Model Inversion(MI) attacks

The Model Inversion(MI) attacks are mainly aimed at the currently popular MLaaS service model, it refers to an attacker extracting information related to training data from the model prediction results.

A simple deployment of a model inversion attack is presented in Fig.1.

To perform a model inversion attack, the adversarial user will be based on his knowledge of the target

* The work was carried out as part of a project to create a trusted environment for the development of artificial intelligence systems [39][40].

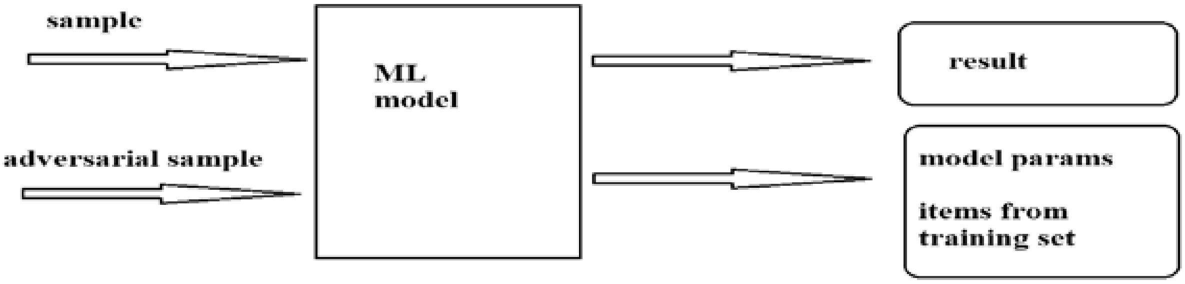


Fig. 1. How adversarial users use ML model

model (0 knowledge - black box, has some knowledge – grey box or white box), and use some well-prepared samples(called adversarial samples) to attack model. As result, adversarial users can get the model's parameters or some samples that are used to train this model.

1.2. Why do we research this kind of attack?

As we said in the Introduction, model inversion attack becomes a huge threat to the ML model, and so far we don't have a mature defensive strategy that can defend against several existing model inversion attacks. So, for the safety and robustness of ML models, we have to learn model inversion attacks, trying to understand how this kind of attack affects the ML model. Then, based on this information and acknowledgment, we might be able to create some effective countermeasures or improve our ML models.

More mathematical definition for MI attack

In [1] has a more mathematical definition of MI attack:

Theorem 1. MI attack is a machine learning problem and specified as a triple (*Z, H, l*), (*Z*, a sample space; *H*, hypothesis space; *l: H*×*Z -> R*, a loss function) with the following notations:

1) Γ: A training algorithm of the learning problem, which outputs a hypothesis Γ(S) ∊ H on an input training set S.
2) D_S : A distribution over the training set S.
3) τ: The objective function computed by the adversary. For now, one can view it simply as some function that maps *Z* to $\{0, 1\}^*$.
4) gen, sgen: Auxiliary information generators. They map a pair (S, z) to an advice string in $\{0, 1\}^*$.

The MI-attack world is described by a tuple (A, gen, τ, S, D_S, Γ), where the adversary (A) is a probabilistic oracle machine. The following game is played between the Nature and the Adversary A.

$$\text{gain}\,(A, \text{gen}, \tau, S, D_S, \Gamma) = \Pr[A^{\Gamma(S)}(\text{gen}(S, z))= \tau(z)] \quad (1)$$

where the probability is taken over the randomness of z~ D_S, the randomness of gen, and the randomness of A.

The simulated world is described by a tuple (A^*, sgen, τ, S, D_S), where the adversary (A^*) is a non-oracle machine and sgen is the second auxiliary information generator. The game between the Nature and A^* is:

$$\text{sgain}(A^*, \text{sgen}, \tau, S, D_S) = \Pr[A^* (\text{sgen}(S,z))= \tau(z)] \quad (2)$$

where the probability is taken over the randomness of z~ D_S, the randomness of sgen, and the randomness of A^*.

1.3. Classification of MI attacks

For the needs of taxonomy, typically, we can classify client-side access as being either *black-box* or *white-box*. In a *black-box* setting, an adversarial customer will create prediction queries against a model, however not transfer the model description. In a *white-box* setting, an adversarial customer is allowed to transfer an outline of the model.

1.4. Attacks & countermeasures

The Fredrikson et al. attack Widely accepted, the first MI attack was realized by *Fredrikson et al.* in [1]. The *Fredrikson et al. attack* is to use auxiliary information and confidence value given by the model prediction to guess the true value of the privacy-sensitive feature. The weakness of the attack is also clearly: 1) adversary knows the target feature(for example, its domain), a large domain means more combinations adversary have to try; 2) adversary have to know as much auxiliary information as he can, same reason as 1), less information about the victim means more combinations to try; 3) adversary has infinite access to the model because he has to correct the guessing value with the help of confidence value, a higher confidence value means our guessing value is closer to the true value. In a word, the *Fredrikson et al. attack* can realize only under many restrictions.

The author gives countermeasures for both decision trees and facial recognition. For decision trees, the level at which the sensitive feature occurs may affect the accuracy of the attack, and it may be possible to design more sophisticated training algorithms that incorporate model inversion metrics into the splitting criteria to achieve resistance to attacks without unduly sacrificing accuracy.

For facial recognition, one possible defense is to degrade the quality or precision of the gradient information retrievable from the model, and also, black-box facial recognition models can produce confidence scores that are useful for many purposes while remaining resistant to reconstruction attacks.

MI attack for deep networks *Fredrikson et al.* established that model inversion attacks include a smart performance on decision tree and face recognition [1]. However, for deep networks, these procedures sometimes cause unidentifiable representations that square measure useless for the opponent [4]. So that they introduced a more realistic definition of model inversion and leveraged properties of generative adversarial networks for constructing a connected lower-dimensional manifold.

MI attack: wherever the opponent is attentive to the final purpose of the attacked model (for instance, whether or not it's an associate degree OCR system or an automatic face recognition system), and the goal is to seek out realistic category representations among the corresponding lower-dimensional manifold (of, separately, general symbols or general faces).

In [4], the approach is based on Generative Adversarial Network. A Generative Adversarial Network(GAN) is a min-max game between two neural networks: generator (G_θ) and discriminator (D_φ). The generator (G_θ) takes random noise z as input and generates ($G_\theta(z)$). The discriminator (D_φ) distinguishes between real samples x and fake samples coming from (G_θ). The objective function for the min-max game between (G_θ) and (D_φ) is:

$$min_\theta max_\phi \mathrm{E}_{x\sim P(x)}\left[\log\left(D_\phi(x)\right)\right]+$$
$$+\mathrm{E}_{z\sim P(z)}\left[1-\log\left(D_\phi\left(G_\theta(z)\right)\right)\right] \qquad (3)$$

where P_x is the real data distribution, and P_z is a noise distribution which is typically a uniform distribution or a normal distribution.

Obviously, different images should belong to their disconnected manifolds without ant paths of "blended" images between them. However, in GAN, the generator function maps from a connected distribution space to all possible outputs, which results in a connected output set of instances. This is an emblematical disadvantage of GANs and various techniques to partition the input into disjoint support sets have been used to address this issue. And [4]'s approach is to leverage this drawback to search in the low-dimensional space P_x (real data distribution) of all possible images.

With some natural knowledge about the underlying target system, an attacker can use this GAN-based approach for retrieving representative and recognizable samples of individual classes. For the countermeasures, the author suggests that a security-based biometric identification system might classify away the larger set of faces so that the faces that are relevant to security verification are effectively hidden sort like a needle in a very rick. The key downside here is to take care of adequate classifier accuracy because the variety of categories will increase.

Also, in the conclusion, the author proposes a prospective research direction is to consider ways to develop a robust defense against model inversion attacks without affecting the model accuracy. This could be difficult since model inversion doesn't involve protecting any specific instance, and the defense should protect all the representative pictures that are part of the manifold used for training.

Generative Model-Inversion Attack For deep neural networks, there is another attack method that uses GAN. In [6], the author presents a novel attack method, termed the generative model-inversion attack, which can reverse deep neural networks with high success rates. Rather than reconstructing private training data from scratch, the author leverage partial public information, which can be generic, to find out a distributional prior via generative adversarial networks (GANs) and use it to guide the inversion method (Fig.2). The author also shows that differential priva-

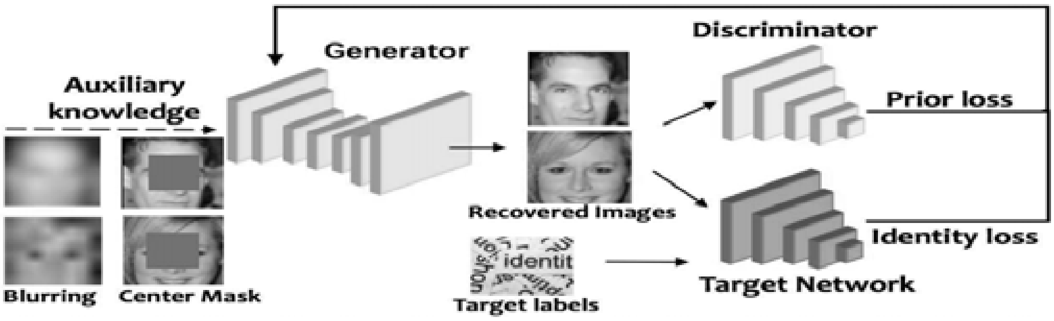


Fig. 2. Overview of the proposed GMI attack method [6]

cy, in its canonical form, is of little avail to defend against their attack.

In this paper, the author focuses on the white-box setting, where the adversary is assumed to have access to the target network *f* and employs some inference technique to discover the options *x* related to a particular label *y*. To recover those images, the reconstruction process consists of two-stage: (1) Public knowledge distillation, which trains the generator and therefore the discriminators on public datasets to encourage the generator to get realistic-looking pictures. The public datasets may be untagged and don't have any identity overlapping with the personal dataset;(2) Secret revelation, in which create use of the generator obtained from the primary stage associated solve an optimization drawback to recover the missing sensitive regions in a picture. In stage 1, the author leverage the canonical Wasserstein-GAN training loss:

$$min_G max_D L_{wgan}(G, D) = \\ E_x[D(x)] - E_z[D(G(z))] \tag{4}$$

when the auxiliary knowledge(blurred or corrupted version of the private image) is available to the attacker, let the generator take the auxiliary knowledge as an additional input. In the second stage, the author solves the following optimization to find the latent vector that generates an image achieving the maximum likelihood under the target network while remaining realistic:

$$\hat{z} = argmin_z L_{prior}(z) + \lambda_i L_{id}(z) \tag{5}$$

where the prior loss L_{prior} (z) penalizes surrealistic pictures and therefore the identity loss L_{id}(z) encourages the generated pictures to own high chance below the targeted network. L_{prior}(z)and L_{id}(z) are defined as follow:

$$L_{prior}(z) = -D(G(z)) \tag{6}$$

$$L_{id}(z) = -log[C(G(z))] \tag{7}$$

where C(G(z)) represents the probability of G(z) output by the target network. The experiments show that this GMI attack has a better performance than the *Fredrikson et al.* attack, which means it is an effective attack method. There are no countermeasures to this GMI from the author.

MI attack that using explanations As the ML models are widely used, people need not only answers from the ML model but also explanations. Explainable artificial intelligence (XAI) provides additional info to assist users to grasp model selections, however further information exposes additional risks for privacy attacks [ref5]. In this paper, the author studies this risk for image-based model inversion attacks and identified several attack architectures with increasing performance to reconstruct private image data from model explanations. And these XAI-aware inversion models were designed to use spatial data in image explanations.

Fig. 3 presents architectures of inversion attack models:

Here, a) Baseline threat model with target CNN model M_t to predict emotion $\tilde{y}_t$ from face *x*, and inversion attack model to reconstruct face $\widetilde{x_r}$ from emotion. Emotion prediction confidences are input to a transposed CNN (TCNN) for inversion attack (d). b) Threat model with explainable target model that also provides instance explanation $\widetilde{E_t}$ of the target prediction, and XAI-aware multi-modal inversion attack model that inputs $\widetilde{E_t}$ via different input architectures: e) Flattened $\widetilde{E_t}$ concatenated with $\tilde{y}_t$, f) U-Net for dimensionality reduction and spatial knowledge, g) combined Flatten and U-Net. c) Threat model with non-explainable target model and inversion attack model that predicts a reconstructed surrogate explanation $\widetilde{E_r}$ from target prediction $\tilde{y}_t$ and uses $\widetilde{E_r}$ for multi-modal image inversion (e-g). Flattened $\widetilde{E_t}$ concatenated with $\tilde{y}_t$, f) U-Net for dimensionality reduction and spatial knowledge, g) combined Flatten and U-Net. c) Threat model with non-explainable target model and inversion attack model that predicts a reconstructed surrogate explanation $\widetilde{E_r}$ from target prediction $\tilde{y}_t$ and uses $\widetilde{E_r}$ for multi-modal image inversion (e-g).

The author divided MI attacks into 3 types: 1) model inversion with Target Explanations; 2) model inversion with Multiple Explanations; 3) model inversion with Surrogate Explanations. For type 1, the author trained the inversion attack model as a Transposed CNN(TCNN) to predict a 2D image from the 1D target prediction vector as input to the attack model. The model is trained with MSE. For type 2, the author exploits Alternative CAMs(Σ-CAM) by concatenating explanations for |*C*| classes into a 3D tensor and training the inversion models on this instead of the 2D matrix of a single explanation. There is no information about countermeasures.

An inversion-specific GAN for MI attack In the paper [7], the author presents a novel inversion-specific GAN that can better distill knowledge useful for performing attacks on private models from public data. In particular, the discriminator is trained to differentiate not only the real and fake samples but the soft labels provided by the target model. Experiments show that the combination of these techniques can significantly boost the success rate of the state-of-the-art MI attacks by 150%, and generalize better to a variety of datasets and models.

Author focus on white-box setting MI attack. The goal of the attacker is to discover a representative

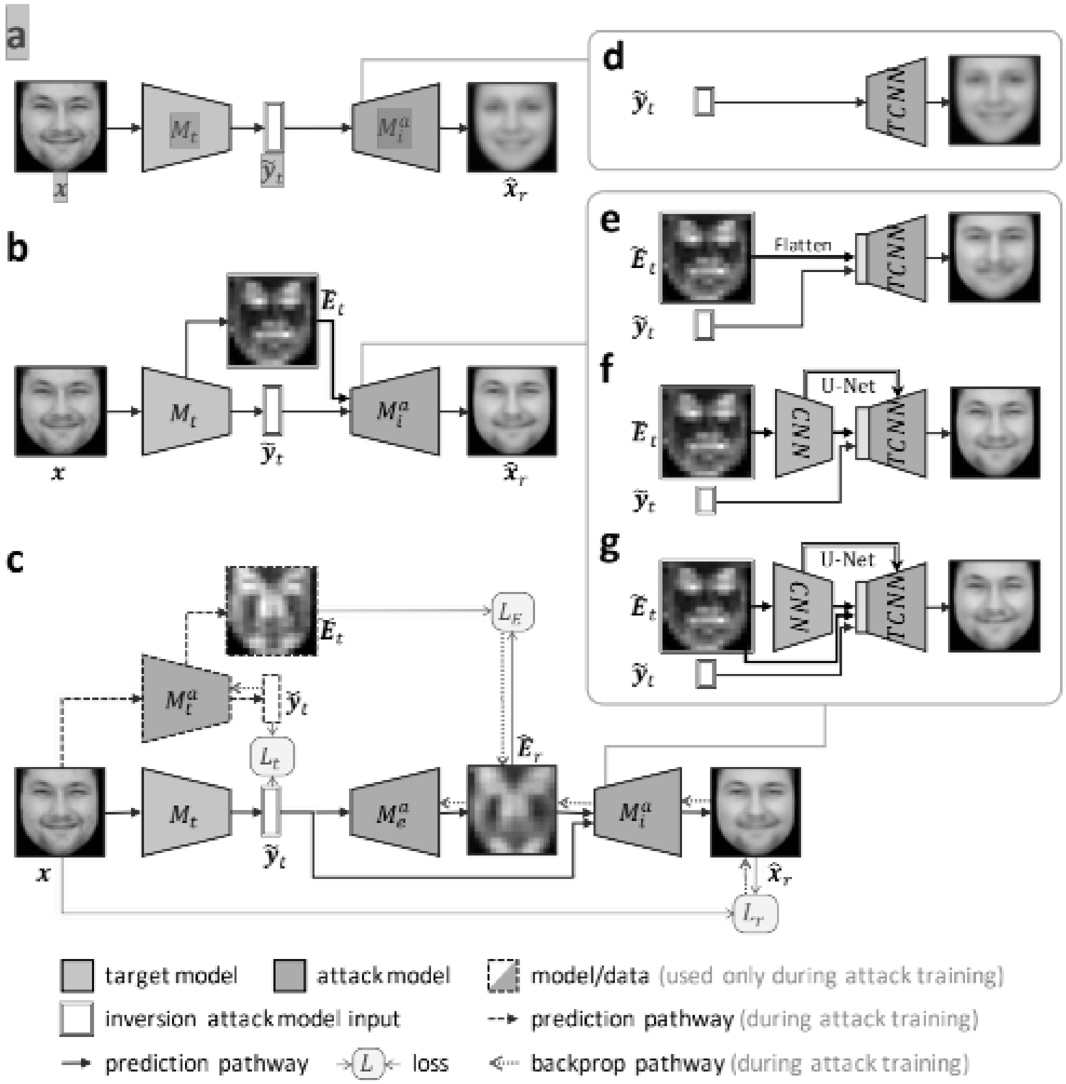


Fig. 3. Architectures of inversion attack models [5]

input feature x associated with a specific label *y*. The author uses facial recognition as a running example for the target network. The proposed attack algorithm consists of two steps. The first step is to train a GAN to have information concerning the personal categories of the target model from public knowledge. Rather than training a generic GAN, the author customizes the training objective for each generator and discriminator thus on higher distill the private-domain data concerning the target model from public knowledge. In the second step, the author makes use of the generator learned in the first step to estimate the parameters of the private data distribution.

Author tested this threat model on several datasets with the baseline. Experiments show that this approach can significantly improve the performance of GMI on all target models. Countermeasures are not mentioned in this paper.

MI attack against collaborative inference Most studies solely targeted knowledge privacy throughout training and neglected privacy throughout illation. During this paper [8], the author devises a brand new set of attacks to compromise inference data privacy in cooperative deep learning systems. Specifically, once a deep neural network and also the corresponding illation task are split and distributed to completely different participants, one malicious participant has the ability to accurately recover any input fed into this system, although he has no access to different participants' information or computations, or to prediction APIs to query this system.

Author considers a collaborative inference system between two participants, P_1 and P_2. The target model is split into two parts: $f_\theta = f_{\theta 2} \cdot f_{\theta 1}$. P_1 performs earlier layers $f_{\theta 1}$, and P_2 performs $f_{\theta 2}$. P_1 is trusted and P_2 is untrusted.

Algorithm 1 White-box model inversion attack

```
function WhiteBoxAttack ($f_{01}, f_{01}(x_0), T, \lambda, \varepsilon$)
/* $f_{01}$ – the target model */
/* $f_{01}(x_0)$ – the intermediate output of sensitive input x0 */
/* $T$ - maximum number of iterations */
/* $\lambda$ – tradeoff between prior and posterior information */
/* $\varepsilon$ – step size if GD */

$L(x) = \|f_{01}(x) - f_{01}(x_0)\|_2^2 + \lambda TV(x)$
$t$=0
$x^{(0)}$ = ConstantInit()
    while (t<T) do
        $x^{t+1} = x^t - \varepsilon * \frac{\partial L(x^t)}{\partial x^t}$
        t = t+1
    end while
  return $x^{(T)}$
end function
```

In the experiments, the results show that different split points can yield different attack effects, so the question is: how to split the neural network in the collaborative system, to make the inference data more secure? Generally, it is observed that the quality of recovered images decreases when the split layer goes deeper. This is straightforward as the relationship between input and output becomes more complicated and harder to revert when there are more layers. Besides, it is also observed that the image quality drops significantly, both qualitatively and quantitatively, on the fully-connected layer (fc1), indicating that model inversion with fully-connected layers is much harder than for convolutional layers. The reason is that a convolutional layer only operates on local elements (the locality depends on the kernel size), while a fully-connected layer entirely mixes up the patterns from the previous layer. Besides, the number of output neurons in a fully-connected layer is typically much smaller than input neurons. So it is relatively harder to find the reversed relationship from the output of the fully-connected layer to the input. And the first defense method is running fully-connected layers before sending out results.

Other possible defenses are making client-side networks deeper, trusted execution on untrusted participants differential privacy, and homomorphic encryption.

Improving robustness to MI attack In the paper [9], the author proposed the Mutual Information Regularization based Defense (MID) against MI attacks. The key idea is to limit the information about the model input contained in the prediction, thereby limiting the ability of an adversary to infer the private training attributes from the model prediction.

The author limits the dependency between X and $\hat{Y}$ to prevent the adversary from inferring the training data distribution associated with a specific label. The author's idea is to quantify the dependence between X and $\hat{Y}$ using their mutual information $I(X;\hat{Y})$ and incorporate it into the training objective as a regularizer. This defense, which is called MID, trains the target model via the loss function:

$$min_{f\in H} E_{(x,y)\sim p_{X,Y}}(x,y)$$
$$[L(y,f(x))] + \lambda I(X,\hat{Y}) \tag{8}$$

where

$$I(X;\hat{Y}) = \int_X \int_Y p_{X,Y}(x,y) log\left(\frac{p_{X,Y}(x,y)}{p_X(x)p_Y(y)}\right) \tag{9}$$

$L(y,f(x))$ is the loss function for the main prediction task, and λ is the weight coefficient that controls the tradeoff between privacy and utility on the main prediction task.

To deconstruct the proposed regularizer, mutual information is as follows:

$$I(X;\hat{Y}) = H(\hat{Y}) - H(\hat{Y}|X) \tag{10}$$

When f is a deterministic model, $H(\hat{Y}|X) = 0$ and introducing the mutual information regularizer effectively reduces the entropy of the model output, i.e., $H(\hat{Y})$. When f is stochastic, the regularizer will additionally promote the uncertainty of the model output for a fixed input, i.e., $H(\hat{Y}|X)$.

For Linear Regression, due to the deterministic nature of the model, the mutual information regularizer is reduced to $H(\hat{Y})$. Approximation of $\hat{Y}$ by a Gaussian mixture is:

$$p(\hat{y}) = \frac{1}{N}\sum_{i=1}^{N} \mathcal{N}(\hat{Y}|Ax_i;\sigma^2) \tag{11}$$

where ${x_i}_{i=1}^{N}$ is the training set and σ is a free parameter. The author utilizes a Taylor-expansion based approximation for the entropy of Gaussian mixtures described in *H*[illegible]*l.* and derive the following approximation to $I(X,\hat{Y})$:

$$\widetilde{I_{Lin}}(X,\hat{Y}) =$$
$$= -\frac{1}{N}\sum_{i=1}^{N} log\left(\frac{1}{N}\sum_{i=1}^{N}\frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{1}{2}\left(\frac{Ax_i - Ax_j}{\sigma}\right)^2\right)\right) \tag{12}$$

For the decision tree, the author modifies ID3 [31] to incorporate the mutual information regularizer. Since decision trees trained with ID3 are deterministic, the mutual information regularizer again reduces to $H(\hat{Y})$. To defend against the MI attacks, the author adds $-H(\hat{Y})$ into the splitting criterion.

For networks, the author gets inspiration from the line of work on information bottleneck [32] [33] and regard the neural network as a Markov chain $Y - X - Z - \hat{Y}$, where X is the feature, Y is the ground truth label, Z is a stochastic encoding of the input X at some intermediate layer and defined by $P(Z|X;\sigma)$, and $\hat{Y}$ is the prediction. The author replaces $I(X,\hat{Y})$ with upper bound $I(X,Z)$ in the training object and train the neural network with loss function:

$$min_\theta - I(Z;Y) + \lambda I(Z,X) \qquad (13)$$

The experiments show that this MID approach can significantly reduce the attack accuracy and effectively protect the ML model from MI attacks.

A simple additive noise method to defend MI attack In [10], the author demonstrates that the attack can be successfully performed with limited knowledge of the data distribution by the attacker, and show that NoPeekNN, an existing defensive methodology, protects completely different info from exposure, suggesting that a combined defense is important to completely shield personal user information.

NoPeekNN may be a technique for limiting knowledge reconstruction in SplitNNs by minimizing the gap correlation between the input data and the intermediate tensors throughout model training [34]. NoPeekNN optimizes the model by a weighted combination of the task's loss and a distance correlation loss, that measures the similarity between the input data and the intermediate data. NoPeekNN's loss weighting is governed by a hyperparameter $\alpha \in [0,\infty)$. While NoPeekNN was shown to cut back autoencoder's ability to reconstruct input information, it's not been applied to adversarial model inversion attack.

Similar to this work, to defend against model inversion attack on one-dimensional ECG data, [35] utilizes noise to the intermediate tensors in a SplitNN. The authors pack this defense as a differential privacy mechanism [36]. However, in that work, the addition of noise greatly impacts the model's accuracy for even modest epsilon values (98.9% to roughly 90% at $\epsilon = 10$). There is also a similar method introduced by [37] called Shredder. To minimize mutual information between input and intermediate data, this method will adaptively generate a noise mask.

In this work, the author considers an honest-but-curious computation server and an arbitrary number of data owners who run the correct computations during training and inference. At least one party attempts to steal input data from alternative parties by employing a model inversion attack. The attack method is as follows: 1) The attackers collect a dataset of inputs (raw data) and intermediate data made by the first model phase. 2) To convert the intermediate information into raw input data, they train an attack model. 3) They collect intermediate information made by some information owners and run it through the trained attack model to reconstruct the raw input information. This attack is considered a "black-box" since the internal parameters of the data owner model segment are not used in the attack. The author assumes that the model training method has been orchestrated by a third party in which there's just one computational server.

MI attack for large language models In [11], the author tries to extract text data from a language model trained on scrapes of the public Internet called GPT-2. Training data extraction attacks are usually seen as theoretical or academic and therefore unlikely to be exploitable in application. This can be even by the prevailing intuition that privacy leakage is correlated with overfitting, and since advanced LMs trained on massive (near terabyte-sized) datasets for a few epochs, they tend to not overfit. This paper proved that training data extraction attacks are viable.

First, is the definition of committal to memory. The author defines eidetic memorization as a special type of memorization. Unofficially, eidetic memorization is data information that has been memorized by a model despite solely showing during a tiny set of training instances. The fewer training samples that contain the information, the stronger the eidetic memorization is.

Theorem 2. A string s is extractable from an *LM* f_θ if there exists a prefix *c* such that:

$$s \leftarrow argmax_{s':|s'|=N} f_\theta(s'|c) \qquad (14)$$

Fig. 4 presents the structure of extraction attack:

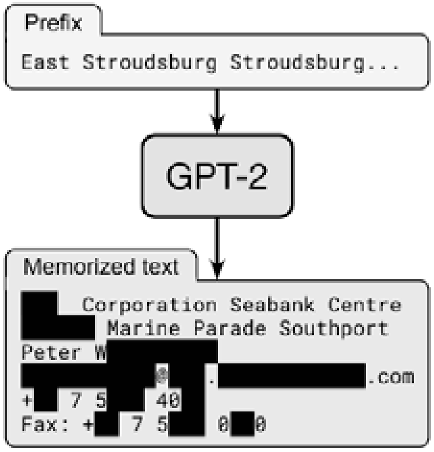

Prefix
East Stroudsburg Stroudsburg...
GPT-2
Memorized text
[redacted] Corporation Seabank Centre
[redacted] Marine Parade Southport
Peter W[redacted]
[redacted]@[redacted].[redacted].com
+[redacted] 7 5[redacted] 40
Fax: +[redacted] 7 5[redacted] 0[redacted]0

Fig. 4. Extraction attack. Given query access to a neural network language model, we extract an individual person's name, email address, phone number, fax number, and physical address. The example in this figure shows information that is all accurate so we redact it to protect privacy [11]

Theorem 3. A string s is k-eidetic memorized (for k ≥ 1) by an $LM\ \ f_\theta$ if s is extractable from f_θ and s appears in at most k examples in the training data $X\colon |x \in X\colon s \subseteq x| \leq k$. This threat model is extremely realistic as several LMs are available through black-box APIs. For example, the GPT-3 model created by OpenAI is available through black-box API access. Auto-complete models trained on actual user information have conjointly been created public, though they reportedly use privacy-protection measures throughout training.

The extraction of training data from a language model has two-step: 1) Generate text. Generate a large quantity of data by unconditionally sampling from the model; 2) Predict which outputs contain memorized text. We next remove the generated samples that are unlikely to contain memorized text using a membership inference attack. These 2 steps correspond on to extracting model information (Theorem 2), then predicting that strings may be k-eidetic memorization (Theorem 3).

Fig. 5 presents the workflow of extraction attack and evaluation:

In total across all strategies, the author identify 604 unique memorized training examples from among the 1,800 possible candidates, for an aggregate true positive rate of 33.5% (the best variant has a true positive rate of 67%).

For countermeasures, the author suggests that needs to be techniques developed to specifically address their attacks. Training with differentially private techniques is one method for mitigating privacy leakage, however, the author believes that it will be necessary to develop new methods that can train models at this extreme scale (e.g., billions of parameters) without sacrificing model accuracy or training time.

Defending MI attack via prediction purification In [12], the author proposes a unified approach, namely a purification framework, to defend data inference attacks. It purifies the confidence score vectors foretold by the target classifier by reducing their dispersion. The setup may be additional specialized in defensive a selected attack via adversarial learning.

The model owner trains a machine learning classifier F on its training dataset D_{train} and test F on validation dataset D_{val}. Both D_{train} and D_{val} are drawn from the same underlying data distribution $p_r(X)$.
The attacker aims at performing data inference attacks against the target classifier F. Consider that the classifier F works as a black-box "oracle" to the attacker, i.e., the attacker can only query F with its data sample x and obtain the prediction scores $F(x)$. The attacker is also assumed to have auxiliary information $\mathcal{A}$. Given a prediction vector $F(x)$ on some victim data point x, the attacker wants to find an attack function $A(F(x), O(F), \mathcal{A})$ for membership inference:

$$A(F(x), O(F), \mathcal{A}) = m \in \{0,1\} \tag{15}$$

for model inversion:

$$A(F(x), O(F), \mathcal{A}) = \hat{x} \tag{16}$$

where $O(F)$ represents the attacker's blackbox access to the oracle classifier F.

For purification, the base of purification is purifier G. The author designed G as an autoencoder and is used to reduce the dispersion of the confidence scores as well as to preserve the utility of the classifier. G is trained on the confidence scores predicted by F on the defender's reference dataset D_{ref}. The author trained G to also produced the label predicted by F by adding a cross-entropy loss function. G is trained to minimize the function:

$$\mathcal{L}(G) =$$
$$= E_{x\sim p_r(x)}\left[\mathcal{R}\left(G\big(F(x)\big), F(x)\right) + \lambda\mathcal{L}\left(G\big(F(x), argmaxF(x)\big)\right)\right] \tag{17}$$

The author also provides specialized G for MI attack.

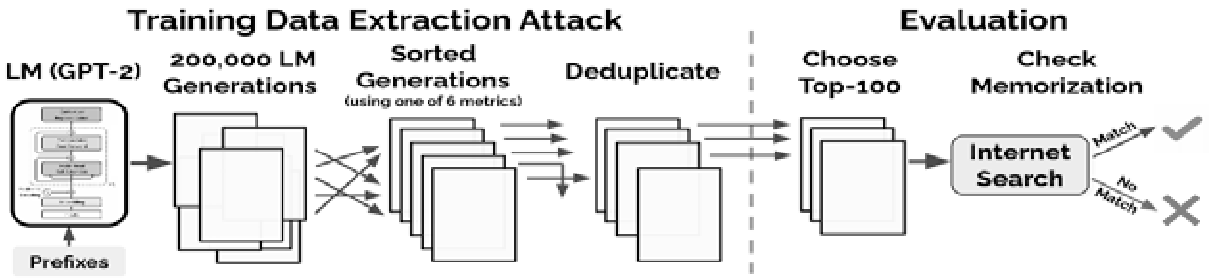


Fig.5. 1) Attack. We begin by generating many samples from GPT-2 when the model is conditioned on (potentially empty) prefixes. We then sort each generation according to one of six metrics and remove the duplicates. This gives us a set of potentially memorized training examples. 2) Evaluation. We manually inspect 100 of the top-1000 generations for each metric. We mark each generation as either memorized or not-memorized by manually searching online, and we confirm these findings by working with OpenAI to query the original training data [11]

For each classification task, we can see that the single purifier is able to concurrently decrease the inference accuracy and increase the inversion error as well as preserve the classification accuracy. The purifier has almost no defense effect against the Label attack where the confidence score information is not leveraged.

MLPrivacyGuard In [18], the author presents MLPrivacyGuard, a countermeasure against black-box MI attack. In this countermeasure, they add controlled noise to the output of the confidence function. The author demonstrates that when noise is injected with a long-tailed distribution, the objectives of low misclassification error with a strong defense can be attained as model inversion attacks are neutralized because numerical approximation of gradient ascent is unable to converge.

MLPrivacyGuard is a measure to forestall model inversion attacks that don't need retraining or any type of modification to the ML system's inner workings. MLPrivacyGuard relies on the injection of long-tailed distributed errors to the output $\tilde{f}$ of the model, so that a model inversion attack can not converge, whilst maintaining the functionality of the ML system reliable for legitimate users. The idea behind MLPrivacyGuard is that when the confidential information has some level of randomness the model inversion attack is unable to converge in black-box systems. The reason for this is straightforward: since a black-box model inversion has to rely on numerical approximations of the gradient $\tilde{f}(x)$, which are obtained by numerical differentiation of $\tilde{f}$ on each of the features, the attack is unable to converge if the value of $\tilde{f}$ is nondeterministic.

To preserve correct classifications for legitimate users, the author guaranteed this with the distributions picked for the random errors. In the experiments, the error size has exponential distribution, i.e. the absolute value of the error injected is x with probability $\lambda e^{-\lambda x}$.

The result of experiments shows that this MLPrivacyGuard approach increases the classification error rate at most by 2% while defeating adversarial model inversion attacks.

MI attack without knowledge of non-sensitive attributes In [15], the author proposes a General Model Inversion (GMI) framework to capture the scenario where knowledge of the non-sensitive attributes is not necessarily provided. This framework also captures the scenario of *Fredrikson et al.*, notably, it enables a new type of model inversion attack that infers sensitive attributes without the knowledge of non-sensitive attributes by modifying the ML model into a target ML model via data poisoning. The GMI attack is defined by a tuple of three algorithms: Setup, Poisoning and ModelInversion [15]. Fig. 6 presents the workflow of GMI attack:

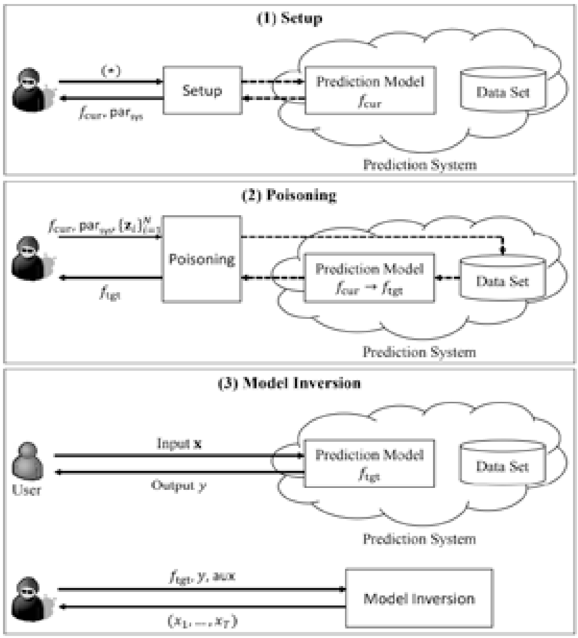


Fig. 6. Workflow of GMI attack [15]

Leverage the functional mechanism to prevent MI attack In [16], the author develops a novel approach that leverages the functional mechanism to perturb coefficients of the polynomial representation of the objective function but effectively balances the privacy budget for sensitive and non-sensitive attributes in learning the differential privacy preserving regression model.

This approach leverages the functional mechanism proposed in [38] but perturbs the polynomial coefficients of the objective function with different magnitudes of noise. This approach can effectively weaken the correlation between the sensitive attributes with the output to prevent model inversion attacks whereas retaining the utility of the released model by decreasing the perturbation effect on non-sensitive attributes.

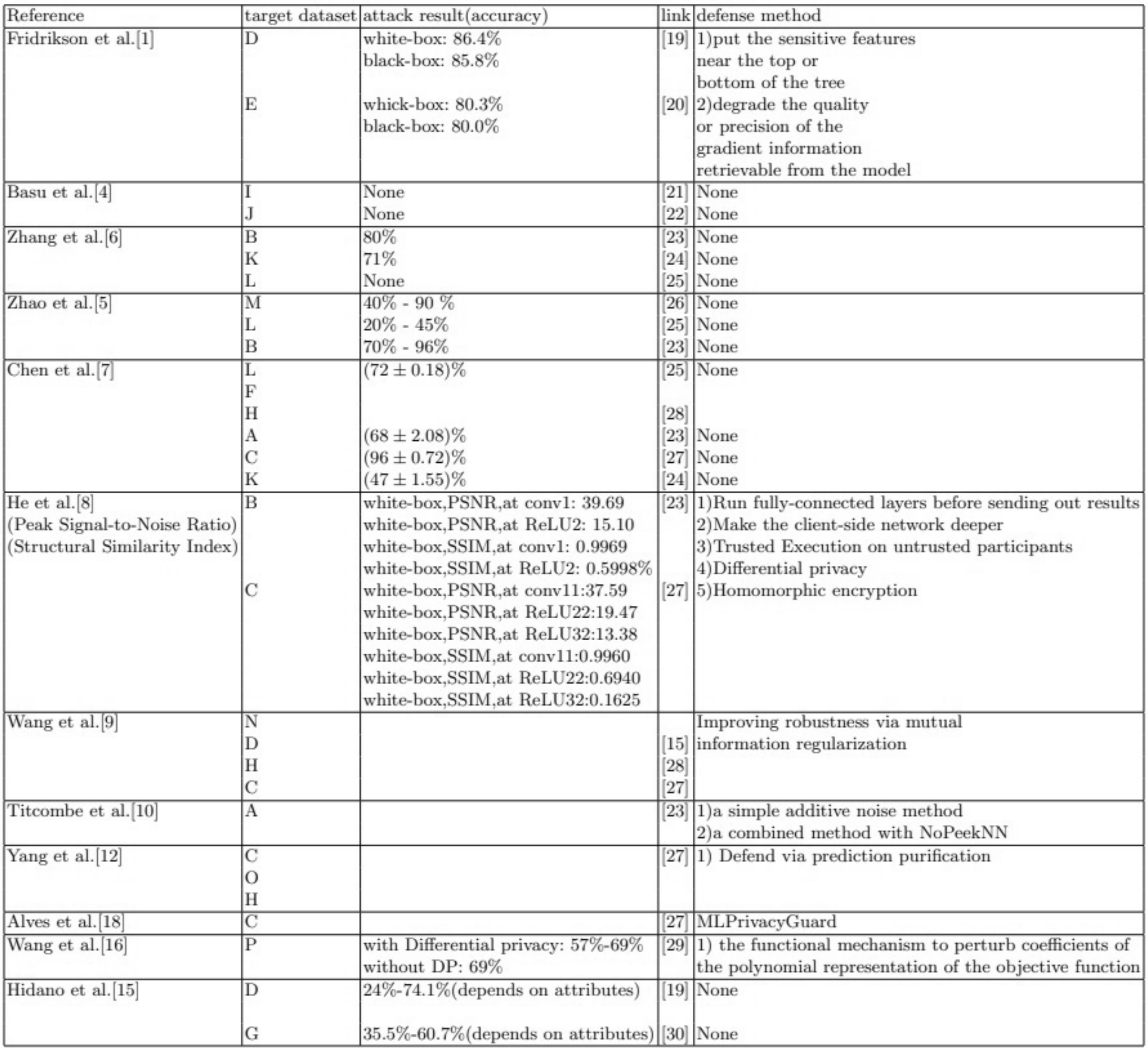
2. Table of MI attacks with their attributes

In this section, we summarize the results of our review (Table 1). Presented datasets are: A - MNIST, B - MNIST handwritten digit, C - CIFAR-10, D - FiveThirtyEight survey (How americans like their steak), E - GSS marital happiness survey, F - Flickr-Faces-HQ(FFHQ), G - MovieLens 1M Dataset, H - FaceScrub, I - Numeric MNIST, J- Fashion MNIST, K - ChestX-ray8, L - CelebA, M - iCV-MEFED, N - IPWC, O - Purchase100, P - Adult dataset.

As can be seen from this table, attacks clearly prevail over defenses. In reality, only the practical feasibility of attacks really protects existing systems. All attacks require multiple polling of models. If this is not

Table 1

Summary of MI attacks



Reference	target dataset	attack result(accuracy)	link	defense method
Fridrikson et al.[1]	D	white-box: 86.4% black-box: 85.8%	[19]	1)put the sensitive features near the top or bottom of the tree
	E	whick-box: 80.3% black-box: 80.0%	[20]	2)degrade the quality or precision of the gradient information retrievable from the model
Basu et al.[4]	I	None	[21]	None
	J	None	[22]	None
Zhang et al.[6]	B	80%	[23]	None
	K	71%	[24]	None
	L	None	[25]	None
Zhao et al.[5]	M	40% - 90 %	[26]	None
	L	20% - 45%	[25]	None
	B	70% - 96%	[23]	None
Chen et al.[7]	L	(72 ± 0.18)%	[25]	None
	F			
	H		[28]	
	A	(68 ± 2.08)%	[23]	None
	C	(96 ± 0.72)%	[27]	None
	K	(47 ± 1.55)%	[24]	None
He et al.[8] (Peak Signal-to-Noise Ratio) (Structural Similarity Index)	B	white-box,PSNR,at conv1: 39.69 white-box,PSNR,at ReLU2: 15.10 white-box,SSIM,at conv1: 0.9969 white-box,SSIM,at ReLU2: 0.5998%	[23]	1)Run fully-connected layers before sending out results 2)Make the client-side network deeper 3)Trusted Execution on untrusted participants 4)Differential privacy
	C	white-box,PSNR,at conv11:37.59 white-box,PSNR,at ReLU22:19.47 white-box,PSNR,at ReLU32:13.38 white-box,SSIM,at conv11:0.9960 white-box,SSIM,at ReLU22:0.6940 white-box,SSIM,at ReLU32:0.1625	[27]	5)Homomorphic encryption
Wang et al.[9]	N			Improving robustness via mutual information regularization
	D		[15]	
	H		[28]	
	C		[27]	
Titcombe et al.[10]	A		[23]	1)a simple additive noise method 2)a combined method with NoPeekNN
Yang et al.[12]	C O H		[27]	1) Defend via prediction purification
Alves et al.[18]	C		[27]	MLPrivacyGuard
Wang et al.[16]	P	with Differential privacy: 57%-69% without DP: 69%	[29]	1) the functional mechanism to perturb coefficients of the polynomial representation of the objective function
Hidano et al.[15]	D	24%-74.1%(depends on attributes)	[19]	None
	G	35.5%-60.7%(depends on attributes)	[30]	None

an MLaaS system, then it will be impossible to carry out an attack directly.

3. Summary and future works

From all this information above we find that several types of MI attacks have been created and successful test on various datasets like CIFAR-10, MNIST, FiveThirtyEight, and so on. But the problem is, those existing countermeasures are passive counter, which means these countermeasures are just been applied in the ML model, and each countermeasure can only defend a specific attack method. Considering there are many attack methods and these methods can also be iterated, if one ML model wants to survive under those attacks, it has to apply many countermeasures simultaneously. We think that this approach may lower not only the efficiency of the ML model but also the accuracy.

So, if we can build a MI attack detector, and this detector can immediately cut off the connection between user and model when it detects a MI attack(or some action similar to an MI attack), it will be great, and it can save much cost for MLaaS provider. In our opinion, this is a promising direction.

For the detector, we want to start from GAN. In GAN there is a generator G and discriminator D, we can use G to simulate existing attack methods and let D discriminate whether one is a MI attack or not. Also, if possible, we can import CNN to our detector. For images, CNN can learn its features; and for MI attacks, maybe we can transform MI attack into a type that can let CNN learn its features.

References

1. *Fredrikson, M., Jha, S., & Ristenpart, T.* (2015, October). Model inversion attacks that exploit confidence information and basic countermeasures. In Proceedings of the 22nd ACM SIGSAC conference on computer and communications security (pp. 1322-1333).
2. *Wu, X., Fredrikson, M., Jha, S., & Naughton, J.F.* (2016, June). A methodology for formalizing model-inversion attacks. In 2016 IEEE 29th Computer Security Foundations Symposium (CSF) (pp. 355-370). IEEE.
3. *Yeom, S., Giacomelli, I., Fredrikson, M., & Jha, S.* (2018, July). Privacy risk in machine learning: Analyzing the connection to overfitting. In 2018 IEEE 31st Computer Security Foundations Symposium (CSF) (pp. 268-282). IEEE.
4. *Basu, S., Izmailov, R., & Mesterharm, C.* (2019). Membership model inversion attacks for deep networks. arXiv preprint arXiv:1910.04257.
5. *Zhao, X., Zhang, W., Xiao, X., & Lim, B.Y.* (2021). Exploiting Explanations for Model Inversion Attacks. arXiv preprint arXiv:2104.12669.
6. *Zhang, Y., Jia, R., Pei, H., Wang, W., Li, B., & Song, D.* (2020). The secret revealer: Generative model-inversion attacks against deep neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 253-261).
7. *Chen, S., Kahla, M., Jia, R., & Qi, G.J.* (2021). Knowledge-Enriched Distributional Model Inversion Attacks. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 16178-16187).
8. *He, Z., Zhang, T., & Lee, R.B.* (2019, December). Model inversion attacks against collaborative inference. In Proceedings of the 35th Annual Computer Security Applications Conference (pp. 148-162).
9. *Wang, T., Zhang, Y., & Jia, R.* (2020). Improving robustness to model inversion attacks via mutual information regularization. arXiv preprint arXiv:2009.05241.
10. *Titcombe, T., Hall, A. J., Papadopoulos, P., & Romanini, D.* (2021). Practical Defences Against Model Inversion Attacks for Split Neural Networks. arXiv preprint arXiv:2104.05743.
11. *Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K. & Raffel, C.* (2021). Extracting training data from large language models. In 30th USENIX Security Symposium (USENIX Security 21) (pp. 2633-2650).
12. *Yang, Z., Shao, B., Xuan, B., Chang, E. C., & Zhang, F.* (2020). Defending model inversion and membership inference attacks via prediction purification. arXiv preprint arXiv:2005.03915.
13. *Ateniese, G., Mancini, L. V., Spognardi, A., Villani, A., Vitali, D., & Felici, G.* (2015). Hacking smart machines with smarter ones: How to extract meaningful data from machine learning classifiers. International Journal of Security and Networks, 10(3), 137-150.
14. *Xu, R., Baracaldo, N., & Joshi, J.* (2021). Privacy-Preserving Machine Learning: Methods, Challenges and Directions. arXiv preprint arXiv:2108.04417.
15. *Hidano, S., Murakami, T., Katsumata, S., Kiyomoto, S., & Hanaoka, G.* (2017, August). Model inversion attacks for prediction systems: Without knowledge of non-sensitive attributes. In 2017 15th Annual Conference on Privacy, Security and Trust (PST) (pp. 115-11509). IEEE.
16. *Wang, Y., Si, C., & Wu, X.* (2015, June). Regression model fitting under differential privacy and model inversion attack. In Twenty-Fourth International Joint Conference on Artificial Intelligence.
17. *Wang, K. C., Fu, Y., Li, K., Khisti, A. J., Zemel, R., & Makhzani, A.* (2021, May). Variational Model Inversion Attacks. In Thirty-Fifth Conference on Neural Information Processing Systems.
18. *Alves, T. A., França, F. M., & Kundu, S.* (2019, May). MLPrivacyGuard: Defeating Confidence Information based Model Inversion Attacks on Machine Learning Systems. In Proceedings of the 2019 on Great Lakes Symposium on VLSI (pp. 411-415).
19. *W. Hickey.* FiveThirtyEight.com DataLab: How americans like their steak. http://fivethirtyeight.com/datalab/how-americans-like-their-steak/, May 2014.
20. *J. Prince.* Social science research on pornography. http://byuresearch.org/ssrp/downloads/GSShappiness.pdf.
21. *Deng, L.* (2012). The mnist database of handwritten digit images for machine learning research. IEEE Signal Processing Magazine, 29(6), 141–142.
22. *H. Xiao, K. Rasul, and R. Vollgraf.* Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. CoRR, abs/1708.07747, 2017.
23. *Yann Lecun, Leon Bottou, Y Bengio, and Patrick Haffner.* Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86:2278 – 2324, 12 1998.
24. *Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers.* Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax

diseases. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 2097–2106,2017.

25. *Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang.* Deep learning face attributes in the wild. In Proceedings of the IEEE international conference on computer vision, pages 3730–3738, 2015.
26. *Christer Loob, Pejman Rasti, Iiris Lusi, Julio CS Jacques, Xavier Baro, Sergio Escalera, Tomasz Sapinski, Dorota Kaminska, and Gholamreza Anbarjafari.* Dominant and complementary multi-emotional facial expression recognition using c-support vector classification. In 2017 12th IEEE International Conference on Automatic Face \& Gesture Recognition (FG 2017), pages 833–838. IEEE, 2017.
27. *Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton.* Cifar-10(canadian institute for advanced research).
28. *H.-W. Ng, S. Winkler.* A data-driven approach to cleaning large face datasets. Proc. IEEE International Conference on Image Processing (ICIP), Paris, France, Oct. 27-30, 2014.
29. *Bache, K. & Lichman, M.* (2013). UCI Machine Learning Repository (Technical report, University of California, Irvine, School of Information and Computer Sciences)
30. GroupLens Research, “MovieLens 1M Dataset,” http://grouplens.org/datasets/movielens/, 2003.
31. *Quinlan, J.R.* (1986) Induction of Decision Trees. Machine Learning, 1, 81-106. http://dx.doi.org/10.1007/BF00116251
32. *Shwartz-Ziv, R., & Tishby, N.* (2017). Opening the Black Box of Deep Neural Networks via Information. ArXiv, abs/1703.00810.
33. *Alemi, Alexander A., et al.* “Deep variational information bottleneck.” arXiv preprint arXiv:1612.00410 (2016).
34. *Vepakomma, P., Gupta, O., Dubey, A., & Raskar, R.* (2019). Reducing leakage in distributed deep learning for sensitive health data.
35. *Sharif Abuadbba, Kyuyeon Kim, Minki Kim, Chandra Thapa, Seyit A Camtepe, Yansong Gao, Hyoungshick Kim, and Surya Nepal.* Can we use split learning on 1d cnn models for privacy preserving training? arXiv preprint arXiv:2003.12365, 2020.
36. *Cynthia Dwork.* Differential privacy: A survey of results. In International conference on theory and applications of models of computation, pp. 1–19. Springer, 2008.
37. Fatemehsadat Mireshghallah, Mohammadkazem Taram, Prakash Ramrakhyani, Ali Jalali, Dean Tullsen, and Hadi Esmaeilzadeh. Shredder: Learning noise distributions to protect inference privacy. In Proceedings of the Twenty-Fifth International Conference on Architectural Support for Programming Languages and Operating Systems, pp. 3–18, 2020.
38. *Zhang, Jun & Zhang, Zhenjie & Xiao, Xiaokui & Yang, Yin & Winslett, Marianne.* (2012). Functional Mechanism: Regression Analysis under Differential Privacy. Proc. VLDB Endowment. 5.
39. *Namiot, D., Ilyushin, E., & Pilipenko, O.* (2022). On Trusted AI Platforms. International Journal of Open Information Technologies, 10(7), 119-127 (in Russian).
40. *Namiot, D., Ilyushin, E., & Chizhov, I.* (2022). On a formal verification of machine learning systems. International Journal of Open Information Technologies, 10(5), 30-34.

Junzhe Song. Student of the magistracy of the faculty of CMC of Lomonosov Moscow State University, MSU, Faculty of Computational Mathematics and Cybernetics, Russia, 119991, Moscow, GSP-1, 1-52, Leninskiye Gory, e-mail: songjz@smbu.edu.cn

D.E. Namiot. Dr. of Sci., senior researcher of the faculty of CMC of Lomonosov Moscow State University, MSU, Faculty of Computational Mathematics and Cybernetics, Russia, 119991, Moscow, GSP-1, 1-52, Leninskiye Gory, e-mail: dnamiot@gmail.com (correspondent author)

Research the Stability of Decision Trees Using Distances on Graphs

N.D. Moskin[I], K.A. Kulakov[I], A.A. Rogov[I], R.V. Abramov[II]

[I] Petrozavodsk State University, Petrozavodsk, Russia
[II] ITMO University, Saint Petersburg, Russia

Abstract. The article deals with the problem of stability of classifiers based on decision trees for the problem of text attribution. Such a task arises, for example, in the study of the authorship of articles from the pre-revolutionary journals “Time” (1861–1863), “Epoch” (1864–1865) and the weekly “Citizen” (1873–1874). The texts were divided into separate parts of different sizes using the sliding window method, then the frequency of n-grams (encoded sequences of parts of speech) in each fragment was determined. Further, these indicators were used to build various classifiers. The resulting decision trees were compared with each other using the tree edit distance. For this purpose, a procedure for processing, comparing and visualizing graphs was implemented in the SMALT software package. As a result of experiments using different weights for editing operations, patterns were revealed between the parameters for constructing text fragments and the decision trees obtained on their basis.

Keywords: *text attribution, n-gram, decision tree, graph matching, tree edit distance, software complex “SMALT”.*

DOI: 10.14357/20790279230111

Introduction

The main way to check the quality of a classifier is that it is checked against a control sample. If two classifiers showed similar results, it is not clear which one to choose. One option is to create a new control sample and conduct a new study. As a result, we get that the choice of a classifier significantly depends on the results on the control sample. When constructing and using a classifier in practice, much attention has to be paid to the quality (representativeness) of the training and control samples. It is very difficult. One of the existing ways out of this situation is to build an ensemble of classifiers. Sometimes it happens that each ensemble classifier is built on its own training sample. Its elements may contain fewer features than in the original data. The final classifier will be a function of the ensemble. Voting is often used as such a function. The element belongs to the class to which it belongs to most of the classifiers from the ensemble. Significant disadvantages of this method are the inability to justify the decision to classify and the complexity.

To eliminate these disadvantages, this paper proposes first splitting the ensemble of classifiers into clusters according to the degree of similarity, and then choosing the most typical (stable) classifier. Note that the training sample for this classifier will be representative. When training a classifier for a similar task, it will be possible to get by with one classifier, using a similarly constructed training set.

Experiments on the construction and comparison of classifiers were carried out on the material of the pre-revolutionary magazines “Time” (1861–1863), “Epoch” (1864–1865) and the weekly “Citizen” (1873–1874). It’s a known fact that F. M. Dostoevsky was their editor. It means that he could have made his own edits to the texts of articles of other authors [1]. The features that distinguish one author from another are the frequency of the occurrence of certain n-grams (encoded sequences of parts of the speech).

1. Decision tree constructing using different samples

In the NLP domain it is essential to have as much data as possible. In order to fulfill those pre-requirements in the circumstances of restricted in a size corpus a few techniques might be considered. Usage of a sliding window is one of those.

Sliding window breaks one text into several chunks which are used further as data (fig. 1). The main idea is to enrich training set through partially duplicating already existing corpus. The following logic might be applied:

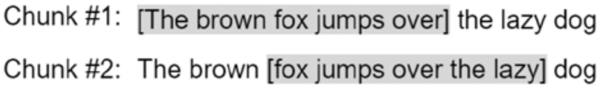
- N and S are chosen where N – sliding window size and S – skip window size.
- K chunks of the original text are created with a distance between start points of adjacent chunks equal to S.

You can calculate K (rounded to the nearest integer) using the following formula:

$$K = \frac{T-N}{S} + 1,$$

where T – size of a text that is being transformed. It is possible to vary both N and S. Basically they can be treated as hyperparameters. Varying N you can obtain more chunks with smaller size or less chunks with larger size.

Adjusting S governs amount of data as well as an intersection rate between two adjacent chunks. For example, a choice of N=1000 and S=100 will cause chunks #1 and #2 to have 900 words in common, chunks #1 and #3 – 800 words in common, etc. It is vital to choose this number high enough to save richness of the data although constructing a larger corpus.



Chunk #1: [The brown fox jumps over] the lazy dog

Chunk #2: The brown [fox jumps over the lazy] dog

Fig. 1. Parameters are N=5 and S=2. Sliding window results with 2 chunks.

The original training-test split was 80%–20%, thus 80% of data is used for training and the other 20% is used for testing algorithm and calculate metrics. Parts of speech are used as features and the purpose of a model is to predict an author of a given text using given features.

The described procedure was applied to a training part of a dataset to enlarge the existing corpus. After a decision tree [2] with default parameters from a Scikit-Learn [3] library is trained with built-in methods using the obtained training and test sets.

The default Decision Tree algorithm in Scikit-Learn is CART [4]. The model is represented as a binary tree with a certain criteria in every node. Prediction is made by going from the root of the tree to one of its leaf. The root is picked by meeting a condition that is represented in the node. A right path is picked for a true statement and a left one is for the false. For this task we use part of speech as data that represents an author, thus we have part of speech criteria in nodes. For example: a bigram “noun-noun” is met in the text more than 49 times?

The tree is built in a following manner:

1. Choose a node of the tree.
2. Calculate information gain for every feature and its threshold that splits data into two child nodes.
3. Choose a feature and its threshold for the node with a maximum purity.
4. Repeat until a stopping criteria is not reached.

The most popular impurity measure for CART algorithm is Gini index [5], which is defined as following:

$$I_{Gini} = 1 - \sum_{i=1}^{j} p_i^2,$$

where j is the total number of classes, p_i is the distribution of the i class in the node. Information gain is calculated as follows:

$$IG(D_p, f) = I(D_p) - \frac{N_{left}}{N_p} I(D_{left}) - \frac{N_{right}}{N_p} I(D_{right}),$$

where D_p, D_{left}, D_{right} – parent, left and right nodes respectively, N_i – number of data samples in node i.

2. Metrics on the set of trees

To compare trees, you can use the methods and algorithms developed within the graph matching direction [6; 7]. On a set of graphs a distance is set, which allows us to estimate how much these or other structures are “similar” to each other. One such distance is a measure based on editing operations (operations of insertion, deletion and replacement of vertices and edges in a graph) [8].

The function of error-correcting graph matching from $G_1=(V_1, E_1, \alpha_1, \beta_1)$ to $G_2=(V_2, E_2, \alpha_2, \beta_2)$ is called a bijective function $f: V_1' \rightarrow V_2'$, where $V_1' \subseteq V_1$ and $V_2' \subseteq V_2$. Cost of f define by the following formula:

$$c(f) = \sum_{u \in V_1'} c_{ns}(u) + \sum_{u \in V_1 - V_1'} c_{nd}(u) +$$

$$+ \sum_{u \in V_2 - V_2'} c_{ni}(u) + \sum_{e \in E_s} c_{es}(e) + \sum_{e \in E_d} c_{ed}(e) + \sum_{e \in E_i} c_{ei}(e),$$

where E_s, E_d and E_i are the sets of edges that are replaced, removed and inserted, respectively, and also:

- $c_{ns}(u)$ – the cost of replacing the vertex $u \in V_1'$ with $f(u) \in V_2'$;
- $c_{nd}(u)$ – the cost of removing a vertex $u \in V_1 - V_1'$ from G_1;
- $c_{ni}(u)$ – the cost of inserting a vertex $u \in V_2 - V_2'$ in G_2;
- $c_{es}(e)$ – the cost of replacing the edge e;
- $c_{ed}(e)$ – the cost of removing the edge e;
- $c_{ei}(e)$ – the cost of inserting the edge e.

At the same time, specific values c_{ns}, …, c_{ei} are selected depending on the specifics of the task being solved. Graph edit distance $d(G_1, G_2)$ is the cost of the optimal function f from G_1 to G_2:

$$d_1(G_1,G_2)=\min_f\left\{c(f):G_1\xrightarrow{f}G_2\right\}.$$

In other words, the graph edit distance $d_1(G_1, G_2)$ is the minimum total cost of editing operations that transform a graph G_1 into a graph G_2. Note that this distance can be used to compare arbitrary graphs, but there are measures designed specifically for trees. These include, for example, a distance based on tree alignment [9; 10]. To align two trees T_1 and T_2 (fig. 2), you need to insert empty vertices 0 into them so that the resulting trees T_1' and T_2' have the same structure (fig. 3).

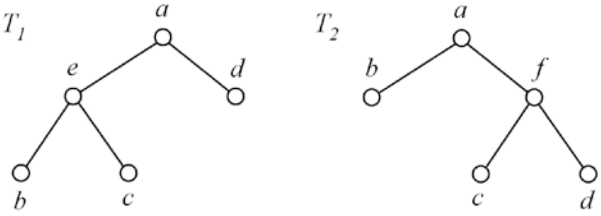


Fig. 2. Trees T_1 and T_2.

In this case, the labels of the corresponding vertices of the obtained trees T_1' and T_2' can be different. After that, we overlay trees T_1' and T_2' on each other and for each pair of vertices we calculate the measure of their difference μ_i. The sum $d_2(T_1,T_2)=\sum_{i=1}^{n}\mu_i$ will determine the distance between the trees if it is the minimum for all possible alignments of the given trees.

Next, consider four metrics for comparing trees, proposed in the work [11]. Let T_1 and T_2 be two trees with vertex set V_1 and V_2, respectively. A bijection $\phi: H_1 \to H_2$ where $H_1 \subseteq V_1$ and $H_2 \subseteq V_2$ is called an isomorphism of subtrees between T_1 and T_2 if ϕ preserves the adjacency relations between vertices and the connectivity of the compared subgraphs.

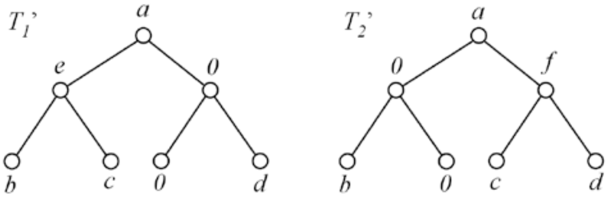


Fig. 3. Trees T'_1 and T'_2 resulting from alignment

Suppose that σ is some measure of similarity between the vertices of the compared trees T_1 and T_2, which, for example, can be based on the values of the attributes of these vertices. Then we define a measure of similarity $W_\sigma(\phi)$ between trees based on the isomorphism of subtrees ϕ and the measure σ:

$$W_\sigma(\phi)=\sum_{u\in H_1}\sigma(u,\phi(u)).$$

A subtree isomorphism ϕ_{12} is called a maximal subtree isomorphism between T_1 and T_2, if $W_\sigma(\phi)$ takes the greatest value among all possible subtree isomorphisms ϕ. For any two trees T_1 and T_2 define the following measures (here $|T|$ denotes the number of vertices in the tree T):

- $d_3(T_1,T_2)=\max(|T_1|,|T_2|)-W(\phi_{12})$;
- $d_4(T_1,T_2)=|T_1|+|T_2|-2W(\phi_{12})$;
- $d_5(T_1,T_2)=1-\dfrac{W(\phi_{12})}{\max(|T_1|,|T_2|)}$;
- $d_6(T_1,T_2)=1-\dfrac{W(\phi_{12})}{|T_1|+|T_2|-W(\phi_{12})}$.

The proof of the fulfillment of the metric properties for can be found in [12]. Another measure of dissimilarity on a set of graphs, based on the biotopic distance of sets, was proposed in [13]. In particular, the author notes that this measure can be used to compare decision trees.

3. Implementation of processing and comparison of decision trees in SMALT

As part of the testing of algorithms and presentation of research results, the implementation of processing and comparison of decision trees in the SMALT information system was carried out. The information system “Statistical methods of literary text analysis” (SMALT) is intended for processing texts in pre-revolutionary graphics and conducting statistical research on texts. The system is focused on processing texts from the magazines “Time”, “Epoch” and “Citizen” edited by F. M. Dostoevsky.

As part of the implementation, functions for working with the decision tree graph and functions for pairwise comparison of decision tree graphs were added. To implement the work with the decision tree graph, the “controller-model” pattern was used. The controller accepted the request, determined the required action, loaded the required model, and formatted the result. The following operations were implemented in the controller: loading a decision tree, editing decision tree metadata, viewing a decision tree, deleting a decision tree, viewing a list of loaded decision trees. Due to the large number of decision trees (for example, several decision trees with different steps can be obtained within one text comparison operation), it was customary to consider all downloaded decision trees private, i.e. not available to unregistered users. The user can add other people's private decision trees to the list of favorites or remove from it. Also, a user with explorer privileges can make decision trees public or remove publicity. These operations will allow you to limit the list of available trees in the pairwise comparison function to your own, selected or public graphs.

The model was responsible for interacting with the database. Decision tree graph data is organized into five tables: graph metadata, graph-related texts, graph vertices, graph edges, and a list of favorite graphs. One of the labor-intensive tasks was the implementation of parsing a text file with a decision tree graph and presenting the data in the database. The complexity of the task was due to the specific file format, the large amount of additional ignored data, and the wide variety of views. For example, the top of the graph should contain a list of parts of speech in the name and weight, but the format allowed for no restrictions on the name and no weight.

For graphical representation of a graph, the graphviz utility and the GraPHP library (https://github.com/graphp/graphviz) are used. As vertices, n-grams with combinations of parts of speech are used (for example, "Verb Pronoun"). If a graph vertex has a weight, then it is additionally displayed in the vertex label. If the weight of a vertex is greater than zero, then it is highlighted in red. By default, the image with the graph is displayed in a reduced version, but it can be opened in an additional window for a more detailed study.

Pairwise comparison of graphs is implemented based on the algorithm from [14]. The algorithm employs a dynamic programming approach and runs in polynomial time. The user can choose one of two comparison methods: based on vertex weights or based on vertex names and weights. To adjust the algorithms, the user can set the following parameters: Cost of adding a vertex, Cost of removing a vertex, Cost of changing a close vertex by weight, Level of proximity of vertices by weight, Cost of changing a far vertex by weight, Cost of changing a vertex with a partial match of parts of speech, Cost of changing a vertex with mismatch of parts of speech. For example, replacing a node (Adverb Verb, weight 0.405) with a node (Union Numeral, weight 0.012) with default settings (proximity level = 0.25, cost of replacing the far node by weight and with mismatch of parts of speech = 2, other costs = 1) will be equal to 4 (2 for the replacement of far vertices by weight + 2 for the replacement with mismatched parts of speech).

As a result of comparing graphs, the user gets the distance between the decision trees. The user can also view the final weight table.

4. Analysis of the regularities between the parameters of F. M. Dostoevsky's text sampling and decision trees

Let's study the regularities between the parameters of constructing a sample of texts (the size of the fragment, as well as the step of the sliding window) and the distances between the decision trees. The complete list of reference texts by F. M. Dostoevsky was taken as a sample [15].

Each article was divided into parts with a length of either x_1=1000 words or x_2=750 words. The limited number of such fragments forces us to resort to data expansion techniques. A sliding window was used as a similar technique: W={100, 200, 300, 400, 500, 750}. Then decision trees were built. Pairwise comparison of graphs was performed using a metric based on editing operations, the calculation of which was implemented based on the algorithm from [14]. Note that graph edit distance is currently one of the most popular similarity measures on a set of graphs [16].

Let's define the following weights of editing operations for G_1 and G_2:

– The cost of adding a vertex $u \in V_2 - V_2'$ to the graph G_2, which is denoted by $c_{ni}(u)$;

– The cost of removing a vertex $u \in V_1 - V_1'$ from the graph G_1, which is denoted by $c_{nd}(u)$;

– The operation of replacing two vertices $u \in V_1'$ with $f(u) \in V_2'$. Then the cost of replacement $c_{ns}(u)$ is the sum of the following components:

- cost of changing close vertex by weight $c_{ns}^1(u)$ (at the same time, the level of proximity of vertices by weight is set $c_{ns}^2(u)$). That is, if the modulus of the difference between the Gini indices is less than $c_{ns}^2(u)$, then the vertex is considered "close", and the cost of the operation is equal to $c_{ns}^1(u)$;
- cost of changing far vertex by weight $c_{ns}^3(u)$. That is, if the modulus of the difference between the Gini indices is greater than $c_{ns}^2(u)$, then the vertex is considered "far", and the cost of the operation is equal to $c_{ns}^3(u)$. Note that if the Gini indices are the same, then nothing is added to $c_{ns}(u)$;
- cost of replacing a vertex with a partial match of parts of speech $c_{ns}^4(u)$, i.e. in bigrams of two pairs of parts of speech, one coincides;
- the cost of replacing a vertex with a mismatch of parts of speech $c_{ns}^5(u)$, i.e. there are no common parts of speech in bigrams. Note that if the n-grams are the same, then nothing is added to $c_{ns}(u)$.

As an example, consider the calculation of the cost $c_{ns}(u)$ of a vertex replacement operation for three cases with the set parameters $c^1_{ns}(u) = 1$, $c^2_{ns}(u) = 0.25$, $c^3_{ns}(u) = 2$, $c^4_{ns}(u) = 1$, $c^5_{ns}(u) = 2$ (as a result, the total cost of the vertex replacement operation can be in the range from 0 to 4 inclusive):

– Let's compare the vertices with the bigrams "Adjective-Noun" (*gini*=0.348) and "Adjective-Noun" (*gini*=0.398), respectively. Since the n-grams are the same, and the difference in *gini* is 0.05, i.e. less than $c^2_{ns}(u) = 0.25$, then the replacement cost is $c_{ns}(u) = 0+1=1$.

– Let's compare the vertices with bigrams "Conjunction-Pronoun" (*gini*=0.262) and "Particle-Pronoun" (*gini*=0.358). Since n-grams coincide in one part of speech, and the difference in *gini* is 0.096, i.e. less than $c^2_{ns}(u) = 0.25$, then the replacement cost is $c_{ns}(u) = 1+1=2$.

– Let's compare the vertices with bigrams "Adjective-Participle" (*gini*=0.032) and "Modal word-Adverb" (*gini*=0.444). Since the n-grams do not completely coincide, and the difference in *gini* is 0.412, i.e. more than $c^2_{ns}(u) = 0.25$, then the replacement cost is $c_{ns}(u) = 2+2=4$.

Table 1 shows an example of a distance matrix obtained with the set parameters $c_{ni}(u) = 1$, $c_{nd}(u) = 1$, $c^1_{ns}(u) = 1, c^2_{ns}(u) = 0.25$, $c^3_{ns}(u) = 2, c^4_{ns}(u) = 1, c^5_{ns}(u) = 2$, where the depth of the tree was not limited. The fragment size was chosen to be 750 or 1000, and the size of the sliding window was 100, 200, 300, 400, 500, and 750 words. The result of cluster analysis is shown in fig. 4.

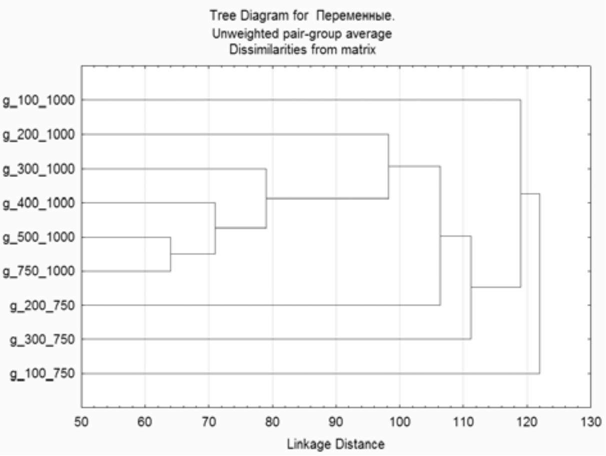


Fig. 4. An example of a dendrogram for a distance matrix from table 1

To represent a set of close graphs as one that would contain basic information about all structures,

Table 1.

An example of a distance matrix between decision trees.

F		1000 words						750 words		
	W	100	200	300	400	500	750	100	200	300
1000 words	100	0	129	118	116	109	102	138	128	131
	200	129	0	107	99	92	95	129	114	119
	300	118	107	0	85	78	74	119	112	112
	400	116	99	85	0	72	70	113	106	106
	500	109	92	78	72	0	64	109	100	103
	750	102	95	74	70	64	0	108	100	103
750 words	100	138	129	119	113	109	108	0	129	131
	200	128	114	112	106	100	100	129	0	124
	300	131	119	112	106	103	103	131	124	0

Table 2.

Median graphs for nine experiments

Experiment	$c_{ni}(u)$	$c_{nd}(u)$	$c^1_{ns}(u)$	$c^2_{ns}(u)$	$c^3_{ns}(u)$	$c^4_{ns}(u)$	$c^5_{ns}(u)$	$\min_{G\in Z} D$	$\hat{G}$
the depth of the decision tree was not limited									
1	1	1	1	0,25	2	1	2	716	g_750_1000
2	2	2	1	0,25	2	1	2	1106	g_500_1000
3	1	1	1	0,25	2	2	4	716	g_750_1000
decision tree depth limited to 5									
4	1	1	1	0,25	2	1	2	565	g_750_1000
5	2	2	1	0,25	2	1	2	849	g_500_1000
6	1	1	1	0,25	2	2	4	568	g_750_1000
decision tree depth limited to 4									
7	1	1	1	0,25	2	1	2	357	g_500_1000
8	2	2	1	0,25	2	1	2	459	g_500_1000
9	1	1	1	0,25	2	2	4	357	g_500_1000

consider the concept of a median graph [17]. Let the distance $d(G_i, G_j)$ be given on the set $Z=\{G_1, G_2, \ldots, G_m\}$. Then we call the median graph on the set Z:

$$\hat{G} = \arg\min_{G \in Z} D = \arg\min_{G \in Z} \sum_{i=1}^{m} d(G, G_i),$$

where $D = \sum_{i=1}^{m} d(G, G_i)$ – the sum of the distances from the selected graph G to all other graphs from the set Z. Note that the median graph search problem is NP-complete. Table 2 describes the parameters for conducting nine experiments to calculate the median graph (experiments with other parameters showed comparable results). As the calculations showed, *g*_750_1000 became such a graph four times (i.e. x_1=1000, y_6=750), in the remaining five cases – *g*_500_1000 (i.e. x_1=1000, y_5=500).

Conclusion

In this article on the material of the pre-revolutionary magazines "Time" (1861–1863), "Epoch" (1864–1865) and the weekly "Citizen" (1873–1874), the problem of the stability of classifiers, which are built to determine the authorship of texts, is considered. The features were the frequency of occurrence of certain n-grams (encoded sequences of parts of speech). The decision trees obtained as a result of applying the sliding window method were compared with each other using the tree edit distance (it is currently one of the most popular similarity measures on a set of graphs). To analyze the results obtained in the SMALT information system (http://smalt.karelia.ru/), tools for storage, processing and comparing trees were implemented. For graphical representation of graphs the graphviz utility and the GraPHP library were used. The most stable decision trees (median graphs) were identified for several text collections using different weights for editing operations.

References

1. *Abramov R. V., Kulakov K. A., Lebedev A. A., Moskin N. D., Rogov A. A.* 2021. Research of features of Dostoevsky's publicistic style by using n-grams based on the materials of the "Time" and "Epoch" magazines. Vestnik of Saint Petersburg University. Applied Mathematics. Computer Science. Control Processes. Saint Petersburg, 17(4): 389–396.
2. *Safavian S. R., Landgrebe D.* 1991. A survey of decision tree classifier methodology. IEEE transactions on systems, man, and cybernetics, 21(3): 660–674.
3. *Pedregosa F., et al.* 2011. Scikit-learn: Machine learning in python. The Journal of Machine Learning Research, 12: 2825–2830.
4. *Lewis R. J.* 2000. An introduction to classification and regression tree (CART) analysis. Annual meeting of the society for academic emergency medicine in San Francisco, California. Citeseer 14.
5. *Coppersmith D., Hong S. J., Hosking J. R. M.* 1999. Partitioning nominal attributes in decision trees. Data Mining and Knowledge Discovery, 3(2): 197–217.
6. *Conte D., Foggia P., Sansone C., Vento M.* 2004. Thirty years of graph matching in pattern recognition. International Journal of Pattern Recognition and Artificial Intelligence, 18(3): 265–298.
7. *Jiang X., Bunke H.* 2008. Graph matching. Case-Based Reasoning on Images and Signals. Vol. 73 of Studies in Computational Intelligence. Springer, 149–173.
8. *Riesen K.* 2015. Structural Pattern Recognition with Graph Edit Distance: Approximation Algorithms and Applications. Advances in Computer Vision and Pattern Recognition. Springer, Heidelberg. 158 p.
9. *Bille P.* 2003. Tree edit distance, alignment distance and inclusion. IT University of Copenhagen. Technical Report Series, TR-2003-23.
10. *Jiang T., Wang L., Zhang K.* 1995. Alignment of trees – an alternative to tree edit. Theoretical Computer Science, 143(1): 137–148.
11. *Torsello A., Hidovic D., Pelillo M.* 2004. Four metrics for efficiently comparing attributed trees. Proc. ICPR'04 – 17th International Conference on Pattern Recognition, 2: 467–470.
12. *Torsello A., Hidovic D., Pelillo M.* 2005. Polynomial time metrics for attributed trees. IEEE Transactions on Pattern Analysis and Machine Intelligence, 27(7): 1087–1099.
13. *Kuznetsov A. V.* 2017. Mera neskhodstva na mnozhestve grafov i ee prilozheniya [A measure of dissimilarity on a set of graphs and its applications]. Vestnik Voronezhskogo gosudarstvennogo universiteta. Seriya: sistemnyj analiz i informacionnye tekhnologii [Bulletin of the Voronezh State University. Series: system analysis and information technology], 1: 125–131.
14. *Isert C.* 1999. The editing distance between trees. Ferienakademie Bäume: Algorithmik und Kombinatorik. Sarntal, Italy.
15. *Rogov A. A., Abramov R. V., Buchneva D. D., Zakharova O. V., Kulakov K. A., Lebedev A. A., Moskin N. D., Otlivanchik A. V., Savinov E. D., Sidorov Y. V.* 2021. Problema atribucii v zhurnalah "Vremya", "Epoha" i ezhenedel'nike "Grazhdanin" [The problem of attribution in the maga-

zines "Time", "Epoch" and the weekly "Citizen"]. Petrozavodsk: Islands, 391 p.

16. *Wu L., Chen Y., Shen K., Guo X., Gao H., Li S., Pei J., Long B.* 2021. Graph Neural Networks for Natural Language Processing: A Survey. ArXiv abs/2106.06090.

17. *Hlaoui A., Wang S.* 2003. A New Median Graph Algorithm. Graph Based Representations in Pattern Recognition (GbRPR 2003). Lecture Notes in Computer Science. Springer, Berlin, Heidelberg, 2726: 225–234.

Moskin N.D. PhD in Technics, Associate Professor, Petrozavodsk State University, 33 Lenin str., Petrozavodsk, 185910, Russia. E-mail: moskin@petrsu.ru (correspondent author)

Kulakov K.A. PhD in Physics and Mathematics, Associate Professor, Petrozavodsk State University, 33 Lenin str., Petrozavodsk, 185910, Russia. E-mail: kulakov@cs.karelia.ru

Rogov A.A. Dr. Sci. in Technics, Professor, Petrozavodsk State University, 33 Lenin str., Petrozavodsk, 185910, Russia. E-mail: rogov@petrsu.ru

Abramov R.V. ITMO University, 49 Kronverksky Pr., bldg. A, Saint Petersburg, 197101, Russia. E-mail: monset008@gmail.com

Economic Cycle Prediction using Machine Learning – Russia Case Study*

B.K. Amos, I.V. Smirnov, I.A. Aidrous, R.R. Asmyatullin, S.G. Glavina

Peoples Friendship University of Russia (RUDN Universiry), Moscow, Russia

Abstract. The long-term development of the world economy is characterized by cyclical development. To date, there is no single accepted approach to describe the nature of the economic cycle. Therefore, studies of economic and political cycles are one of the key areas of economic theory. Econometrics and machine learning have a common goal: to build a predictive model, for a target variable, using explanatory variables. This research aims to identify economic cycle in Russian Federation using collective factors. It uses a different approach, concerning classical econometric techniques, and shows how machine learning (ML) techniques can improve the accuracy of forecasts. We used three machine learning algorithms such as k-Nearest Neighbors (kNN), Random Forests (RF) and Support vector machines (SVM). The research is based on 30 economic factors for the period 1990-2020 from FRED, World Bank, WTO, Federal State Statistics Service, Bank of Russia etc. The results indicate that the Russian economy would be very active (peak) in the next quarters. This result could be a new approach to provide policy recommendations to authorities and financial institutions in particular.

Keywords: *macroeconomics, Machine Learning, Econometric Forecasting, Russian economy, Economic cycle.*

DOI: 10.14357/20790279230112

Introduction

One of the most important characteristics of a developed market economy is its cyclical development - periodic ups and downs in business activity. Economic cyclicality can be traced in the change in the main macroeconomic indicators – the volume of real GDP, the intensity of investment, the unemployment rate, etc.

The cyclical nature was influenced by the deepening of the integration and liberalization processes; strengthening the international division of labor; mass

Application and rapid development of information and communication technologies, software, nanotechnology, composite materials. Digitalization of the economy and covid-19 crisis in the last few years have played an important role in the emergence of many other services and activities: e-commerce, remote exchange trading, mobile banking and others creating tons of information to manage hence big data.

The current forecasting literature has focused on matching specific variables and horizons with a particularly successful algorithm. The intersection of machine learning (ml) with econometrics has become an important research land- scape in economics. Machine learning has gained prominence due to the availability of large data sets, especially in microeconomic applications (coulombe et el., 2020). De- spite the growing interest in ml, understanding the properties of ml procedures when they are applied to predict macroeconomic outcomes remains a difficult challenge.

There are many techniques and methods to analyze and process this information. Machine learning lies at the intersection of mathematics and statistics, computer science and big data. Simply put, machine learning is a way to put big data analysis into service. With the help of such technology, the computer can learn to identify certain patterns, to "meet", with which it will perform certain actions - buying or selling stocks, segmenting potentially high-yielding customers, or identifying faulty products on the conveyor belt.

Econometrics and machine learning share a common goal, namely the construction of a predictive model, for a variable of interest, using explanatory variables (or features). However, these two approaches have developed in parallel, creating two different cultures. The first aims to build probabilistic models to de-

* This paper has been supported by the RUDN University Strategic Academic Leadership.

scribe economic phenomena. The second uses algorithms that learn from their errors, for the purpose of prediction or classification. Recently, learning models have proven to be more efficient than traditional econometric techniques and are able to handle much larger data sets. In this context, it becomes necessary for econometricians to understand what these two cultures are, what opposes them and especially what brings them together, in order to appropriate tools developed by the statistical learning community to integrate them into econometric models.

1. Related work

Cyclic booms and busts have been a feature of economic life and one of the major arguments against capitalism since the eighteenth century. (Ayres, R. 2020). The cyclical nature of the Russian economy, its susceptibility to alternating booms and booms, is still poorly understood, although the recent crisis has generated a flurry of publications covering various aspects of the crisis in Russia. These publications, however, are largely journalistic in nature and do not yet allow to form a holistic idea of the nature of the crisis unfolding in the world economy, its driving forces and possible consequences, as well as the mechanism for overcoming the crisis and the degree of influence of these processes on the Russian economy.

The theoretical influence of economic cycles on time-varying risk premiums is then explained based on two key economic concepts: nominal GDP and adaptive expectations. (Raffinot, T. (2017) The economic cycle is thus a fundamental yet ambiguous concept, since it can refer to conceptually different global economy fluctuations.

On the functioning of the economy, many theories exist on the variables that need to be analyzed in order to keep a quick overview of economic trends. Ensemble machine learning algorithms, referred to as random forest (Breiman (2001)), are applied to quickly and accurately detect economic turning points in the United States and in the euro area. Genetic programming (GP) has been successfully used as a machine learning tool for automatic problem- solving in many areas such as image processing (Harding et al., 2013), but very seldom for macroeconomic modelling and forecasting (Alvarez-Diaz, 2019, Claveria et al., 2019, 2020). This article uses an extended set of indicators for the forecast: GDP in constant 2010 US $, population in absolute values are added to macroeconomic parameters, data are added not only on unemployment, but also on labor force, as well as manufacturing production. Exports and imports of goods are added to the indicators of the international economy, the analysis includes not only foreign direct investment, but also portfolio investment, current account balance. More financial indicators are used, related to the exchange rate, stock markets and cross-border transactions. This study does not consider factors that are stable over the analyzed period (such as the area of the country, forest areas), generalized indicators for the general situation in certain sectors of the economy, as well as evaluative indicators (such as the political and social atmosphere) unlike the study conducted by Chukiat et al, (2019) which uses qualitative information, quantitative trends and social movement activities. However, the Random Forest (RF) algorithm showed high accuracy in predicting GDP with an accuracy of 0.70 and a Kappa coefficient of 0.41 in the study by Chukiat et al.(2019) and in ours an accuracy of 0.78 and a Kappa coefficient of 0.49.

2. The methods of the research

The objective of this paper is to computationally predict economic structural trends in Russia by applying machine learning. The data used consists of mixed observations such as qualitative survey details and time trend data series from 1990 to 2020, that are used to perform econometric estimation by artificial intelligent approaches. All variables are described and presented in Table 1. Technically, the collective variables used in this paper were collected from reliable sources that have managed to store information from the world trends for convenient access, for example, FRED (Federal Reserve Economic Data), Federal Budget of the Russian Federation, World Bank Database, World Trade Organization, Bank of Russia, etc.

The modern economic theory is based on the idea that the cycle is primarily about output and employment fluctuations. The trend sets the level of potential output corresponding to the production capacity of the economy at full employment. The growth in potential output is a consequence of the growth of production opportunities and is described by the models of economic growth. Growth is associated with the accumulation of factors of production (capital accumulation, growth in the labor force, productivity growth). Growth rates depend on long-term trends; therefore, a significant change in potential output occurs only in the long run, while in the short run it is considered constant. Manufacturing capabilities affect aggregate supply.

The cyclical component describes short-term fluctuations in the observed GDP around a trend. The economy grows faster or slower than the trend, de- pending on how intensively the resources available in it are used. With excessively intensive use, it grows faster than the trend, with insufficient intensive use it grows slower.

In the following research we selected the following group of factors:

– National Currency

With an increase in the exchange rate, on the contrary, exports become unprofitable, which can lead to a reduction in export industries and national production in general, and the volume of imports increases. The import of foreign investments is growing. The real amount of external debt, expressed in depreciated foreign currency, is decreasing. The devaluation of the ruble taking place today is an ambiguous process. In the short term, of course, this affects the population and its income. But, on the other hand, this creates more favorable conditions for Russian manufacturers, because their products are becoming more competitive in the domestic and international markets. Traditionally, it is believed that the weakening of the ruble is beneficial to exporting companies. But the benefits from the weakening ruble for oil companies as exporters of raw materials does not outweigh their losses because of the growth of the corresponding tax burden tied to the national currency exchange rate.

– State finance

This group includes expenses, income, external debt, account balance and military expenses, etc. Russia is characterized by a high level of state participation in the economy. Large companies, with state participation dominate on the national market. Government funding and their effectiveness directly determine economic growth and diversification. With a significant amount of public external debt, most of the federal budget funds are spent on reducing the deficit in the consolidated budget, therefore, there is a decrease in expenditures for the development and expansion of production within the country, for social needs, as a result, everything affects the living standards of people. Thus, there is a slowdown in the country's economic growth.

– Stock Market

This group includes share market, stock market indices, number of registered companies, etc. The stock index is the main indicator of the stock market, which is based on the prices of a group of securities, reflects the state and dynamics of the securities market. Stock markets around the world are interconnected through communication channels, and information can spread to investors very quickly. The change in the index over time allows us to judge the general di- rection of price movement. Pessimism in the stock market causes a decline in quotations, and optimism or a low level of pessimism in the market contributes to high trading volumes in the stock market and higher incomes. Pessimism and optimism equally affect the stock market in accordance with the theory of investor sentiment and the theory of timing for an IPO (market timing theory).

– Central Bank policy

Includes interest rate, Central Bank assets, etc. The interest rate is one of the most effective instruments for national economy regulation. The rate reduction will stimulate the economy. Borrowing becomes cheaper, in the beginning for banks, then for corporations and then for consumers. The demand for goods and services is growing. However, the global economy develops in cycles. And at a certain stage of economic recovery, the market overheats. In order to avoid landslide falls, regulatory authorities are taking measures to smooth out cyclical fluctuations by slowing down of economic processes by raising the interest rate.

– Macroeconomic factors

This group includes GDP, Industrial production index, inflation, international trade, etc. Essentially, GDP reflects the health of an economy, which can directly affect investor sentiment. You should also use the values of real GDP, which considers the extent to which GDP growth is determined by real growth in production, and not by price increases. The Industrial Production Index is an indicator of business cycles that affect price fluctuations in the stock market. The growth of the industrial production index reflects the development of existing sectors of the economy, which may cause an increase in the number of IPOs, and, on the other hand, may indicate the development of knowledge-intensive industries, whose innovative companies enter the capital markets in search of investments for development. Thus, they contribute to the creation of a higher level of industrial production. As for structural growth, it has been insignificant in the past seven years due to the orientation of the economy towards the export of raw materials while the share of innovative sectors has decreased. Due to the rigidity of prices, the general equilibrium in the economy in a short period is ensured by changes in quantity. When demand rises, firms increase production and hire additional workers, and when demand falls, they reduce production and lay off workers. Therefore, the observed GDP is greater or less than the potential GDP, respectively. In response to high inflation, it becomes unprofitable to make savings, open deposits, so bank clients withdraw money from their accounts. Russia is a major participant in the system of international trade relations. The development of international trade in the XX century. turned it into a decisive factor in economic growth for most countries in the world.

– Population

Rapid aging of the population is forcing the world to rethink fiscal, social and migration policies. This question is especially relevant for Russia, where, due to the "demographic trap" of the 1990s, the birth rate is declining faster than in developed countries. In Russia, the echoes of the Second World War remain in the demographic pyramid. As well as the consequences of the demographic slump of the 90s. Labor force affects labor productivity and the scale of production. As long as the labor market can absorb labor, labor productivity will rise. This creates the so-called "demographic dividend" of economic growth, which contributes to an increase in savings, savings and investment.

– Investments

The stock market indicators' outstripping economic fluctuations is explained by the desire of investors to predict the direction of economic development in order to sell assets at their peak or buy them at the lowest price. That is why the cyclical development of the stock market is ahead of the development of the economic cycle.

Table 1

The different variables and their sources

Variables	**Range of time series**	**Sources**
Real Effective exchange rate	1990-2008	Federal Reserve Economic Data
Government Expenditures	1990-2008	Federal Budget of the Russian Federation
Government Revenues	1990-2008	Federal Budget of the Russian Federation
Public debt (% per GDP)	1990-2008	Trading Economics
Military expenditures	1990-2008	Trading Economics
External debts stocks total (DOD US $) Russian Federation	1990-2008	The World Bank
Use of IMF credit (DOD current US$) Russian Federation	1990-2008	The World Bank
Stock Market Capitalization to GDP for Russian Federation	1990-2008	Federal Reserve Economic Data
Long-Term Government Bond Yields: 10-year: Main for the Russian Federation	1990-2008	Federal Reserve Economic Data
3-Month or 90-day Rates and Yields: Inter-bank Rates for the Russian Federation	1990-2008	Federal Reserve Economic Data
Interest Rates. Government Securities. Treasury Bills for Russian Federation	1990-2008	Federal Reserve Economic Data
Central Bank Assets to GDP for Russian Federation	1990-2008	Federal Reserve Economic Data
Total Share Prices for All Shares for the Russian Federation	1990-2008	Federal Reserve Economic Data
National Currency to US Dollar Exchange Rate: Average of Daily Rates for the Russian Federation	1990-2008	Federal Reserve Economic Data
Stock Market Total Value Traded to GDP for Russian Federation	1990-2008	Federal Reserve Economic Data
Number of Listed Companies for RF, Number of Listed Companies per Million People, Annual, Not Seasonally Adjusted	1990-2008	Federal Reserve Economic Data
Total Share Prices for All Shares for the Russian Federation, Index 2015=100, Annual, Not Seasonally Adjusted	1990-2008	Federal Reserve Economic Data
External debt stocks, long-term (DOD, current US$)	1990-2008	Federal Reserve Economic Data
GDP (constant 2010 US$)	1990-2008	The World Bank
GDP Growth annual %.	1990-2008	The World Bank
Inflation, %, consumer prices (annual %)	1990-2008	The World Bank
Population	1990-2008	United Nations
Population (growth rate)	1990-2008	Federal State Statistics Service
Labour force, total	1990-2008	The World Bank
Unemployment	1990-2008	The World Bank
Industrial value added. Industry, value added (% of GDP)	1990-2008	The World Bank
Agricultural value added. Agriculture, forestry, and fishing, value added	1990-2008	The World Bank
Total Manufacturing Production for the Russian Federation	1990-2008	The World Bank
Total merchandise exports (Russia), annual	1990-2008	World Trade Organization
Total merchandise imports (Russia), annual	1990-2008	World Trade Organization
Commercial services exports (Russia),	1990-2008	World Trade Organization
Commercial services imports (Russia), (Million US dollar)	1990-2008	World Trade Organization
Foreign direct investment, net inflows	1990-2008	The World Bank
Portfolio Investment, net (BoP, current US$)	1990-2008	The World Bank
Current account balance (BoP, current US$)	1990-2008	The World Bank

Total reserves (includes gold, current US$)	1990-2008	The World Bank
Official exchange rate	1990-2008	The World Bank
External debt stocks	1990-2008	The World Bank
Net international investment position	1990-2008	The World Bank
Cross-border transactions of individuals,Transfers from Russia, million USD	1990-2008	Bank Of Russia
Cross-border transactions of individuals,Receipts into Russia, mln USD	1990-2008	Bank Of Russia
Transfers from Russia made through money transfer systems, mln USD	1990-2008	Bank Of Russia
Transfers to Russia made through money transfer systems, mln USD	1990-2008	Bank Of Russia
Volume of money transfers, RUB billion (for the period)	1990-2008	Bank Of Russia
Transaction volume, RUB billion	1990-2008	Bank Of Russia
Volume of payments by credit institutions, RUB billion	1990-2008	Bank Of Russia

3. Forecasting methods and tools

3.1. Optimization of hyper-parameters

Before discussing the forecasting models, it is important to detail how the hyper- parameters are selected. Let us take the example of the auto-regressive model AR

$$y_{th} = \rho(L)Y_t + e_t \tag{1}$$

where the order $\rho(L)$ is the only hyper-parameter. To fix $\rho(L)$, the standard approach is the selection criterion. The Bayesian Information Criterion (BIC) is used in this case:

$$\boldsymbol{log}\left(\frac{SCR_{pj}}{T}\right) + p_i.\boldsymbol{log}\left(\frac{T}{T}\right) \tag{2}$$

Where scr_{pj} is the sum of squares of the residuals and where pj denotes the choice of auto-regressive order. Since the first term is decreasing in pj , the second term allows the BIC to regularize the over-fitting by penalizing with the number of parameters to be estimated.

Another way of optimizing parameters that is especially popular in machine learning is cross-validation (CV). Like BIC, CV also selects the optimal order p, but regularizes using out-of-sample prediction performance, whereas BIC selection is based solely on in-sample performance. The popularity of CV is also due to its simplicity, since it can be practiced even when the information criterion is not available. There are several approaches to CV, but the most popular is based on random resampling (K -fold) in the training period. Assume that the number of folds (fold) is fixed at five (another second-order hyper-parameter). This is equivalent to dividing the sample period into five equally sized sub-samples. Then, four sub-samples are used in turn to estimate model j of the previously presented grid (forming the training set in ML language), and one sub-sample is used to evaluate the out-of-sample performance with, usually, the mean square error (MSE) as metric. The element j of the grid producing the minimum MSE will be the optimal order estimate. In this paper we used cross-validation.

3.2. The proposed forecasting models

In this section of the article different models of the learning machine are used. These models are: Random Drills (RF), K-Nearest neighbors (KNN), Support Vector Machine (SVM).

- **Random Forests (RF)**

A random forest is machine learning algorithm that is made up of multiple decision trees to predict a result, and this collection of trees is often called an ensemble. It's powerful tool that is used in many industries to help companies make better decisions, reduce risk and maximize success. It's a very popular machine learning algorithm that can be used for classification and regression. The first model using a non-linear approximation is the random forest model (de Breiman, 2001).

A random forest consists of many decision trees, and every tree is built using a four-step process.

Step 1. Create a 'Random' Dataset : Bootstrapping
Step 2. Select 'Random' Attributes
Step 3. Select best attribute to split Step 4. Split the attribute
Step 2-4 are repeated until a tree is fully constructed.

Technically, the Random Forest algorithm is often shown as follows:

For b = 1 to B

1. Create a bootstrap sample, with replacement, B training examples from x, y. Label these x_b, y_b
2. Train the tree, f_b on x_b, y_b
3. Average the predictions or take the majority vote to arrive a final prediction.

'b' represents a single tree

'B' represents the entire forest

Like decision tree, each tree in random forest is no different, except that is built from a 100 % unique training dataset.

This is accomplished by selecting random examples from the original training dataset and recreating the dataset for each tree. Technically, this is called bootstrapping, which is a statistical technique that makes use of sampling. A major benefit of random sampling is that every piece of training data is likely to be included, or represented, in at least one tree in the forest.

After a random forest has created a new dataset, it randomly selects at- tributes to split. To do this, algorithm does two things:

First, it restricts how many attributes it will use for any given branch. To do this, it uses the square root of the number of attributes as the maximum of attributes it will consider.

Second, it randomly selects the number of attributes.

Next, the algorithm selects the best attribute to split. To do this, it makes use of entropy and information gain to determine the 'purity' of each split. The split with the greatest level of purity is selected.

To make a prediction, there are two ways that a random forest predicts: majority vote or mean.

Predicting with a majority vote - One option for making a prediction is to simply take a majority vote, which means that the class that is predicted most often within the forest is selected as the algorithm's final prediction. This approach works very well for categorical classes, such as 'Yes' and 'No' or 'Dog' and 'Cat'

Predicting with the mean - Second option for predicting is to use the mean, or average, of the result of all the trees in the forest. This is calculated by using the following formula:

$$\hat{y} = \frac{1}{T}\sum_{b=1}^{B} y_b(x') \tag{3}$$

ŷ: is the final prediction for the random forest.

y_b: represents a single tree.

T: represents the total number of trees in the random forest.

x′: represents the prediction of each class for a single tree

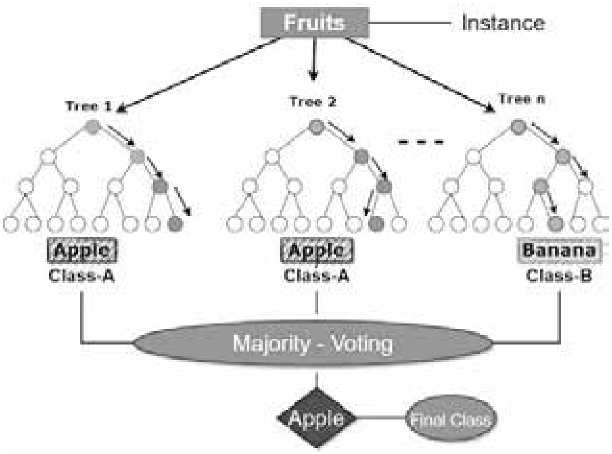


Fig. 1. Decision tree with depth of 4.

- **K-Nearest neighbors (KNN)**

The K-Nearest Neighbors (KNN) classifier is one of the most widely used classification algorithms in machine learning that belongs to the supervised learning category. This algorithm can be used in regression problems. It records all valid attributes and classifies new attributes based on their similarity dimension. KNN is a statistical recognition model method for detecting different classes in a model. A tree data structure is used to determine the distance between the point of interest and the points in the training data set.

The attribute is classified by its neighbors. In the classification method, the value of k is always a positive integer closest to the neighbor. The nearest neighbors are selected from a set of classes or property values of the object.

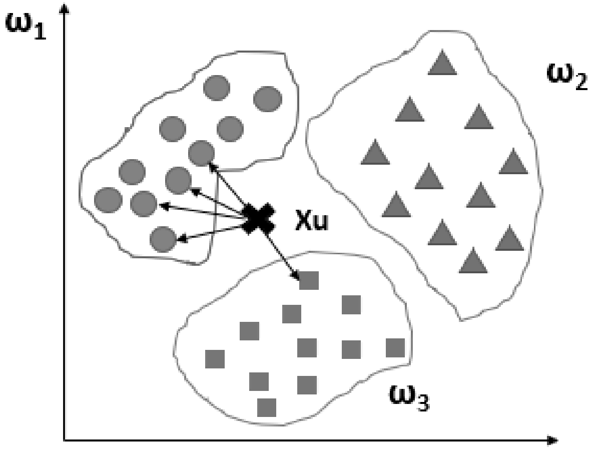


Fig. 2. Assignment of observations to clusters of classes w1, w2 and w3

- **Support Vector Machine (SVM)**

The support vector machine algorithm is used to solve classification problems and regression [3]. This algorithm is a relatively new approach and has shown good performance in recent years. The support vector machine algorithm is based on linear classifiers and in line-separated data, this algorithm isolates objects into specified classes [4]. It can also identify and classify instances that are not supported by the data. The only extension of this algorithm is to perform a regression analysis to obtain a linear function, and another extension teaches how to classify items to obtain a classification of individual items.

3.3. Model validation

As in econometrics, it is difficult to give better results in machine learning. For the validation of the model in this article we will use Cohen's Kappa coefficient (k). [5] Cohen's Kappa coefficient is a statistical tool that measures the agreement between two raters, determining to which category a finite number of items belong, and represents the degree of accuracy and reliability in a statistical classification. The value of Kappa can be negative, i.e., less than 0. A score of k=0 means that there is random agreement between the raters, while a score of k=1 means that there is complete agreement between the raters. Therefore, a score of k less than 0 means that there is less agreement than random. Cohen's co-efficient Kappa remains calculated by this formula:

$$k = \frac{p_0 - p_e}{1 - p_e} \quad (4)$$

Where p_0 is the observed relative agreement between raters, which identifies accuracy, and is the hypothetical probability of strong agreement. The observed data are used to calculate the probabilities of each randomly observable view in each category [5]. For categories k, a number of elements n and is the number of time evaluators n_{ki} is the number of time evaluators i predicted category k,

$$p_e = \frac{1}{N^2} \sum_k n_{k1} \cdot n_{k2} \quad (5)$$

4. Empirical results

4.1. Descriptive information

The figure below shows Russia's economic trends through its annual collective GDP from 1990 to 2008 between actual and forecast. The ranking of the reality trends clearly shows that the Russian economy fluctuates considerably. This fluctuation makes prediction more difficult, as traditional econometric tools cannot provide the best model, so Newton's method was used to extend the explanatory power of the data. The empirical results show that the peak period is defined at the level, which is above the optimal value (5%). The graph below shows the Russian GDP from 1990 to 2008 between actual and forecast values.

4.2. The results of the machine learning algorithm

This step is the comparison of different machine learning models. For this com- parison we used cross test validation. Three machine learning algorithms were used namely K-NN, Random Forest and SVM. The data set will be divided into two parts. 75% of the data is used to train the models, and 25% of the data is used to perform tests.

The results are presented in Table2. In this calculation Random Forest (rf) is the best model and contains the highest parametric values when selecting by accuracy values and Kappa coefficient, which are 0.78 and 0.49 respectively.

The Random Forest algorithm has the highest score among the machine learning algorithms used for the prediction of macroeconomic variables in Russia. Moreover, the mean absolute error (MAE) of this algorithm is 0.012595 and the root mean square error (RMSE) is 0.017032.

Predictions show the following peaks – 1998 and 2008. We can give a proper economic interpretation. The recovery of the economy after reforms began in 1996-1997. Price liberalization, capital movements and large-scale privatization of enterprises at undervalued prices led the economy to default in 1998- a deep economic recession, the depreciation of the ruble, and a drop in income. The devaluation and freed-up capacity gave the economy a boost, growing 6.4% in 1999 and 10% in 2000. From 1999 to 2008, the Russian economy showed GDP growth at an average of 7% per year. By 2008, Russia's GDP had almost doubled, the poverty rate had halved, foreign direct investment had risen from $14.3 billion in 2001 to $121.1 billion in 2007. As a result, Russia continued to attract the attention of investors.

Table 2

Summary table of results

Models	Hyper-parameters	Accuracy	Kappa's Coefficient

Random Forest (rf)	Max-depth = 3
	N-estimators = 100
	Criterion = gini
	Max-features = 0.0693
		0.7810	0.4999

SVM (svm)	C = 246.4819
	Gamma = scale
	Class-weight = balanced
		0.7190	0.5

KNN (knn)	N-neighbors = 14, p = 3	0.5571	0.125

Conclusion

We use several economic variables in order to make a prediction. Machine learning systems can handle a huge number of informative details in databases, including qualitative data, quantitative factors and even time series trends. In this paper, 30 time series variables concerning the economic structures of Russia from 1990 to 2008 were included to predict the reasonable future trend. due to relative economic stability, to exclude external influence on the results. The obtained results show that the Random Forest is the best selected model. As a conclusion, we can say that machine learning is the appropriate solution for money-econometric research in Russia. In an unfavorable institutional environment, even low inflation can have a negative impact on economic growth, allowing more to save rather than spend. Let us also pay attention to the traditionally ignoring the business cycle nature of Russian monetary and budgetary policy, which in modern conditions is actually becoming pro-cyclical.

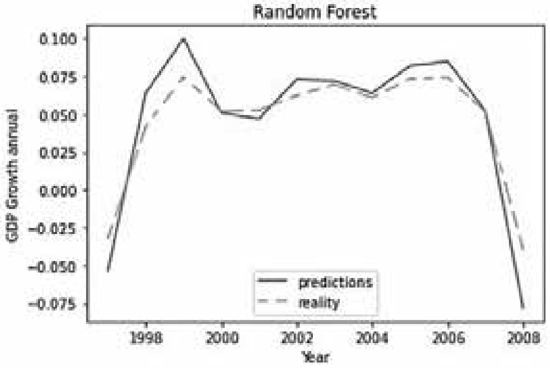


Fig. 3. Russian economic forecasts from 1998 to 2008.

The dependence of the Russian economy on external demand for raw materials and on world trends formed by world economic and political centers is one of the reasons for the absence in Russia of the need for autonomous forecasts of cyclical fluctuations. Implementing the ML techniques, we were able to increase the accuracy of our results. With advanced artificial calculations, the empirical result is very precise to real situations. We believe that the bunch of ML techniques will ensure and support further research in this field.

References

1. *Dalibor, S.* : Prévision macroéconomique dans l’ère des données massive et de l’apprentissage automatique.
2. *M. Khichane* : Le machine Learning par la pratique. Edition ENI
3. *Chaiboonsri, Chukiat Wannapan, Satawat.* (2019). Big Data and Machine Learning for Economic Cycle Prediction: Application of Thailand’s Economy. https://doi.org/10.1007/978-3-030-14815-7_29.
4. *Pontius, R.G. and Millones, M.* (2011) Death to Kappa: Birth of Quantity Disagree- ment and Allocation Disagreement for Accuracy Assessment. International Journal of Remote sensing, 32, 4407-4429.
5. *Cortes, C. and V. Vapnik,* Support-vector networks. Machine learning, 1995. 20(3): p. 273-297.
6. *Raikwal, J. and K. Saxena,* Performance evaluation of SVM and k-nearest neighbor algorithm over medical data set. International Journal of Computer Applications, 2012. 50(14).
7. *Abdi, M.J. and D. Giveki,* Automatic detection of erythemato-squamous diseases using PSO–SVM based on association rules. Engineering Applications of Artificial Intelligence, 2013. 26(1): p. 603-608.
8. Ministry of finance of Russian federation: http://minfin.gouv.ru
9. *Wannapan, S., Chaiboonsri, C., Sriboonchitta, S* : Prévision macro-économétrique pendant les périodes du cycle économique à l’aide d’un algorithme bayésien d’optimisation des valeurs extrêmes. Dans : Kreinovich, V., Sriboonchitta, S., Chakpitak, N. (eds.) TES 2018. SCI, vol. 753, p. 706 à 723. Springer, Cham (2018).
10. *Chakraborty, C., Joseph, A.* : L’apprentissage automatique dans les banques centrales. Document de travail n° 647 des services de la Commission. Banque d’Angleterre (2017).
11. *Bholat, D.:* Big data and central banks. Q. Bull. Q1, pp. 86-93 (2015). https://www.researchgate.net/publication/276101527_Big_Data_and_central_banks
12. *Chakraborty, C., Joseph, A.*: Machine learning at central banks. Stuff Working Paper No 647. Bank of England (2017).
13. *Hinge, D.*: Big Data in Central Banks. Published by Infopro Digital services Ltd. Central Banking Publications, London (2017).
14. *Signorini, L.F.* Harnessing big data machine learning technologies for central banks. The Printing and Publishing Division, Bank of Italy, Rome (2018).
15. *Coulombe, Philippe Leroux, Maxime Stevanovic, Dalibor Surprenant, Stéphane.* (2020). How is Machine Learning Useful for Macroeconomic Forecasting?
16. *Lahiri, S.N.* (2003). Resampling Methods for Dependent Data. Springer.
17. *Raffinot, Thomas.* (2017). Asset Allocation, Economic Cycles and Machine Learning.

18. *Hull, Isaiah.* (2021). Machine Learning and Economics. https://doi.org/10.1007/978-1-4842-6373-0_2.
19. *Statistics, L.B. Breiman, L.* (2001). Random forests. Machine Learning. 5-32.
20. *Ayres, Robert.* (2020). Economic Cycles, in Principle. https://doi.org/10.1007/978-3-030-39651-0_16.

Amos B.K. PhD student in IT, Department of Information Technologies, Peoples' Friendship University of Russia (RUDN), 6 Mikluho-Maklaya St, Moscow, 117198, Russian Federation, e-mail: broukouameamos9@gmail.com

Smirnov I.V. Associate Professor, Department of Information Technologies, Peoples' Friendship University of Russia (RUDN), 6 Mikluho-Maklaya St, Moscow, 117198, Russian Federation and also Head of Department No. 72, FIC "Informatics and Control" RAS, Institute for Artificial Intelligence Problems, Moscow, e-mail: ivs@isa.ru

Aidrous I.A. PhD. in Economics, Associate Professor at the Institute of World Economy and Business, Peoples' Friendship University of Russia (RUDN), 6 Mikluho-Maklaya St, Moscow, 117198, Russian Federation, e-mail: aidrous@mail.ru

Asmyatullin R.R. PhD. in Economics, Associate Professor at the Institute of World Economy and Business, Peoples' Friendship University of Russia (RUDN), 6 Mikluho-Maklaya St, Moscow, 117198, Russian Federation, e-mail: rav.asmyatullin@gmail.com

Glavina S.G. PhD. in Economics, Head of the Digital Economy Programme at the Institute of World Economy and Business, Peoples' Friendship University of Russia (RUDN), 6 Mikluho-Maklaya St, Moscow, 117198, Russian Federation, e-mail: sofiya.glavina@gmail.com

Quantitative large-scale study of school student's academic performance peculiarities during distance education caused by COVID-19*

V.A. Yunusov, A.F. Gilemzyanov, F.M. Gafarov, P.N. Ustin, A.R. Khalfieva

Kazan Federal University, Kazan, Russia

Abstract. The paper presents the large-scale analysis results of the distance learning impact caused by COVID-19 and its influence on school student's academic performance. This multidisciplinary study is based on the large amount of the raw data containing school student's grades from 2015 till 2021 academic years taken from "Electronic education in Tatarstan Republic" system. The analysis is based on application of BigData and mathematical statistics methods, realized by using Python programming language. Dask framework for parallel cluster-based computation, Pandas library for data manipulation and large-scale analysis data is used. One of the main priorities of this paper is to identify the impact of different educational system's factors on school student's academic performance. For that purpose, the quantile regression method was used. This method is widely used for processing a large-scale data of various experiments in modern data science. Quantile regression models are designed to determine conditional quantile functions. Therefore, this method is especially suitable to exam conditional effects at various locations of the outcome distribution: e.g., lower and upper tails. The study-related conditional factors include such factors as student's marks from previous academic years, types of lessons in which grades were obtained, and various teacher's parameters such as age, gender and qualification category.

Keywords: *Data Science, Big Data, Python, Dask, Quantile Regression, Conditional Quantile Functions, COVID-19.*

DOI: 10.14357/20790279230113

Introduction

The future of people and their future living standards are closely related to the education they receive. A higher education level is essential for achieving higher living standards. The education system should provide equal opportunities for equal success for everyone, regardless of individual and socio-cultural characteristics, socio-economic status, health status, pandemics and other factors. Failure to achieve educational goals will negatively affect a person throughout his life. Therefore, it is necessary to take a comprehensive approach to the problem of inequality in education. The COVID-19 pandemic is fundamentally changed the society, often exacerbating social and economic inequalities [2]. It is necessary to develop quantitative models based on modern methods of mathematical statistics in combination with BigData methods [12] to quantify the impact of the COVID-19 pandemic on educational systems.

Empirical quantitative analysis in education, psychology, and the social sciences is typically based on linear statistical models such as least squares regression (OLS), analysis of variance or covariance, or weighted linear models (e.g., at multiple levels) to calculate either mean scores of associations between dependent and independent variable or group differences in the dependent variable that controls the other independent variables included to the model. The regression coefficients obtained with such linear modeling approaches are averages, usually corrected for estimates of the effects of covariates present in the model, especially when analyzing observational or quasi-experimental data. In OLS regression-based modeling approaches, the regression coefficients show the "influence" of the independent variable x on the mean of the dependent variable y, taking into account the influence of the remaining independent variables [7,11].

* The study (all theoretical and empirical tasks of the research presented in this paper) was supported by a grant from the Russian Science Foundation, project № 22-28-00923, "Digital model for predicting the academic performance of school-children during school closings based on big data and neural networks".

Quantile regression was introduced nearly 30 years ago as an extension of the typical regression model (OLS) and addresses the shortcomings of the typical regression model by allowing to conditionally estimate different points (called quantiles) in a score distribution [10]. This method has become a comprehensive approach in linear and non-linear response models for conditional quantile functions. The quantile regression method, based on minimizing the residual of the "testing function", allows to evaluate all conditional quantile functions, just as the classical linear regression methods, based on least squares estimation, offer a mechanism for estimating conditional means of functions. In this sense, the regression median is a special case of the quantile regression model because the median is the 0.50 quantile (or 50th percentile). Thus, the quantile regression method is gradually becoming a unified statistical methodology and is widely used in education, economics, biology, ecology, finance, econometrics, statistics and applied mathematics [18].

In addition to assessing the impact of variables on different parts of distribution, quantile regression method has a number of other advantages over OLS. Firstly, it gives less weight to outliers of dependent variable compared to OLS. Secondly, it is a more reliable method because it allows to distinguish marginal effects of independent variables in quantiles of the dependent variable. Thirdly, when the errors are not normal, quantile regression estimates can be more efficient than OLS estimates. Finally, the semi-parametric nature of the approach weakens the restrictions on the constancy of parameters throughout the distribution of the dependent variable [6,7].

In this study, we used quantile regression method to study the impact of COVID-19 pandemic on school student's achievements by using a large data set covering data from all schools of Tatarstan Republic [19]. The data includes student's marks for main subjects for grades from 1 to 11, as well as the results of homework, tests, laboratory work, practical works, essays, tests, answers during the lesson, essays, presentations, dictations, etc. The whole database (records from 2015 till 2021) contains more than two billion information units, including information on the progress of more than a million students and the professional activities of more than 120,000 teachers.

The main purpose of this paper is to expand the quantile regression methodology by using BigData methods for using in educational analytics field. We've used quantile regression methodology to build a models of academic performance dependence for the entire academic year on many different factors: on grades for the last academic year and previous mouth, and on various teacher characteristics. This work is aimed to assess the quantitative factors of influence (regression parameters), caused by introduction of distance learning on the school student's academic performance. To solve this problem, we evaluated quantile regressions in the 10th, 25th, 50th, 75th, and 90th quantiles. The distance learning format caused by the COVID-19 epidemic was carried out in schools of the Republic of Tatarstan in April and May of the 2019-2020 academic year, therefore we studied this period in detail. In the 2020-2021 academic year, the schools of the Republic Tatarstan no longer switched to the distance learning and studied full-time as usual.

1. Review of Literature

Recently, a huge amount of various data has been accumulated in various in-formational systems and such databases exists in educational systems too. The analysis of such a large amount of data in the field of education analytics has becoming widespread in order to improve the educational and methodological process [13]. For example, in [17] authors use machine learning to perform an analysis of a large amount of student's data, including their academic performance, in order to identify those who are at risk of being expelled before the end of grade 9. Special emphasis is placed on the use of modern data analysis technologies: classification methods based on trees and support vector machines, thanks to which the forecast efficiency exceeds 90% [17].

Quantile regression is widely used in statistical analysis of educational data. This method is used to investigate the dependence of student's academic achievement in mathematics on factors like family background [18]. The analysis was based on a large sample of data from 2000-2002 years. As a result of the analysis, the authors determine the factors that significantly affect certain quantiles of the distribution of student grades, which makes possible to determine the impact of various indicators on their academic performance [18]. Quantile regression is an appropriate method for evaluating effects in different quantiles, including points in the upper and lower tails of the achievement distribution [14]. Therefore, quantile scores for multiple predictors can be obtained separately in the upper tail of the distribution at the 75th, 80th, 90th, or 95th percentiles [11].

A distinctive feature of quantile regression in the educational field is that it can be used to assess the degree of difference in the influence of factors on weak and strong students [20]. In work [1], by using quantile regression, the analyzed influence of grades received by students in the final year of the school on their performance in the university. As a result of the analysis of a large amount of data, the authors concluded that the average grade for the first half of the graduating class of the school and the grade in the certificate on academic performance at the university has a significant impact. Moreover, for students in the highest quantiles (based on the weighted average score for university courses), this effect is stronger than for students with poor performance. The authors also compare the results obtained with the results of a linear regression model and concluded that quantile regression allows to consider some important aspects of the relationship under study in more detail [1].

Also, quantile regression can be used to assess the influence of students on each other. The work [15] reports the effect of peers on PISA scores by analyzing a large sample of statistical data. The average score of peers in the class and the heterogeneity of their grades is used as independent variables of the regression model, and as the dependent variable is the result of the PISA exam. The authors conclude that for underperforming students these factors have a greater influence, than for students in the highest quantiles (for whom the effect is absent or becomes negative), which suggests that diversity is needed to achieve higher average results.

In addition, the work [4] used the method of quantile regression to analyze da-ta from large-scale studies of educational data of Italian schools conducted by the INVALSI institute. The regression model investigated the dependence of academic performance in mathematics and reading on various characteristics of both the students themselves and on geographical factors. As a result, in this work authors established characteristics that significantly affect student performance: gender, immigrant status and the distribution of performance across all regions of Italy. The authors also concluded that quantile regression is a powerful tool for building a model for multivariate statistical analysis of a large amount of data [4].

Quantile regression is also used to analyze Big Data, because big volumes of datasets make the estimation of regression parameters extremely difficult due to the vigorous computation and the limited storage space. The solution of this problem is described in paper [3]. Authors propose an approach, which simply saves the compact statistics of each data block and uses them to obtain an estimate of all the data with an asymptotically small approximation error, instead of processing all the data together. Another solution is proposed in work [21]. Authors use subsampling algorithm for the following use of composite quantile regression — an improved quantile regression method. Data is split into subsamples, and then the optimal subsampling is used for computing the resulting estimators. To deal with high-dimensional data in work [8] authors developed a new approach by using distributed computing. In this case, only the master machine computes penalized quantile regression estimations, while the other machines only compute subgradient of the local data. The efficiency of the proposed meth-od was confirmed on both the numerical simulation and prerecorded Big Data analysis [8].

2. Methods

2.1. Quantile regression

Quantile regression [10] can be considered as an extension of the least squares method for estimating conditional mean models to estimate an ensemble of models for multiple conditional quantile functions, by taking into account the effect of a set of covariates on the response variable [4] While the classical linear regression model detects the change in the conditional mean of the dependent variable associated with the change in covariates, the quantile regression model detects changes in the conditional quantiles. There-fore, since multiple quantiles can be modeled, a better understanding of how response distributions are affected by predictors can be gained by gaining information about changes in location, distribution, and shape. By analogy with the classical linear regression structure, the linear regression model for the θ-th conditional quantile y_i can be expressed as

$$Q_{y_i(\theta)|x_i} = x_i^T \beta_\theta \tag{1}$$

where y – is a scalar dependent variable, x_i^t - vector of k × 1 independent variables, β – coefficients vector, θ – the conditional quantile of interest, and it is assumed that

$$Q_\theta(u_{i,\theta} \mid x_{i,\theta}) = 0 \tag{2}$$

$u_{i,\theta}$ – residual term of the regression model in θ-*th* quantile.

From Equation 1, it turns out that in comparison to classical linear regression methods based on minimizing sums of squared residuals, quantile regression methods are based on minimizing asymmetrically weighted absolute residuals:

$$\min_\beta \sum\nolimits_{y_i \geq x_i^T \beta} \theta \mid y_i - x_i^T \beta \mid + \sum\nolimits_{y_i < x_i^T \beta} (1-\theta) \mid y_i - x_i^T \beta \mid \tag{3}$$

Substituting θ=3, equation (3) gives the median solution, and by using any θ from 0 to 1 allows to study the structure of the dependence anywhere in the conditional distribution of the response variable [4].

The estimation of the coefficients for each quantile regression is based on the weighted data of the entire sample [7].

The $\hat{\beta}_\theta$ coefficient in linear quantile regression models has the same interpretation as in other linear models, i.e.

$$\hat{\beta}_\theta = \frac{dQ_\theta(y_i \mid x)}{dx} \qquad (4)$$

means that each coefficient $\hat{\beta}_\theta$ can be interpreted as the rate of change of the θ-th quantile of the distribution of the dependent variable per unit change in the value of the corresponding regressor, keeping the rest unchanged.

However, important differences between least squares regression and quantile regression models relates to monotonic equivariance and robustness to distribution assumptions in conditional quantiles compared to these properties in the conditional mean setting [4].

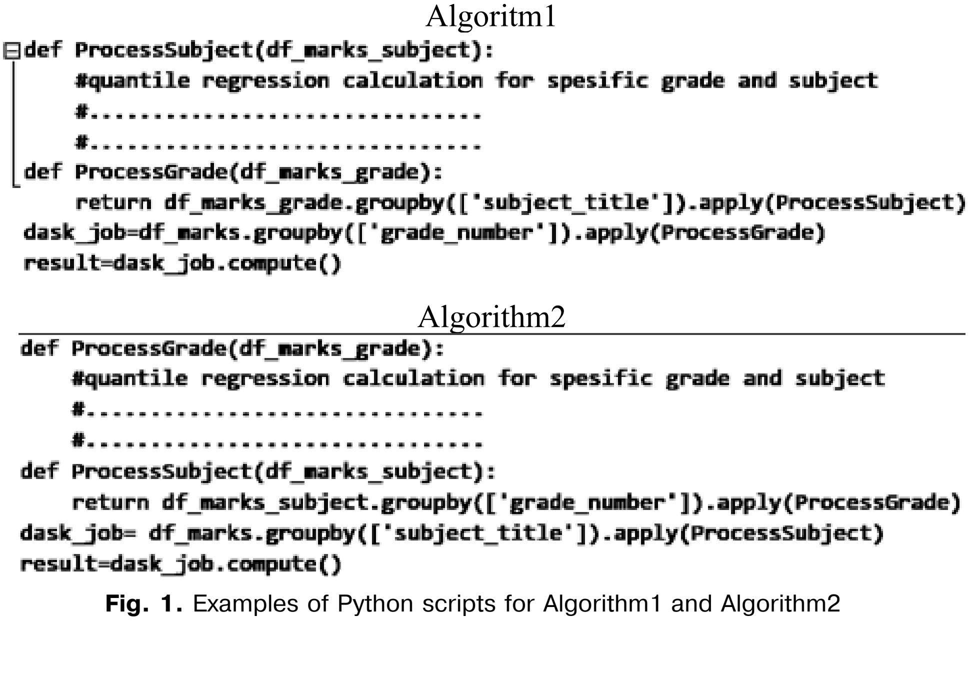
2.2. Dask-based HPC computational framework

For efficient process of a large amount of unstructured data we have to use Big Data methods, based on the power of computing clusters. In this work we used computational cluster containing 4 virtual machines (each VM has 1TB HDD, 32 GB RAM, 16 CPU cores), with parallel computing framework Dask installed. Dask is a flexible parallel big data processing library, designed to provide scalability and to extend the capabilities of existing Python packages and libraries [16]. The computational framework is based on Dask framework, because it very suitable for processing large datasets and Dask is able to perform computations with data volumes that are larger than the available memory of single computer [9, 5]. Dask has a dynamic task scheduler optimized for cluster based HPC computation, and "Big Data" collections like parallel lists, dataframes and arrays, and extend common interfaces like NumPy, Pandas, or Python iterators running on top of dynamic task schedulers [16].

We obtained anonymized datasets, describing different entities (grades, les-son topics, timetable, information about teachers and students) as separate csv of xml files. The total size of raw data files is more than 120 GB. At the initial pre-processing stage, the raw datasets (csv of xml files) were loaded into data structures called Dask DataFrames. We used Dask's DataFrame. merge() method to merge by some key the data frames obtained by loading different data files, be-cause the raw data has been scattered in different files. Grouped, reduced and aggregated DataFrames obtained from raw datasets subsequently were processed by using quartile regression methods in parallel mode (for different grades, sub-jects, etc). The analysis was carried out by using the quantile regression method implemented in the statsmodels.formula.api library.

As an example, here we briefly present Python scripts and a diagram describing the process of Dask framework based parallel calculation of quantile regression coefficients for one case (quantile regression coefficients calculation for specific grade and subject) (Fig. 1).

Data processing in parallel mode is started by calling student's mean marks containing dataframe's apply method, and by specifying the corresponding method name (ProcSubjects), which must be executed in parallel mode as a parameter of dataframe's apply method. To perform parallel processing for distinct grades, the following call to the apply method Process-Grade is used. The example of one kind of pipeline



Algoritm1

```
def ProcessSubject(df_marks_subject):
    #quantile regression calculation for spesific grade and subject
    #.............................
    #.............................
def ProcessGrade(df_marks_grade):
    return df_marks_grade.groupby(['subject_title']).apply(ProcessSubject)
dask_job=df_marks.groupby(['grade_number']).apply(ProcessGrade)
result=dask_job.compute()
```

Algorithm2

```
def ProcessGrade(df_marks_grade):
    #quantile regression calculation for spesific grade and subject
    #.............................
    #.............................
def ProcessSubject(df_marks_subject):
    return df_marks_subject.groupby(['grade_number']).apply(ProcessGrade)
dask_job= df_marks.groupby(['subject_title']).apply(ProcessSubject)
result=dask_job.compute()
```

Fig. 1. Examples of Python scripts for Algorithm1 and Algorithm2

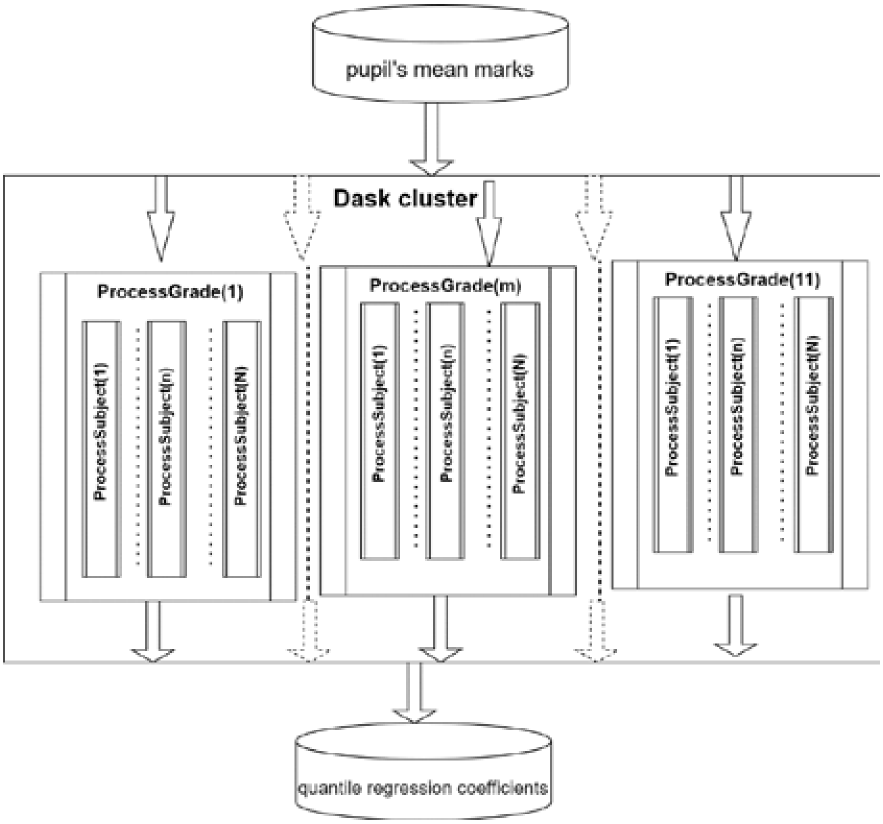


Fig. 2. Dask distributed based computational algorithm architecture for Algorithm 1

(Algorithm1) on Dask-based distributed data processing frame-work is shown schematically in Fig. 2.

The calculation process started in Dask cluster effectively executed in parallel mode (Fig. 2). The results presented in these graphs show the acceleration of the computing process when using a computing cluster in comparison with system based only 1 VM. For evaluation of the computational efficiency computations performed by cluster, we conducted a comparative analysis of the speed of performing calculations for Algorithm1 and Algorithm2, using only one node (1 VM) and full computational cluster (4 VMs). We also analyzed the influence of the npartitions parameter, which sets the number of data partitions into which Dask splits the initial dataframe at the beginning of processing (Fig. 3). By using a parallel algorithm

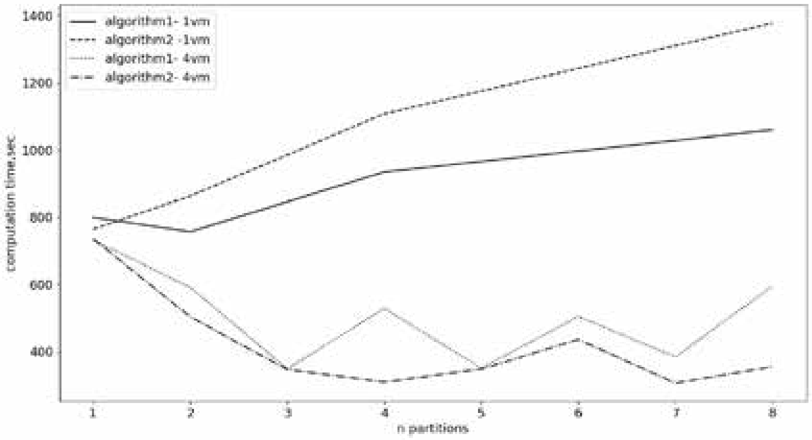


Fig. 3. Comparative analysis of computation time for Algorithm1 and Algoritm2 on a computing cluster consisting of 4 virtual machines and one virtual machine

based on the Dask cluster (with 4 VMs) speeds up the calculation process by almost 3 times (from 786 seconds to 280 seconds for Algorithm 2). The value of the npartitions parameter greatly influences the speed of calculations, with an increase this parameter value decreases the calculation time if the full cluster (4 VM) is used. Increasing this parameter value in the computational system containing only using 1 VM leads to an increase in the calculation time. The graphs also show that the use of Algorithm2 gives slightly higher performance compared to Algorithm1.

3. Results

The analysis of the data was carried out in several stages, divided according to the objects of the distinct study, and mostly performed in parallel mode by using Dask framework. Basically, we have maximally carefully analyzed 25% and 75% quantiles. The most significant results are presented in Fig. 4-7. The central line shows the values of the regression coefficients, dotted lines show 95 % confidence intervals for the regression coefficient's distribution.

At the first stage, we studied the dependence of school student's marks on the characteristics of teachers: age, qualification category, and gender for different subjects (Fig. 4). For comparison, we also provided the values of the quantile regression coefficients for the 2018-2019 academic year (full academic year without distance learning) Basically, before the of the distance learning begin in April month, the distribution of the quantile regression coefficient is the same as for all previous months, which indicates a similar distribution of marks for all students in these months.

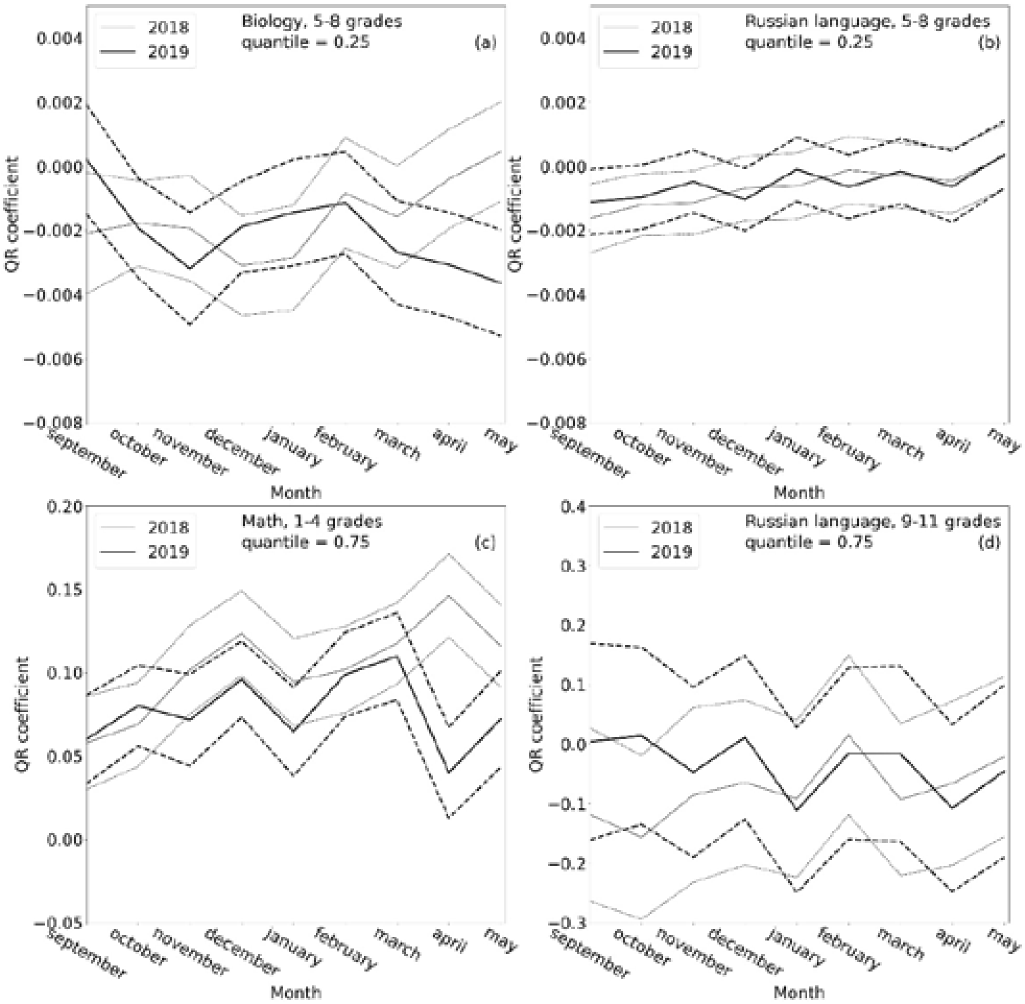


Fig. 4. Distribution of quantile regression coefficient describing the influence of the teacher's characteristics on student's grades for 2018-2019 and 2019-2020 academic years: (a) regressor variable - teacher age, 25% quantile, grades 5-8, subject Biology; (b) regressor variable - teacher age 25% quantile, grades 5-8, subject Russian language; (c) regressor variable - teacher category 75% quantile, grades 1-4, subject Mathematics; (d) regressor variable – teacher's gender, 75% quantile, grades 9-11, subject Russian language

Significant differences in quantile regression coefficient appears in April and May months for some subjects, and quantiles. In the case of a teacher's age as a regressor variable, differences appear in biology (Fig. 4(a)) and foreign language for 25% quantile (grades 5–8, not shown), i.e., for low-achieving students. But in other groups considered or in other subjects, there is no statistically significant differences (for example see Fig. 4(b)). By using qualification category as a regressor variable significant differences appeared in one of the main subjects: mathematics (Fig. 4(c)), geometry and algebra for all age groups and quantiles of 25% and 75%, this is especially noticeable for the April month. By using teacher's age as a regressor variable we discovered, that the distribution of the regression coefficient did not change significantly during distance education (for example see Fig. 4(d)).

We also noticed, that the values of the regression coefficients in the models for 2018-2019 academic year, describing the dependence of school student's marks on the characteristics of teachers, are usually higher than in the such models for the 2019-2020 academic year. This fact indicates that after the introduction of distance learning, the teacher's features began to play a smaller role in the distribution of regression coefficients for all students.

At the next stage, the dependence of marks for all subjects for the 2019–2020 and 2020–2021 academic years on marks for the previous academic year was studied (student mean marks for the previous academic year were taken as a regressor variable). In Figure 5 we present of the regression coefficients values for individual months, and for different subjects. It can be seen that immediately after beginning of the distance education (April-May 2020), the dependence of

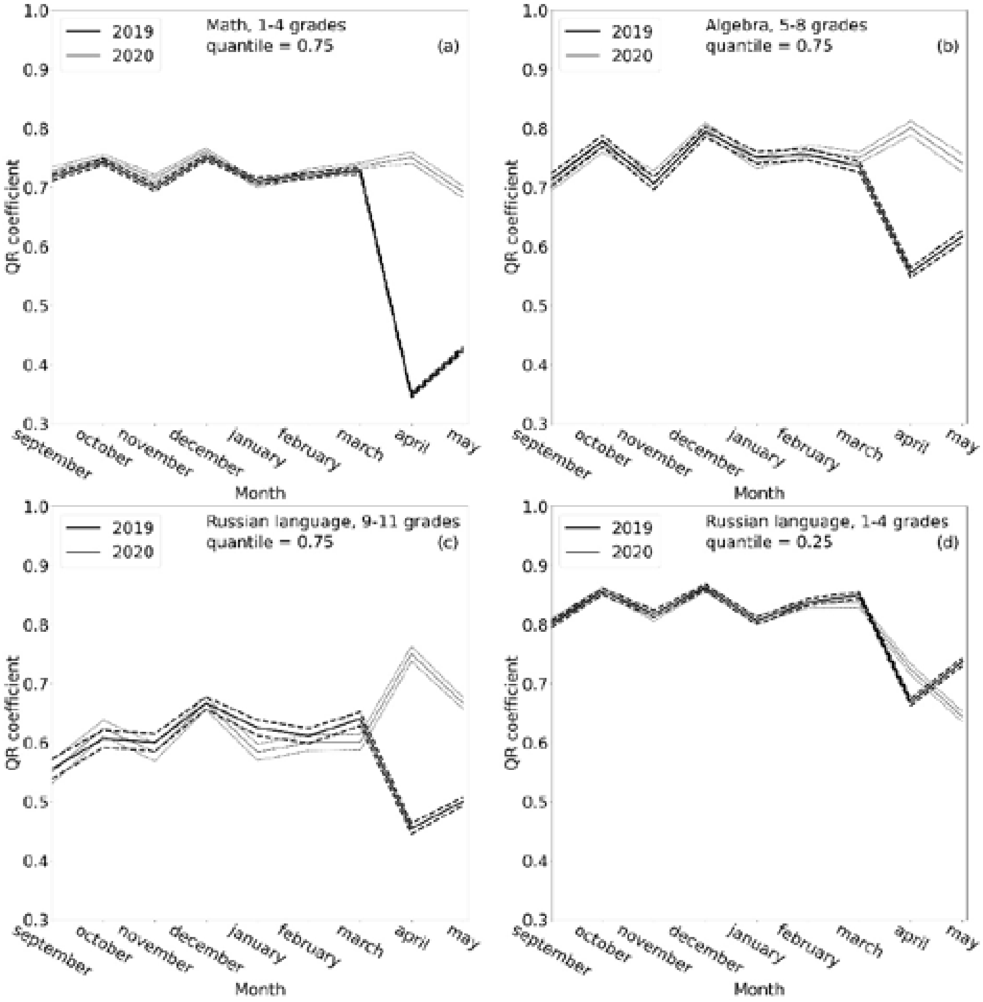


Fig. 5. The quantile regression coefficients, describing the dependence of marks on the previous academic year marks, for all subjects for the 2019–2020 and 2020–2021: (a) 75% quantiles, grades 1-4, subject Mathematics; (b) 75% quantile, grades 5-8, subject Algebra; (c) 75% quantile, grades 9-11, subject Russian language; (d) 25% quantile, grades 1-4, subject Russian language

grades on the grades of the previous year decreased significantly (the drop was up to 0.4, see Fig. 5(a)), which can be explained by the period of adaptation to the new format of education. Maximal drop of regression coefficient values is observed for 75% quantile (Fig. 5(a, b, c)), whereas for 25% quantile the lower values of this is drop observed (Fig. 5(d)). This feature is observed for all subjects and for different age groups of students. For the 2020-2021 academic year, there is no such sharp decline, what indicates the normalization of the educational process for this time.

A similar analysis of quantile regression coefficients conducted also for different types of assessment without division to subjects (Fig. 6). Here we observed a sharp decrease in the quantile regression coefficient for students of all age groups for "classwork" and "homework", 75% quantile (Fig. 3(a, c)) in April and in May of 2019-2020 academic year. At the same time, the regression coefficients for "control work" remained practically unchanged (Fig. 3(b)). Also, the regression coefficients for lover quantiles does not changes significantly (Fig. 3(d)). But just a year after the introduction of distance learning, the regression coefficient did not decrease, and in some cases even increased significantly (the increase was up to 0.35, see Fig. 3(d)).

At the final stage of the study, we analyzed the coefficients of quantile regression model describing the dependence of marks in the month of distance education start (April) on marks in the previous month (February). The detailed analysis was performed for distinct subjects and distinct types of assessment (the most characteristic results are shown in Figure 7). According these data, we concluded that during distance education for students from the 10%, 25% and 50% quantiles, nothing has changed in terms of academic subjects and types of assessment. However, for students in 75% and 90% quantiles, the transition to distant education has a critical impact, due to which the dependence of grades during distance education on grades for the pre-distance period decreased significantly (see, Fig. 7).

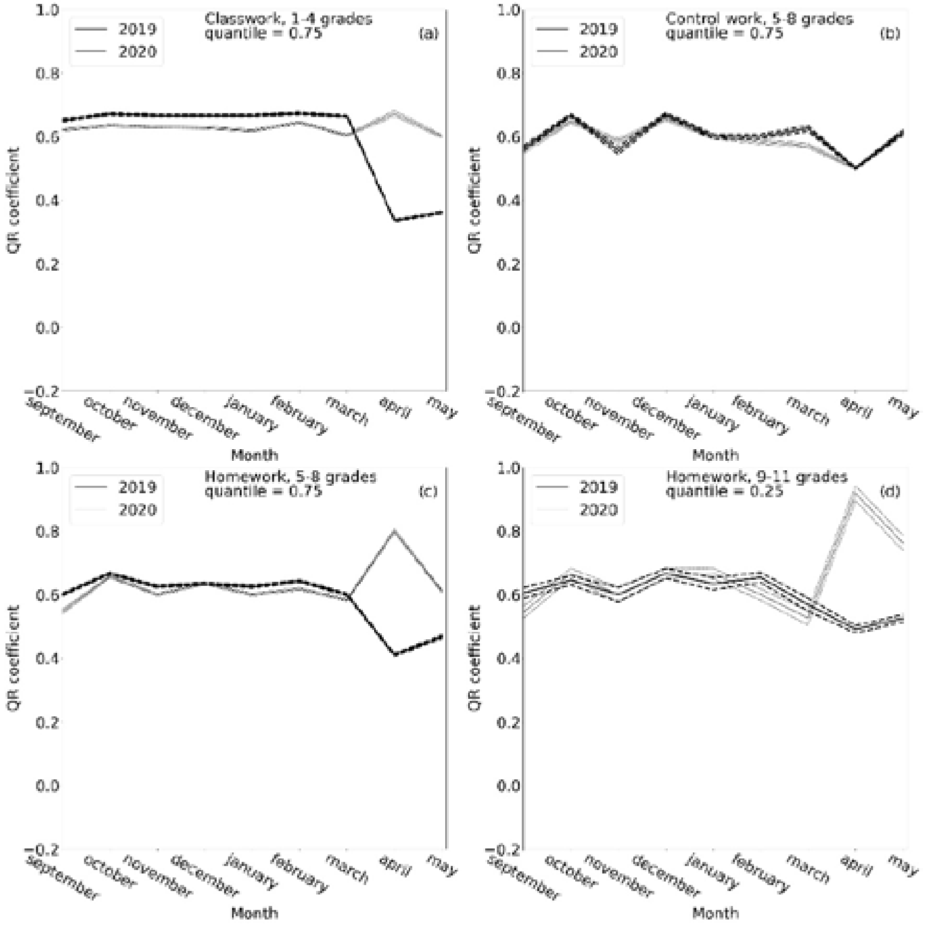


Fig. 6. The quantile regression coefficients for different types of assessment for 2019-2020 and 2020-2021 academic years: (a) 75% quantile, grades 1-4, type of assessment «class work»; (b) 75% quantile, grades 5-8, type of assessment «control work»; (c) 75% quantile, grades 5-8, type of assessment «homework»; (d) 25% quantile, grades 9-11, type of assessment «homework»

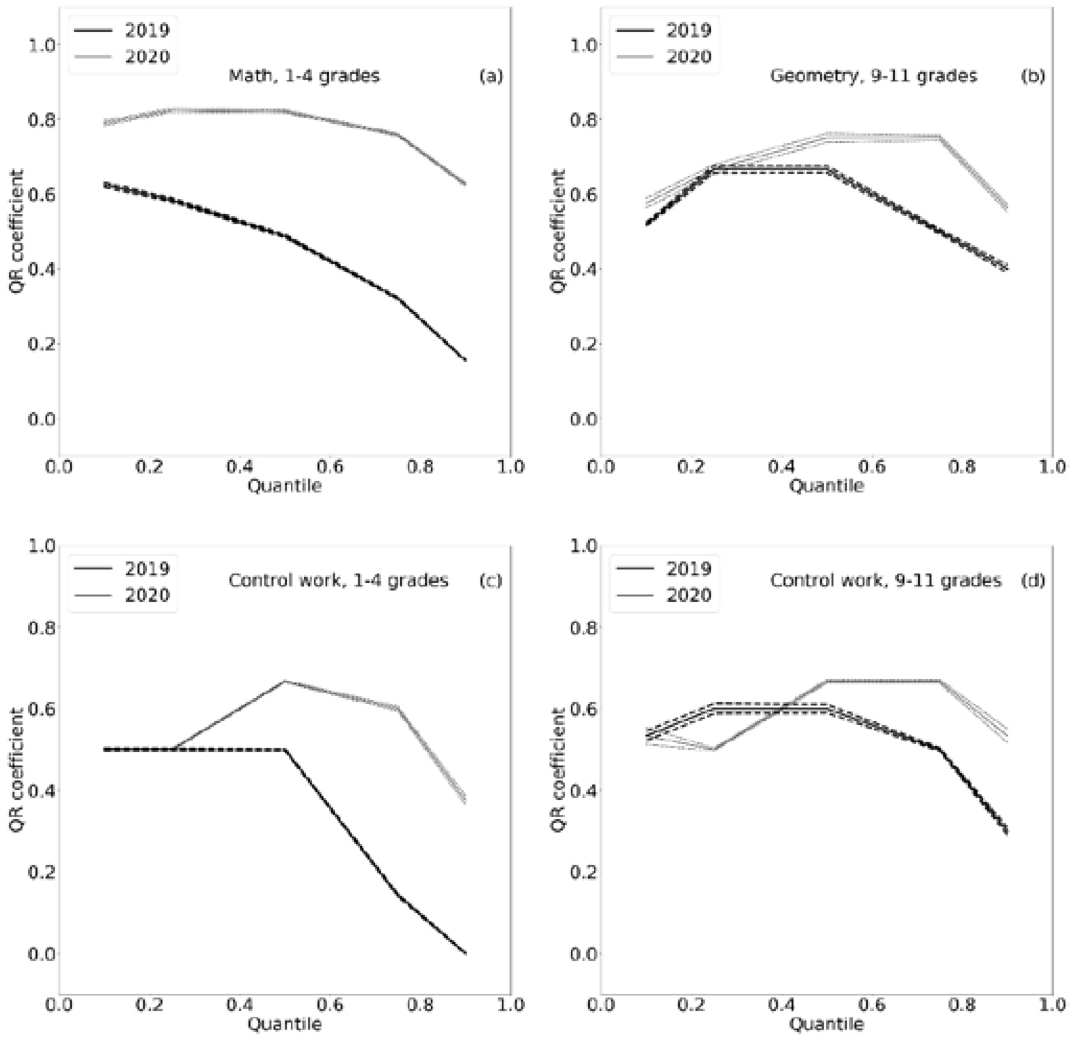


Fig. 7. Distribution of the 2018-2019 and 2019-2020 quantile regression coefficients de-scribing the dependence of April marks on the corresponding February marks: (a) grades 1-4, subject Mathematics; (b) 9-11 grades, subject Geometry; (c) grades 1-4, type of assessment «control work» ; (d) 9-11 grades, type of assessment «control work» . The values of quantile regression coefficients were plotted out against to the corresponding quantiles.

Conclusions

In this paper, we analyzed school student's academic performance in the peri-od before and during distance learning caused by COVID-19. The analysis per-formed on the basis of data obtained from "Electronic education in the Republic of Tatarstan" system. The analysis and interpretation of distance education effect is performed by using quantile regression approach. The Big Data processing framework Dask is used as a basis of data processing systems computational architecture. We developed high-performance cluster-based data processing pro-gram scripts for efficient quantile regression coefficients calculation in parallel mode, and performed analysis of the proposed algorithm's computational speed.

In the course of the study, we established that after transition to a distance learning, the first two months (April and May of 2019-2020 academic year) showed the greatest differences in the parameter values of quantile regression model, what indicate a period of adaptation to the new learning model. Statistically significant differences are existing both in the dependence of marks on the teacher's features, and for regression model coefficients for different years of study. Also, the parameters of quantile regression models significantly differ for different quantiles. Thus, it was possible to establish that during transitional moment (the period from February to April), the quantile regression coefficients for the group of students with high academic performance dropped sharply. This means that the dependence of April grades on

February grades has decreased in the 2019-2020 academic year, what means that if a student received high grades before, then new grades are less determined by old ones. These findings suggests that there are significant alterations in academic performance in different groups of students immediately after transition to distance learning format in April 2020. Moreover, there may be differential study-related factors effects at different points in the conditional distribution of school students' academic performance.

Marks in mathematics, geometry, algebra began to depend more on the teachers' qualification category for "weak" and "strong" students (changes are insignificant for "average" students). This result shows that in the face of unexpected and fundamental changes in the educational process, teachers' qualification is one of the stable factors influencing the assessment of the academic success of students in the field of technical disciplines. We also conclude that older teachers were worse adapted to the distance learning format during the period of mass digitalization of the educational process and assessed the progress of students in a simplified format.

Students did not immediately join the distance learning format, this caused them difficulties, therefore, their grades in the most difficult and important sub-jects decreased. It is easier for students to learn mathematics offline. As for the humanities and natural sciences, they are easier to perceive in a distance learning format. It was more difficult to learn such subjects as physics, algebra, geometry and the Russian language for school students in a distance learning format. Stronger students who are accustomed to express themselves in the classroom lose this opportunity in the distance learning format and therefore have lower grades compared to last year. Here we can talk about how limited the possibilities of the distance learning for-mat in the educational process are. This is also evidenced by the fact that by the new academic year the educational process has normalized.

During the distance learning format, the opportunities for the manifestation of each student are limited. Most likely, there is a distraction factor, when online students are more distracted, less focused on the educational process. While in the offline learning format, each student is in front of the teacher, everyone is included in the process of education, and strong students have more opportunities to actively manifest themselves, get involved in the process, and compete with each other. Here we can say about the importance of the influence of the educational environment on the manifestation of strong students, so on their grades. This confirms the fact that the test scores have not changed much. With the introduction of distance learning, strong students have sharply decreased their motivation to learn. Probably, for successful students, the opportunity to express themselves, communicate, and be included in the educational process as much as possible plays a big role. With the introduction of a distance learning format, these opportunities are reduced. Therefore, those students who had high and very high academic performance were no longer included in the learning process due to a decrease in interest in it. This was especially pronounced in basic subjects, as mathematics and the Russian language.

Thus, it can be concluded that the introduction of quarantine and distance learning format had a greater impact on students with high academic performance. This is especially true for such subjects as algebra, geometry, Russian language. It can be assumed that the online format does not allow to fully integrate into the educational process, and those students who can be actively involved in the traditional format, are interested in the learning process, lose this opportunity and the ability to actively participate in the online format.

References

1. *Amerise, I.L.* Predicting Students Academic Achievement: A Quantile Regression Approach. International Journal of Statistics and Systems 13(1), 9–14 (2018).
2. *Aspachs O, Durante R, Graziano A, Mestres J, Reynal-Querol M, et al.* (2021) Tracking the impact of COVID-19 on economic inequality at high frequency. PLOS ONE 16(3): e0249121. https://doi.org/10.1371/journal.pone.0249121
3. *Chen, L., Zhou, Y.* Quantile regression in big data: A divide and conquer based strategy. Computational Statistics & Data Analysis 144, 106892 (2020). https://doi.org/10.1016/j.csda.2019.106892 .
4. *Costanzo, A., Desimoni, M.* Beyond the mean estimate: a quantile regression analysis of inequalities in educational outcomes using INVALSI survey data. Large-scale Assess Educ 5, 14 (2017). https://doi.org/10.1186/s40536-017-0048-4
5. *Gafarov F, Minullin D, Gafarova V.* Dask-based efficient clustering of educational texts. CEUR Workshop Proceedings, 3036, 362–376 (2021).
6. *Gürsakal, Necmi & Murat, Dilek.* (2018). Assessment of PISA 2012 Results With Quantile Regression Analysis Within The Context of Inequality In Educational Opportunity. alphanumeric journal. 4. 41-54. https://doi.org/10.17093/aj.2016.4.2.5000186603 .
7. *Hao, L., Naiman, D.* Quantile regression. Sage, London (2007).

8. *Hu, A., Li, Ch., Wu, J.* Communication-Efficient Modeling with Penalized Quantile Regression for Distributed Data. Complexity, 2021, 6341707 (2021). https://doi.org/10.1155/2021/6341707
9. *Henriques, J., Caldeira, F., Cruz, T., Simões, P.* Combining K-Means and XGBoost Models for Anomaly Detection Using Log Datasets. Electronics 9, 1164 (2020). https://doi.org/10.3390/electronics9071164
10. *Koenker, R., Basset, G.* Regression quantiles. Econometrica, 46, 33–50 (1978). https://doi.org/10.2307/1913643
11. *Konstantopoulos S., Li W., Miller S., van der Ploeg A.* Using Quantile Regression to Estimate Intervention Effects Beyond the Mean. Educational and Psychological Measurement 79(5), 883–910 (2019). https://doi.org/10.1177/0013164419837321
12. *Li J., Jiang Y.* The Research Trend of Big Data in Education and the Impact of Teacher Psychology on Educational Development During COVID-19: A Systematic Review and Future Perspective. Front. Psychol. 12, 753388 (2021). https://doi.org/10.3389/fpsyg.2021.753388
13. *Park Y.-E.* Uncovering trend-based research insights on teaching and learning in big data. Journal of Big Data 7 (93), 1–17 (2020). https://doi.org/10.1186/s40537-020-00368-9
14. *Porter, S.R.* Quantile regression: Analyzing changes in distributions instead of means. In: M. B. Paulsen (Ed.), Higher education: Handbook of theory and research, vol. 30, 335–381. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-12835-1_8
15. *Rangvid, B.* School composition effects in Denmark: quantile regression evidence from PISA 2000. Empirical Economics 33, 359–388 (2007). https://doi.org/10.1007/s00181-007-0133-6
16. *Rocklin M.* Dask: Parallel Computation with Blocked algorithms and Task Scheduling. In: Proceedings of the 14th Python in Science Conference, pp. 126–132, (2015) https://doi.org/10.25080/Majora-7b98e3ed-013
17. *Sorensen, L.* “Big Data” in Educational Administration: An Application for Predicting School Dropout Risk. Educational Administration Quarterly 55, 404–446 (2019). https://doi.org/10.1177/0013161X18799439
18. *Tian, M.* A Quantile Regression Analysis of Family Background Factor Effects on Mathematical Achievement. Journal of Data Science 4, 461–478 (2006). https://doi.org/10.6339/JDS.2006.04(4).283
19. *Ustin, P., Sabirova E., Alishev T., Gafarov F.* Key Factors of Teacher’s Professional Success in the Digital Educational Environment. ARPHA Proceedings 5: 1747-1761 (2022) https://doi.org/10.3897/ap.5.e1747
20. *Yu, K.* Quantile Regression: Applications and Current Research Areas. Journal of the Royal Statistical Society Series D (The Statistician) 52(3), 331–350 (2003). https://doi.org/10.1111/1467-9884.00363
21. *Yuan, X., Li, Y., Dong, X., Liu T.* Optimal subsampling for composite quantile regression in big data. Statistical Papers (2022). https://doi.org/10.1007/s00362-022-01292-1

Yunusov V.A. Kazan Federal University, Kremlyovskaya St, 18, Kazan, Respublika Tatarstan, Russia, 420008, e-mail: valentin.yunusov@gmail.com

Gilemzyanov A.F. Kazan Federal University, Kremlyovskaya St, 18, Kazan, Respublika Tatarstan, Russia, 420008, e-mail: gilemal59@gmail.com

Gafarov F.M., PhD, Kazan Federal University, Kremlyovskaya St, 18, Kazan, Respublika Tatarstan, Russia, 420008, e-mail: fgafarov@yandex.ru (correspondent author)

Ustin P.N. PhD, Kazan Federal University, Kremlyovskaya St, 18, Kazan, Respublika Tatarstan, Russia, 420008, e-mail: pavust@mail.ru

Khalfieva A.R. PhD, Kazan Federal University, Kremlyovskaya St, 18, Kazan, Respublika Tatarstan, Russia, 420008, e-mail: khalfieva@inbox.ru

Методы и модели в естественных науках

Структура и эволюция рассеянных звездных скоплений: теория и наблюдения на основе данных Gaia*

Е.С. Постникова[I], О. Л. Рябухина[I], А.В. Тутуков[I],
С.В. Верещагин[I], Н.В. Чупина[I], А. П. Демидов[II]

[I] Федеральное государственное учреждение «Институт астрономии Российской академии наук», г. Москва, Россия
[II] Align Technology, Inc., г. Москва, Россия

Аннотация. Рассмотрено строение и эволюция рассеянных звездных скоплений (РЗС) на примерах РЗС Плеяды и группы РЗС в области Меча Ориона. Проведен отбор звезд по данным Gaia. Прослежена связь скоплений Меча Ориона с молекулярными облаками по данным космического аппарата «Гершель». Показано место рассмотренных объектов в общей схеме эволюции, составленной нами ранее. Сделан вывод о назревшей необходимости расширения классификации РЗС. Рассмотренная звездная система Плеяд показала наличие обширного звездного гало. Найденный в окрестности РЗС звездный поток Рыб − Эридана, вероятно, генетически связан с Плеядами и совместно с ним представляет остатки распавшейся ОВ ассоциации. В области Меча Ориона наблюдающиеся молодые РЗС, по всей вероятности, связаны с молекулярными облаками. Меч Ориона является видимой с ребра дисковой структурой продуктом столкновения двух гигантских молекулярных облаков. Данные о РЗС быстро пополняются, число РЗС растет за счет их выявления в обзорах Gaia. Анализ данной области может повторяться и расширяться по увеличивающемуся со временем объему данных с помощью проверенных методик, укладываясь в концепцию управления данными в областях с интенсивным использованием данных.

Ключевые слова: *информационная поддержка, рассеянные звездные скопления, Плеяды, область Меча Ориона, аналитика, управление данными.*
DOI: 10.14357/20790279230114

Введение

РЗС представляют собой звездные системы, связанные общим происхождением и гравитацией. РЗС являются необходимым инструментом для изучения эволюции как звезд, так и Галактики. Исследование звездных скоплений позволило предложить схему их классификации по внешнему виду (распределение плотности звезд по радиусу скопления с выделением ядра и короны) и свойствам, основанным на виде их диаграммы цвет-величина (ярчайшие звезды, возраст). Данные Gaia позволили детализировать эту схему, уточнить возрасты и расстояния скоплений от Солнца и

* Работа выполнена при финансовой поддержке РФФИ и в рамках научного проекта «Теоретическая химико-динамическая модель диска Галактики: рассеянные скопления как маркеры динамической эволюции», грант № 20-52-12009.

изучить структуру как центральной части скопления, так и далекой от центра периферию (гало, шлейфы (“копья”), удаленные шлейфы (“суперкопья”). Возросшая надежность данных о скоростях позволила более точно (на уровне точности менее ~1 км/с) изучать кинематику звездных скоплений, что важно для наиболее полного понимания того, как формируются и развиваются РЗС. Эта задача актуальна на данный момент, так как получив больше данных, возможно наиболее глубоко заглянуть в особенности движения молодых скоплений и находящегося вблизи них газа, что может дать ключ к наиболее точному понимаю того как образуются звездные скопления, как ведут себя после потери газа и насколько остаточное присутствие газо-пылевой структуры влияет на их свойства. Особенно наглядна для этого рассмотренная здесь область с несколькими близко расположенными друг к другу РЗС. Кроме того, стало возможно обнаруживать двойные и кратные скопления. Актуальность задачи обоснована и тем, что результаты, полученные КА Gaia [1], [2] и наиболее крупными космическими ИК телескопами “Гершель” и “Джеймс Уэбб” [3] делают возможным изучать эволюцию РЗС путем сопоставления данных об РЗС и газе. Задача актуальна и тем, что данные Gaia активно используются для открытия новых, ранее неизвестных РЗС [4], [5]. Число выделенных РЗС может увеличиться на порядок и достигнуть нескольких десятков тысяч.

1. Краткая история классификации РЗС

По мере накопления наблюдательных данных стали понятны различия между отдельными РЗС, которые позволили поставить вопрос об описании их многообразия с помощью разработанной классификации. Поначалу такая классификация строилась на различии внешнего вида скоплений. Классификация Бэйли 1908 [6] ориентирована на учет степени концентрации звезд, пространственную форму скопления (правильной, сравнительно правильной, неправильной, рассеянной) и параметрам ярчайших звезд. С помощью каталога РЗС, составленного Мелотте (1915) [7], построена классификация РЗС на основе предложенной схемы Бэйли [6]. Позднее работа была продолжена Рааб (1922) [8]. Вскоре Трюмплер (1925) [9] предложил физически обоснованную классификацию, основанную на диаграмме Герцшпрунга-Рессела (Г Р). При этом был использован спектральный класс самой яркой звезды главной последовательности. Эта классификация была развита в работах Маркаряна [10]. Таким образом, основой классификации стали внешний вид РЗС и диаграмма цвет – звездная величина. Современный взгляд на устройство РЗС показан на Рис.1.

2. Преимущества данных Gaia

На текущий момент поиск новых РЗС осуществляется в автоматическом режиме. Так как большие объемы данных наблюдений не позволяют быстро и качественно отождествлять такие источники визуально, к тому же слабонаселенные звездами скопления часто сливаются с общим фоном и их удобнее отождествить по совокупности нескольких параметров одновременно. Визуально нереально просмотреть и выделить на небесной сфере РЗС, рассматривая до 1.8 млрд. звезд Gaia. Однако, автоматизация поиска новых РЗС может приводить к тому, что одно и тоже или известное РЗС выделяются повторно. При этом рассматриваются звезды до 20 звездной величины, что на несколько величин слабее рассмотренных ранее (в каталоге MWSC, Milky Way Star Clusters, Харченко и др. 2013 [11]). Использование высокоточных данных Gaia позволило говорить о дополнительных параметрах классификации скоплений, включая не только форму РЗС и особенности диаграммы Г-Р, но следующие показатели. Это 1) двойственность и эволюционный статус, 2) наличие экзопланет, 3) степень развития шлейфов, 4) принадлежность к группе скоплений, представляющей родительскую OB ассоциацию.

3. Современные представления о строении РЗС

РЗС состоит из центральной части (ядра), короны (полости Роша), шлейфов диссипативной природы, или “звездного копья”, включающего потерянные звезды и АКП (астероиды, кометы, планеты) объекты, а также и “суперкопья”. Последние включают звезды, покинувшие распавшиеся РЗС, и выжившие РЗС распадающейся родительской звездной ассоциации. Все перечисленные структуры и строение РЗС показано на Рис.1.

4. Звездные и АКП потоки

Природа движущихся звездных потоков долгое время оставалась неопределенной. Вероятно, М. Фламмарион первым обратил внимание на возможное существование линейной цепочки звезд “celestial road” Плеяд (Холмс 1894) [12]. Проктор (1869) [13] нашел сходство скоростей звезд в Тель-

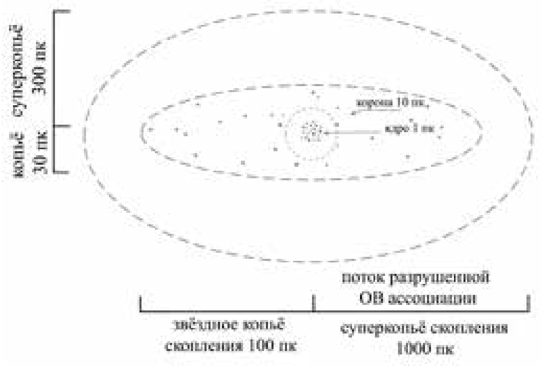

копьё 30 пк
суперкопьё 300 пк
корона 10 пк,
ядро 1 пк
поток разрушенной ОВ ассоциации
звёздное копьё скопления 100 пк
суперкопьё скопления 1000 пк

Рис. 1. Структура РЗС с массой ~10^3 $M_\odot$ с указанием масштабов

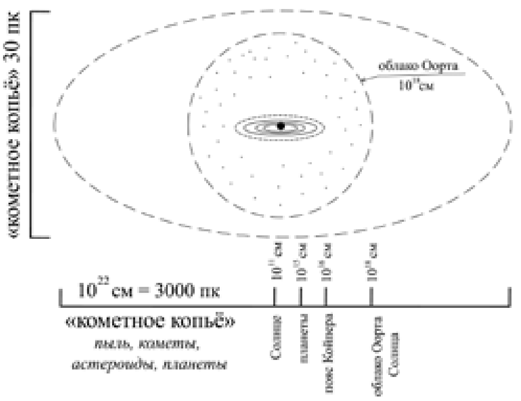


Рис.3. Структура Солнечной системы

це со скоростью звезд Плеяд. С самого начала было понятно, что для надежного отождествления звездных потоков в Галактике вблизи Солнца необходима надежная информация о скоростях большого количества звезд и определение их расстояний от Солнца (Эддингтон 1909) [14].

Скорее всего, большинство звезд рождаются в звездных скоплениях. Последние, как правило, распадаются после потери ими газовой компоненты Тутуков (1978), [15], Рис.2. В работе Тутуков и др. (2021) [16] с помощью простых численных моделей проведен анализ формирования и эволюции потоков, состоящих из астрофизических объектов различной природы, включающих звезды, звездные скопления, астероиды, кометы и планеты (Рис.3). Практически все обозначенные потоки сейчас наблюдаются. В работе Тутуков и др. (2021) [16] построены численные модели потоков, генерируемых кометами, астероидами, звездами и их скоплениями, галактиками в их скоплениях и рассмотрены условия разрушения исходных объектов.

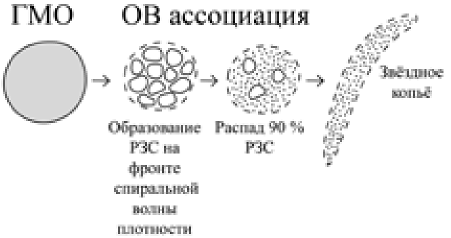


Рис.2. Сценарий эволюции РЗС и формирования звездных потоков

5. Природа звездных потоков

Назовем возможные причины разрушения астрономических объектов, ведущие к появлению потоков и их элементов. Все планетные системы пронизывают астероидно-кометные и/или метеороидные потоки. Они являются продуктами разрушения астероидов и комет. Следует отметить, что такие потоки необязательно являются продуктами разрушения астероидов и комет. Они могут образовываться и путем конденсации газопылевых продуктов, яркий пример – планетезимали.

Как уже говорилось выше, планетные системы обладают так называемыми “копьями”, представляющими собой астероидно-кометные фрагменты (пример планетной системы Солнца на Рис. 3). Длина копий определяется возрастом их родительских звезд и может достигать галактических масштабов, Тутуков и Смирнов [17]. Звездные скопления обладают АКПЗ копьями, то есть наряду с АКП объектами включают в свой состав и звезды. Распад ОВ ассоциаций ведет к появлению широких АКПЗ копий и потоков, включающих в свой состав и отдельные скопления, сохранившиеся после исходной потери ими газа. Разрушение спутников массивных галактик ведет к появлению звездных потоков галактических масштабов. Разрушение галактик в ходе столкновений внутри скоплений галактик превращает эти галактики в широкие звездные потоки, составляющие, в конечном итоге, непрерывный звездный фон скоплений галактик.

Назовем основные процессы разрушения наблюдаемых объектов и систем. Ледяные ядра комет разрушаются в процессе испарения льда и выброса пыли и камней давлением водяного пара на горячих орбитах вокруг своих звезд. Астероиды разрушаются в ходе их взаимных столкновений. Массивные, далекие от своих звезд планеты ускоряют АКП своей гравитацией. Зоны HII разрушают большинство молодых звездных скоплений за счет ослабления гравитационной связи скопления в целом и за счет выметания газа давлением излучения, а вместе с ним и части массы системы. Кроме того, гравитационное взаимодействие звезд внутри скоплений ведет к постепенному испарению скоплений. Вспышки звездообразования и сверхновые звезды в сфероидальных галактиках, ведущие к быстрой потере газа этими галактиками, могут вести к их разрушению.

6. Межгалактические звездные потоки

Известно, что столкновения галактик между собой и приливные взаимодействия между ними могут сопровождаться частичным или их полным распадом, Вшивков и др. (2011) [18]. Приливные хвосты взаимодействующих галактик стали еще одним примером звездных потоков, возникающих в ходе разрушения их периферийных областей (Танаки 1981 [19], Борне и др. 2004 [20]). Со временем стало ясно, что часть линейных структур галактик являются звездными потоками – остатками галактик низкой плотности, разрушенных приливами массивных галактик (Гилмар 2006 [21]). Исследование плотных частей и гало Галактики показало присутствие в нем многих звездных потоков – следов разрушения ее близких спутников (Малхан и др. 2021 [22]) или же вспышками звездообразования в них. Такие структуры составляют в итоге «непрерывную» звездную среду скоплений галактик (Тутуков и др. [16]). Это явление было признано общим в мире галактик. В звездных потоках найдены следы звездных скоплений.

7. Звездная система Плеяд

Известны наблюдаемые приливные структуры, которые обнаружены вблизи шаровых скоплений, например, Pal 5 (Оденкирхен и др., 2003 [23]) и Pal 12 (Гриллмайр и др., 1995 [24])). Обнаружено, что РЗС Плеяды и альфа Персея представляют собой два потока – остатки ОВ ассоциации (Эгген 1998 [25]).

Нами проведен поиск приливных структур в районе 200 пк вокруг центра скопления Плеяды на основе Gaia DR2. В качестве инструмента поиска использовался метод сходящихся точек, описанный ван Лювен 2009 [26]. Он эффективен для обнаружения звезд с тангенциальными скоростями, близкими к скорости РЗС Плеяды. При отборе также учитывалось положение звезды и направление ее скорости в пространстве. Определен возраст скопления и размер находящегося рядом с ним звездного потока, который составляет приблизительно 100 пк. Определено его расположение относительно скопления. Получен и каталогизирован список использованных звезд. Процедура отбора применена для наиболее качественных измерений, описанных Линдегрен и др. (2018) [27], для получения выборки, очищенной от возможных артефактов. Кроме того, значение перенормированной ошибки единичного веса (RUWE), определяемое формулой Линдегрен и др. (2018) [27] было принято <1.4, в результате чего исключены звезды, которые в силу своей неразрешенной двойственности или проблемы с определением астрометрических параметров, не подходят для нашего исследования. Вдобавок к этому, согласно процедуре, приведенной в Gaia DR2 documentation [28], применено «сокращение коэффициента избытка потока». В результате получена фотометрически и астрометрически чистая выборка. Для повышения точности результатов пришлось пожертвовать звездами слабее G = 15 mag. Таким образом, определив центр скопления и применив описанную выше процедуру отбора, на начальном этапе было выделено 610 548 звезд. На Рис.4 показано распределение этой выборки по направлениям пространственных скоростей и распределение в проекции на плоскость Галактики XY. В последнем случае заметно, что пространственная форма отличается от сферически симметричной. Также на правой панели Рис.4 заметен звездный поток, расположенный близко к Плеядам и сходный по кинематике и возрасту с ними. Его звезды достаточно сложно отделить от звезд скопления в центральной части, но сам поток выделяется как сгущение звезд вдоль оси X по обе стороны от скопления.

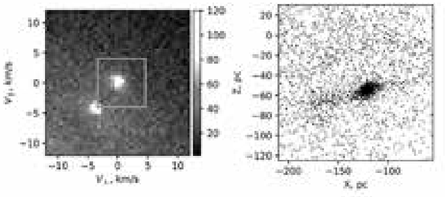


Рис. 4. Левая панель: Распределение звезд в плоскости тангенциальных скоростей. Квадратом в центральной части рисунка показана область отбора звезд РЗС Плеяды. Правая панель: распределение звезд в ZX-плоскости после выделения по направлению тангенциальных скоростей

На Рис. 4 нулевая точка — это точка схождения векторов тангенциальных скоростей звезд. Прямоугольником ограничена область, в которой производится поиск звезд скопления. Избыток плотности в левом нижнем углу обусловлен звездами РЗС альфа Персея. Избытки плотности звезд слева и справа от скопления Плеяды – части связанного с ним потока.

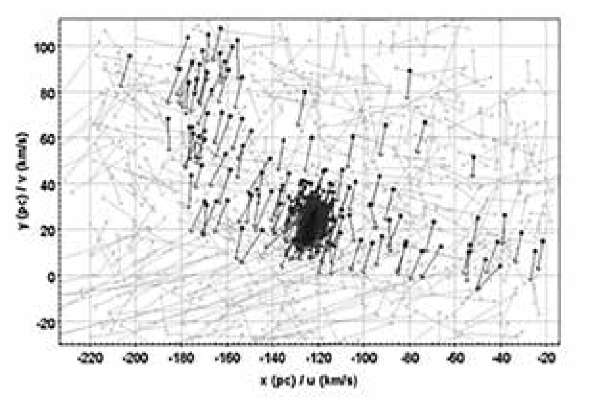


Рис. 5. Движение звезд Плеяд и связанного с ними потока (жирные стрелки) среди фона звезд (серые стрелки) в галактической плоскости

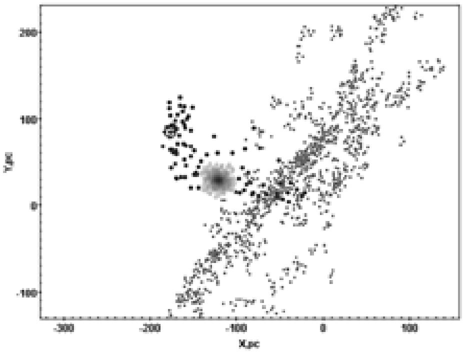


Рис. 6. Распределение звезд скопления Плеяды (область точек в центре – выборка звезд из [29]), связанного с ними потока (жирные точки) и потока Рыб-Эридана (серые точки справа от Плеяд, звезды из [29]) на плоскости XY

На Рис.5 видна вытянутая пространственная форма потока, связанного с РЗС Плеяды, а на и Рис.6 также потока Рыб-Эридана, (Розер и др. 2020 [29]). Однако эти две структуры имеют неодинаковое положение относительно друг друга и, скорее всего, являются разными структурами.

Для получения распределения звезд, представляющих звездные потоки («копья» и «суперкопья») в пространстве, представленные на Рис. 5 и Рис. 6 применена следующая методика.

Сначала был определен центр скопления Xc, Yc, Zc = (-120.8, 29.1, -54.3) по данным из Конте-Годе и др. (2018) [30] и собственные движения также из [30], а также средняя лучевая скорость из Галли и др. (2017) [31]. Поиск звезд приливных шлейфов (“копий”) осуществляется в пространстве Галактики, поэтому параметры центра Плеяд, переведены в барицентрические галактические декартовы координаты.

Вокруг указанного центра Xc, Yc, Zc (где X направлена на ось вращения Галактики (l=0°, b=0°), Y – в направлении вращения Галактики (l=0°, b=0°), Z – на северный галактический полюс (b=0°) были отобраны звезды из GDR2 в радиусе 200 пк.

Одновременно была применена процедура отбора наиболее качественных измерений, описанных в [27] Линдегрен и др. (2018, глава 4.3 и Приложение С, рисунки С.1 и С.2), чтобы получить выборку, очищенную от возможных ошибок и артефактов. Кроме того, значение перенормированной ошибки единичного веса (RUWE), определяемой в [27] было принято <1.4, в результате чего из выборки были исключены звезды, у которых возможна неразрешенная двойственность или присутствуют проблемы с определением астрометрических параметров, которые также не укладываются в модель одиночной звезды. Кроме того, в соответствии с процедурой, описанной в Браун и др. (2018) [32], была применена отсечка «коэффициента избытка потока». В результате осталась фотометрически и астрометрически чистая выборка, также для большей точности результатов пришлось пожертвовать звездами слабее G = 15mag.

После такого отбора звезд по их расположению и по качеству был использован метод точки схождения, описанный в ван Лювен (2009) [26]. Этот метод хорошо подходит для близких скоплений и ранее успешно применялся для поиска шлейфов скоплений Гиад и Ясли [33] Розер и др. (2019), также он применим и к Плеядам. Основным преимуществом метода является его применимость при недостатке лучевых скоростей, поскольку он основан на тангенциальной составляющей скорости.

Для выбора наиболее надежных членов скопления были отобраны только те звезды, лучевые скорости которых известны. Это условие позволяет наиболее полно рассмотреть их перемещение в пространстве. Для предварительной оценки дисперсии скоростей скопления с учетом шлейфов была сделана выборка из членов скопления по Лоди и др. (2019) [34] среди звезд со всеми параметрами определения скоростей. По этим данным значение дисперсии скоростей составляет около 10 км/с с учетом, как звезд скопления, так и шлейфов. Это значение кажется завышенным по сравнению со средней оценкой дисперсии в рассеянных звездных скоплениях, которая обычно принимается равной 1-3 км/с, Чумак и Расторгуев (2006) [35], а с учетом дополнительной дисперсии в шлейфах порядка 1 км/с может достигать 4 км/с. Также полученный здесь результат может иметь место из-за больших ошибок в определении лучевых скоростей

Компоненты пространственных скоростей звезд скопления также сводятся к экваториальному галактическому преобразованию, объясненному в

Джонсон и др. (1987) [36]. Таким образом, средние компоненты пространственной скорости скопления Uc, Vc, Wc = (6.71, -28.54, -14.18) км/с, где U направлено к антицентру Галактики, V – в направлении вращения галактики, а W – к северному галактическому полюсу.

На рис. 5 и рис. 6 показаны векторы пространственных скоростей UVW на галактической плоскости XYZ. Отбор по пространственным скоростям задал ограничение на дисперсию скоростей звезд не более10 км/с от средней пространственной скорости скопления. Также отклонение направления векторов скорости (U,V,W) от направления центрального вектора направления (Uc, Vc, Wc) скопления не должно отклоняться более чем на 10 град.

8. Звездно-газовая структура области Меч Ориона

Некоторые РЗС содержат, кроме звёзд, облака газа и / или пыли. Нами составлена синтетическая карта звездно-газового состава области Меч Ориона. Она позволила изучить возможные случаи связи скоплений и газовых облаков.

Метод получения звездной структуры области Меча Ориона. Сделана выборка звезд из каталога Gaia EDR3 в области Меча Ориона. В этой области находятся РЗС NGC 1981, NGC 1977, NGC 1976, NGC 1980, Рис.7 (левая панель). Взяты звезды Gaia EDR3, расположенные на расстояниях от Солнца от 408 до 377 пк (рассчитаны по параллаксам Plx, приведенным в каталоге Gaia [1,2]).

Метод получения газовой структуры области Меча Ориона. Тепловое излучение холодной пыли лежит в дальнем инфракрасном диапазоне и его анализ может быть использован для получения физических параметров, например, температуры вещества и лучевой концентрации [38, 39]. На Рис. 7 (средняя панель) и более детально на Рис. 8 показана полученная в данной работе карта кинетической температуры пыли (от 15К до 35К) по данным наблюдений космического телескопа «Гершель». Для этого было выполнено моделирование спектрального распределения энергии излучения пыли по данным «Гершель». Из архивов были скачены карты излучения на 160, 250, 350 и 500 мкм, все карты приведены к единицам Jy/pixel, $S_\nu = 2h\nu^3/c^2\,(\exp(h\nu/kT) - 1)\,(1 - \exp(-\tau_\nu))$, $\tau_\nu = \mu_{H2}\, m_H\, \kappa_\nu\, N(H_2)$, здесь μ_{H2} – средний молекулярный вес = 2.8 [41], m_H масса водорода, $N(H_2)$ – лучевая концентрация, κ_ν непрозрачность пыли.

Проведена процедура конволюции, все карты приведены к единому разрешению и размеру пикселя как на карте−500 мкм. С использованием 4-х карт попиксельно (количество пикселей 576050) была вписана модель абсолютно черного тела со свободными параметрами T_d и $N(H_2)$ – температура пыли и лучевая концентрация водорода [40]. В результате получены карты распределения данных параметров. Размер пикселя в получившихся картах 14”, размер получившихся карт 3° * 3.5°, что на расстоянии в 400 пк соответствует площади 21 * 24 пк с центром в точке ($5^h33^m36^s$, -5°02′27″). Получившаяся лучевая концентрация водорода лежит в пределах от $5.6\cdot10^{17}$ до $1.9\cdot10^{23}$ с медианным значением $5.30\cdot10^{20}$, среднеквадратичное отклонение $2.9\cdot10^{21}$, rms = $3.25\cdot10^{21}$. Медианная температура 19 K, среднеквадратичное отклонение 4 K, rms = 20 K.

Построены по картам телескопа Гершель. Наблюдения на телескопе ИСЗ Гершель проводились на нескольких частотах в диапазоне от 160 до 500 микрон. Нами получены карты лучевой концентрации водорода $(1\text{-}10)\cdot10^{21}$ см$^{-3}$ и кинетической температуры пыли от 15 до 35 К. Выполнена следую-

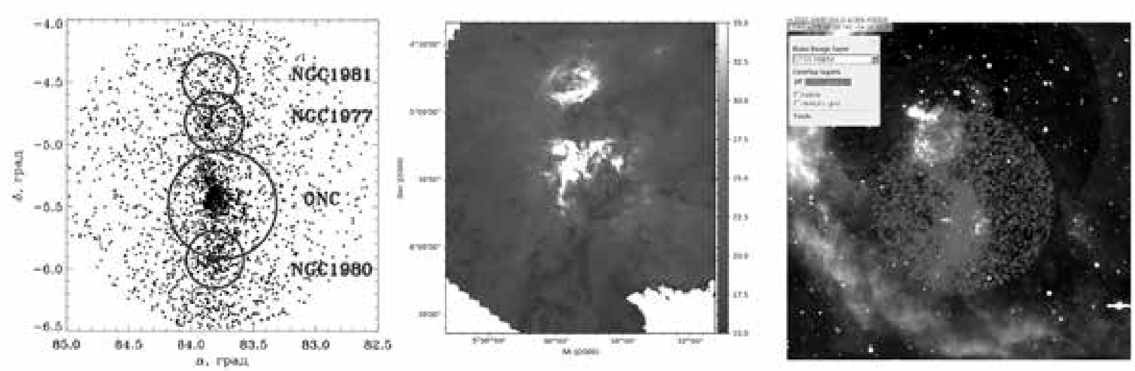


Рис.7. Распределение звезд, пыли и горячего газа в экваториальной системе координат. Левая панель показывает положение рассмотренных РЗС на небесной сфере. Средняя панель − распределения кинетической температуры пыли, полученные нами по данным космического телескопа “Гершель”, правая панель − распределение горячего газа в по данным VTSS [37], совмещено со звездами левой панели

щая обработка данных: 1)взяты карты в указанных 4-х диапазонах, 2)приведены к единому разрешению, 3)для каждого пикселя вписано функция излучения абсолютно черного тела для определения температуры и плотности для каждого пикселя.

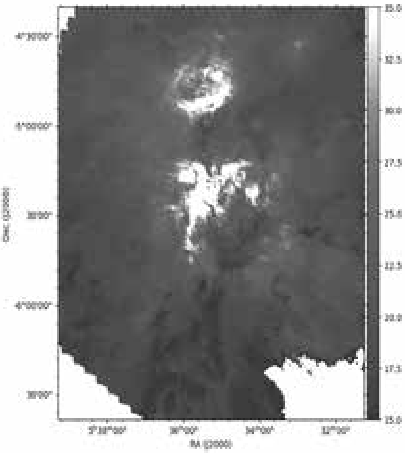


Рис. 8. Полученное нами детальное распределение кинетической температуры пыли (справа шкала температуры) по данным наблюдений «Гершель»

Рассмотренная группа РЗС в пространстве образует вытянутую с севера на юг структуру Меча Ориона размером в проекции на небесную сферу около 20 пк. Это плоская структура видимая с ребра. Ее расстояние от Солнца составляет ~400 пк. Нами рассчитаны температура и концентрация частиц на луче зрения для каждого пикселя. Сопоставление распределений звезд и указанных параметров излучения межзвездного газа и пыли (Рис.7 и Рис.8) позволили сделать выводы об эволюционной стадии цепочки РЗС, входящих в состав структуры Меча Ориона. Пыль наблюдается в РЗС NGC 1981, NGC 1977, облака горячего газа совпадают по положению на небе с областью занятой NGC 1976. РЗС NGC 1980 вероятно не содержит больших масс газа. Пыль связана с самыми молодыми скоплениями, в которых еще нет массивных звезд. Горячий газ появляется после образования в РЗС массивных звезд и связанных с ними областей ионизованного водорода HII. Мы видим рождение ОВ ассоциации в результате столкновения двух газовых облаков.

9. Двойные скопления

Двойственность РЗС отражает физические процессы во время фрагментации газового облака. Мы разделили наблюдаемые пары скоплений на визуально двойные (ВДРС) и тесные пары (ТДРС), Рис.9. Тесные пары – это физически связанные скопления, то есть такие объекты, массы которых, расстояния между которыми и относительные скорости допускают их гравитационную связанность. Анализ современных каталогов РЗС позволяет отобрать потенциально физически связанные пары и классифицировать РЗС согласно областям, выделенным на диаграмме Рис. 9, Верещагин и др. (2022) [42]. Также на диаграмме Рис. 9 показаны значения массы пар РЗС (выделены M = 100, 1000, 10000 $M_{\odot}$). Подставляя указанные значения для массы M, последовательно получены зависимости $\delta R = GM/(\delta V)^2$, где расстояние между скоплениями в паре - δR и разность их пространственных скоростей - δV, G – гравитационная постоянная.

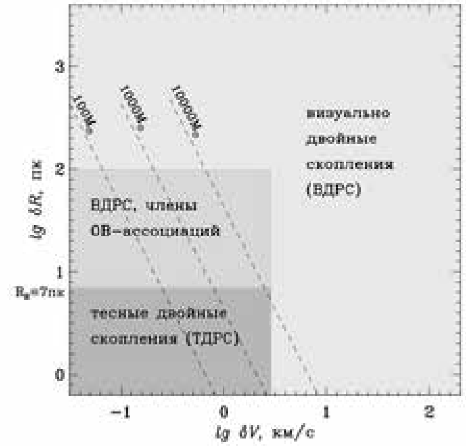


Рис. 9. Диаграмма для пар скоплений, позволяющая классифицировать типы двойственности РЗС. Подписаны области, внутри которых находятся пары РЗС различных типов. Наклонные прямые, показанные штрих-пунктиром, дают информацию о суммарной массе скоплений. Основными параметрами, позволяющими классифицировать пары скоплений, являются расстояние между скоплениями в паре *δR* и разность их пространственных скоростей *δV*

Заключение

Для изучения звездной структуры на начальном этапе применяются машинные методы обработки массовой информации. Например, в [43] использован метод (UPMASK) для определение членства звезд в скоплениях. Он подготовлен для использования фотометрии и пространственных положений, но может учитывать другие типы данных. Подход, используемый для оценки членства, основан на итерационном процессе, уменьшении размерности, алгоритме кластеризации и оценке плотности ядра РЗС. Другой пример – [44] использовали метод OCfinder для поиска 628 новых РЗС в Gaia EDR3 с использованием среды больших данных. Как первый шаг, OCfinder идентифицировал звездные статистические сверхплотности в пятимерном астрометрическом пространстве (положение, параллакс и собственные движения) с использованием алгоритма кластеризации DBSCAN. Затем эти сверхплотности классифицировались на случайные статистические сверхплотности и реальные физические РЗС с использованием глубокой искусственной нейронной сети, обученной на хорошо охарактеризованных диаграммах G, GBP – GRP цвет-величина. Далее применяются методы фильтрации по параметрам звезд, детально рассматривются: положение на небе, параллаксы, цвет, звездная величина и детальные модели распределения звездной плотности. Ясно, что эволюция скоплений тесно связана с пониманием природы звездных потоков, АКП и АКПЗ (АКП + звезды) копий. Можно выделить структуру данных, представленных в разных разделах астрофизики, сочетающихся на различных этапах эволюции РЗС, Табл. 1.

Табл. 1

Структура данных РЗС (n – ожидаемое число в Галактике, $n_{кат}$ –каталогизировано)

объект	состав			n	$n_{кат}$
ГМО				10^7	10^3
звезды		АКП		10^{11}	10^9
РЗС	звезды	АКП		10^8	10^4
двойные РЗС			РЗС	10^7	10^2
ОВ ассоциации	звезды	АКП	РЗС	10^5	10^3
звездные потоки	звезды	АКП	РЗС	10^8	10^3

Оценки числа РЗС, ОВ ассоциаций и звездных потоков в Табл. 1 сделаны, исходя из общей оценки числа звезд в Галактике. При этом учтены именно все РЗС, а не только выжившие.

В области Меча Ориона наблюдающиеся РЗС, по всей вероятности, связаны с газово-молекулярными облаками. По данным космического телескопа «Гершель» мы выделили скопления, связанные с пылевыми облаками (T = 15 – 35K) и с горячим (10000K) газом. Ранее построенная диаграмма позволяет выделить типы двойных скоплений. Все перечисленные в Таблице 1 объекты укладываются в предложенную схему эволюции скоплений, Рис. 2.

Нами изучена звездная система Плеяд по данным Gaia DR2, Рис.4 – Рис.6. Показаны особенности пространственной формы РЗС Плеяды и расположенного вместе с ними потока. Выделенный в [29] в окрестности РЗС звездный поток Рыб-Эридана генетически, вероятно, связан с Плеядами и не содержит другого скопления. Скорее всего, совместно они представляют остатки распавшейся ОВ ассоциации.

Результаты обработки изображений газовой структуры по данным КА «Гершель», Рис. 7 и Рис. 8 свидетельствуют о том, что различные газовые структуры по положению на небесной сфере совпадают с РЗС в области Меча Ориона. При этом выделяются сочетания РЗС различного возраста с соответствующей газовой структурой.

Наша работа позволяет рассчитывать, что на основе накопления данных и их интенсивной обработки классическая классификация РЗС может быть дополнена параметрами «АКП копья» и «суперкопья» скопления, характеристиками газа, связанного со скоплением и характеристиками двойственности РЗС. Также можно говорить о изучении происхождения звездных потоков, которые могут являться как галактическими, так и внегалактического происхождения из-за разрушения галактик-спутников нашей Галактики. Перспективно обнаружение и исследование экзопланет в РЗС различных типов. Таблица 1 показывает возможности применения интенсивной обработки накопляемых со временем данных о разных объектах, связанных с эволюцией РЗС. На сегодня данные каталога Gaia являются основой для эволюционной пространственно-кинематической классификации РЗС Галактики, которых на данный момент обнаружено около 10 тыс.

Благодарности. В этой работе использовались данные миссии Gaia Европейского космического агентства (ESA) (https://www.cosmos.esa.int/gaia), обработанные Консорциумом обработки и анализа данных Gaia (DPAC, https://www.cosmos.esa.int/ web/gaia/dpac/consortium). Финансирование DPAC было предоставлено национальными учреждениями, в частности учреждениями, участвующими в Многостороннем соглашении Gaia. Веб-сайт миссии Gaia: https://www.cosmos.esa.int/gaia. Веб-сайт архива Gaia: https://archives.esac.esa.int/gaia. В этом исследовании использовалась база данных SIMBAD (http://cds.u-strasbg.fr), работающая в CDS, Страсбург, Франция. Авторы благодарят рецензентов за полезные замечания.

Литература

1. *Gaia Collaboration* (2016), Prusti, T., de Bruijne, J. H. J., Brown, A. G. A., Vallenari, A., et al. The Gaia mission (provides a description of the Gaia mission including spacecraft, instruments, survey

and measurement principles, and operations) // Astronomy and Astrophysics. 2016. Volume 595. id.A1. 36 pp.

2. *Gaia Collaboration* (2022), Vallenari, A., et al. Gaia DR3: data release content and main properties // Astronomy and Astrophysics. 2022. in prep.
3. The James Webb Space Telescope and Herschel, https://esahubble.org/images/jwst_herschel/
4. *Cantat-Gaudin, T., Anders, F., Castro-Ginard, A., et al.* Painting a portrait of the Galactic disc with its stellar clusters // Astronomy and Astrophysics. 2020. Volume 640, p. 1.
5. *Pang, X., Tang, S-Y., Li, Y., et al.* 3D Morphology of Open Clusters in the Solar Neighborhood with Gaia EDR3 II: Hierarchical Star Formation Revealed by Spatial and Kinematic Substructures // preprint astro-ph/2204.06000. 2022.
6. *Baily, S.I.* A catalogue of bright clusters and nebulae // Annals Harvard College Observatory. 1908. Volume 60. pp. 199-230.
7. *Melotte, P.J.* A Catalogue of Star Clusters shown on Franklin-Adams Chart Plates // Memoirs of the Royal Astronomical Society Mem. 1915. Volume 60. pp. 175-186.
8. *Raab, S.* Research on open clusters // Meddelanden fran Lunds Astron. Obs. Series II. 1922. 28. 3-48.
9. *Trumpler, R.J.* Spectral Types in Open Clusters // Publications of the Astronomical Society of the Pacific. 1925. Volume 37. № 220. p. 307.
10. *Markarian, B.E.* On Classification of Open (Galactic) Star Clusters. Communications // Byurakan Astrophys. Obs. (ComBAO). 1950. Volume 5. p. 3-34.
11. *Kharchenko, N.V., Piskunov, A.E., Roeser, S., Schilbach, E., Scholz, R.-D.* Global survey of star clusters in the Milky Way. II. The catalogue of basic parameters // Astronomy and Astrophysics 2013. Volume 558. A53.
12. *Holmes, E.* Star Streams // Journal British Astron. Assoc. 1894. Volume 5. pp. 26-28.
13. *Proctor, R.* Preliminary Paper on Certain Drifting Motions of the Stars // Proc. Roy. Soc. London. 1869. 18. 169-171.
14. *Eddington, A.* Aberration, in relation to two star-streams // Monthly Notices of the Royal Astronomical Society. 1909. Volume 69. p. 57.
15. *Tutukov, A.* Early Stages of Dynamical Evolution of Star Cluster Models // Astronomy and Astrophysics. 1978. Volume 70. p. 57.
16. *Tutukov, A.V., Vereshchagin, S.V., Sizova, M.D.* Destruction of Galaxies as a Cause of the Appearance of Stellar Streams // Astronomy Reports. Volume 65. № 11. pp. 1085-1101 (2021). doi: 10.1134/S106377292111007X
17. *Tutukov, A, Smirnov, M.* Peripheral Structures of Planetary Systems // Solar System Research. 2004. Volume 38, № 4, pp. 279-287. doi: 10.1023/B:SOLS.0000037463.42719.71
18. *Vshivkov, V., Lazareva, G., Snytnikov, A. et al.* Hydrodynamical Code for Numerical Simulation of the Gas Components of Colliding Galaxies // The Astrophysical Journal Supplement Series. 2011. Volume 194 p. 47. doi: 10.1088/0067-0049/194/2/47
19. *Tanaka, K.* The Magellanic Stream and the Interacting Galaxies // Publications of the Astronomical Society of Japan. 1981. Volume 33. P. 247.
20. *Bournaud, F., Duc, P., Amram, P. et al.* Kinematics of tidal tails in interacting galaxies: Tidal dwarf galaxies and projection effects // Astronomy and Astrophysics. Volume. 2004. 425, p. 813. doi: 10.1051/0004-6361:20040394
21. *Grillmair, C.* Detection of a 60°- long Dwarf Galaxy Debris Stream // The Astrophysical Journal. 2006. Volume 645, p. 37. doi: 10.1086/505863
22. *Malhan, K., Yuen, Z., Ibata, R. et al.* Evidence of a dwarf galaxy stream populating the inner Milky Way Halo // astro-ph/2104.09523. 2021.
23. *Odenkirchen, M., Grebel, E.K., Dehnen, W., Rix, H-W. et al.* The Extended Tails of Palomar 5: A 10° Arc of Globular Cluster Tidal Debris // The Astronomical Journal. 2003. Volume 126, 5, pp. 2385-2407. doi: 10.1086/378601
24. *Grillmair, C.J., Freeman, K.C., Irwin, M., Quinn, P.J.* Globular Clusters with Tidal Tails: Deep Two-Color Star Counts // The Astronomical Journal. 1995. Volume 109, p. 2553. doi: 10.1086/117470
25. *Eggen, O.* The Pleiades and alpha Persei Clusters // The Astronomical Journal. 1998. Volume 116, № 4, pp. 1810-1815. doi: 10.1086/300533
26. *van Leeuwen, F.* Parallaxes and proper motions for 20 open clusters as based on the new Hipparcos catalogue // Astronomy and Astrophysics. 2009. Volume 497, p. 1. doi: 10.1051/0004-6361/200811382
27. *Lindegren, L., Hernández, J., Bombrun, A. et al.* Gaia Data Release 2. The astrometric solution // Astronomy and Astrophysics. Volume 616, A2. doi: 10.1051/0004-6361/201832727
28. *van Leeuwen, F., de Bruijne, J.H.J., Arenou, F. et al.* Gaia DR2 documentation. Gaia DR2 documentation, European Space Agency; Gaia Data Processing and Analysis Consortium. Online at https://gea.esac.esa.int/archive/documentation/GDR2/. 2018.
29. *Röser, S., Schilbach, E.* A census of the nearby Pisces-Eridanus stellar stream. Commonalities

with and disparities from the Pleiades// Astronomy and Astrophysics. 2020. Volume 638, A9. doi: 10.1051/0004-6361/202037691

30. *Cantat-Gaudin, T., Jordi, C., Vallenari, A., Bragaglia, A., et al.* A Gaia DR2 view of the open cluster population in the Milky Way // Astronomy and Astrophysics. 2018. Volume 618. id.A93, 16 pp.
31. *Galli, P.A.B., Moraux, E., Bouy, H., Bouvier, J., Olivares, J., Teixeira, R.* A revised moving cluster distance to the Pleiades open cluster // Astronomy and Astrophysics. 2017. Volume 598. id.A48, 22 pp.
32. *Gaia Collaboration, Brown, A.G.A., Vallenari, A., Prusti, T., et al.* Gaia Data Release 2. Summary of the contents and survey properties // Astronomy and Astrophysics. 2018. Volume 616. id.A1, 22 pp.
33. *Röser, Siegfried, Schilbach, Elena, Goldman, Bertrand*: Hyades tidal tails revealed by Gaia DR2 // Astronomy and Astrophysics. 2019. Volume 621. id. L2, 5 pp.
34. *Lodieu, N., Pérez-Garrido, A., Smart, R.L., Silvotti, R.* A 5D view of the α Per, Pleiades, and Praesepe clusters // Astronomy and Astrophysics. 2019. Volume 628. id.A66, 26 pp.
35. *Chumak, Ya.O., Rastorguev, A.S.* Analysis of the structure and dynamics of the stellar tails of open star clusters // Astronomy Letters. 2006. Volume 32, № 3, p.157-165.
36. *Johnson, Dean R.H., Soderblom, David R.* Calculating Galactic Space Velocities and Their Uncertainties, with an Application to the Ursa Major Group // The Astronomical Journal. 1987. Volume 93, p.864.
37. VTSS. The Virginia Tech Spectral-Line Survey Arcminute-Resolution Digital Images of Interstellar H-alpha and [SII] Emission http://www1.phys.vt.edu/~halpha/.
38. *Battersby, C., Bally, J., A. Ginsburg, A. et al.* Characterizing precursors to stellar clusters with Herschel // Astronomy and Astrophysics. 2011. 2011. Volume 535. A128. https://doi.org/10.1051/0004-6361/201116559
39. *Launhardt, R., Stutz, A.M., Schmiedeke, A., et al.* The Earliest Phases of Star Formation (EPoS): a Herschel // Astronomy and Astrophysics. 2013. Volume 551. A98. https://doi.org/10.1051/0004-6361/201220477
40. *Mallick, K.K., Ojha, D.K., Tamura, M., et al.* Study of morphology and stellar content of the Galactic H II region IRAS 16148−5011// Monthly Notices of the Royal Astronomical Society. 2015. Volume 447. № 3. Pp. 2307–2321. https://doi.org/10.1093/mnras/stu2584
41. *Kauffmann, J., Bertoldi, F., Bourke, T.L., Evans, II, N.J., and Lee, C.W.* MAMBO mapping of Spitzer c2d small clouds and cores // Astronomy and Astrophysics. 2008. Volume 487. №. 3. Pp. 993 – 1017 pp. https://doi.org/10.1051/0004-6361:200809481
42. *Vereshchagin, S.V., Tutukov, A.V., Chupina, N.V., Postnikova, E.S. and Sizova, M.D.* Binary Clusters: Theory and Observations // Astronomy Reports. 2022. Volume 66. № 5. pp. 361–386. doi: 10.1134/S1063772922060063
43. UPMASK: Unsupervised Photometric Membership Assignment in Stellar Clusters, Linking: https://CRAN.R-project.org/package=UPMASK
44. *Castro-Ginard, A., Jordi, C., Luri, X., Cantat-Gaudin, T., et al.* Hunting for open clusters in Gaia EDR3: 628 new open clusters found with OCfinder // Astronomy and Astrophysics. 2022. Volume 661. A118.

Постникова Екатерина Сергеевна. Федеральное государственное учреждение «Институт астрономии Российской академии наук», г. Москва, Россия. Младший научный сотрудник, кандидат физико-математических наук. Количество печатных работ: 12. Область научных интересов: звездная астрономия, рассеянные звездные скопления, информационные системы. E-mail: es_p@list.ru

Рябухина Ольга Леонидовна. Федеральное государственное учреждение «Институт астрономии Российской академии наук», г. Москва, Россия. Младший научный сотрудник. Количество печатных работ: 10. Область научных интересов: структура газовых облаков, информационные системы. E-mail: ryabukhina@inasan.ru

Тутуков Александр Васильевич. Федеральное государственное учреждение «Институт астрономии Российской академии наук», г. Москва, Россия. Главный научный сотрудник, доктор физико-математических наук. Количество печатных работ: более 400. Область научных интересов: астрофизика, звездная астрономия, информационные системы. E-mail: atutukov@inasan.ru

Верещагин Сергей Викторович. Федеральное государственное учреждение «Институт астрономии Российской академии наук», г. Москва, Россия. Старший научный сотрудник, кандидат физико-математических наук. Количество печатных работ: 100. Область научных интересов: звездная астрономия, рассеянные звездные скопления, информационные системы. E-mail: svvs@ya.ru (ответственный за переписку).

Чупина Наталия Викторовна. Федеральное государственное учреждение «Институт астрономии Российской академии наук», г. Москва, Россия. Старший научный сотрудник, кандидат физико-математических наук. Количество печатных работ: 40. Область научных интересов: звездная астрономия, рассеянные звездные скопления, информационные системы. E-mail: chupina@inasan.ru

Демидов Андрей Павлович. Align Technology, Inc., г. Москва, Россия. Инженер по разработке ПО. Область научных интересов: информационные системы. E-mail: the-admax@yandex.ru

The Structure and Evolution of Open Star Clusters: Theory and Observations Based on Gaia Data

E.S. Postnikova[I], O. L. Ryabukhina[I], A. V. Tutukov[I], S.V. Vereshchagin[I], N.V. Chupina[I], A.P. Demidov[II]

[I] Institute of Astronomy Russian Academy of Sciences (INASAN), Moscow, Russia.

[II] Align Technology, Inc., Moscow, Russia.

Abstract. The structure and evolution of open star clusters (OSCs) are considered using the Pleiades OSCs and the OSC group in the Orion Sword region as examples. The stars were selected according to the Gaia data. The relationship between the Orion Sword clusters and molecular clouds is traced according to the data of the Herschel spacecraft. The place of the considered objects in the general scheme of evolution compiled by us earlier is shown. It is concluded that there is an urgent need to expand the OSC classification. The considered Pleiades star system showed the presence of an extensive stellar halo. The stellar stream Pisces - Eridanus found in the vicinity of the Pleiades is probably genetically related to the Pleiades and, together with it, represents the remnants of the disintegrated OB association. In the Orion Sword region, the observed young OSCs are most likely associated with molecular clouds. Young clusters stand out associated with dust (15 - 35 K) and hot (10000 K) gas. Data on OSCs are rapidly replenishing, and the number of OSCs is growing due to their detection in the Gaia surveys. Analysis in this area can be iterated and extended over time with proven methodologies to fit data management concepts in data-intensive areas.

Keywords: *operational information support, open star clusters, Pleiades, Orion Sword region, analytics, data management*

DOI: 10.14357/20790279230114

References

1. *Gaia Collaboration* (2016), Prusti, T., de Bruijne, J. H. J., Brown, A. G. A., Vallenari, A., et al. The Gaia mission (provides a description of the Gaia mission including spacecraft, instruments, survey and measurement principles, and operations) // Astronomy and Astrophysics. 2016. Volume 595. id.A1. 36 pp.
2. *Gaia Collaboration* (2022), Vallenari, A., et al. Gaia DR3: data release content and main properties // Astronomy and Astrophysics. 2022. in prep.
3. The James Webb Space Telescope and Herschel, https://esahubble.org/images/jwst_herschel/
4. *Cantat-Gaudin, T., Anders, F., Castro-Ginard, A., et al.* Painting a portrait of the Galactic disc with its stellar clusters // Astronomy and Astrophysics. 2020. Volume 640, p. 1.
5. *Pang, X., Tang, S-Y., Li, Y., et al.* 3D Morphology of Open Clusters in the Solar Neighborhood with Gaia EDR3 II: Hierarchical Star Formation Revealed by Spatial and Kinematic Substructures // preprint astro-ph/2204.06000. 2022.
6. *Baily, S.I.* A catalogue of bright clusters and nebulae // Annals Harvard College Observatory. 1908. Volume 60. pp. 199-230.
7. *Melotte, P.J.* A Catalogue of Star Clusters shown on Franklin-Adams Chart Plates // Memoirs of the Royal Astronomical Society Mem. 1915. Volume 60. pp. 175-186.
8. *Raab, S.* Research on open clusters // Meddelanden fran Lunds Astron. Obs. Series II. 1922. 28. 3-48.
9. *Trumpler, R.J.* Spectral Types in Open Clusters // Publications of the Astronomical Society of the Pacific. 1925. Volume 37. № 220. p. 307.
10. *Markarian, B.E.* On Classification of Open (Galactic) Star Clusters. Communications // Byurakan Astrophys. Obs. (ComBAO). 1950. Volume 5. p. 3-34.
11. *Kharchenko, N.V., Piskunov, A.E., Roeser, S., Schilbach, E., Scholz, R.-D.* Global survey of star

clusters in the Milky Way. II. The catalogue of basic parameters // Astronomy and Astrophysics 2013. Volume 558. A53.

12. *Holmes, E.* Star Streams // Journal British Astron. Assoc. 1894. Volume 5. pp. 26-28.
13. *Proctor, R.* Preliminary Paper on Certain Drifting Motions of the Stars // Proc. Roy. Soc. London. 1869. 18. 169-171.
14. *Eddington, A.* Aberration, in relation to two star-streams // Monthly Notices of the Royal Astronomical Society. 1909. Volume 69. p. 57.
15. *Tutukov, A.* Early Stages of Dynamical Evolution of Star Cluster Models // Astronomy and Astrophysics. 1978. Volume 70. p. 57.
16. *Tutukov, A.V., Vereshchagin, S.V., Sizova, M.D.* Destruction of Galaxies as a Cause of the Appearance of Stellar Streams // Astronomy Reports. Volume 65. № 11. pp. 1085-1101 (2021). doi: 10.1134/S106377292111007X
17. *Tutukov, A, Smirnov, M.* Peripheral Structures of Planetary Systems // Solar System Research. 2004. Volume 38, № 4, pp. 279-287. doi: 10.1023/B:SOLS.0000037463.42719.71
18. *Vshivkov, V., Lazareva, G., Snytnikov, A. et al.* Hydrodynamical Code for Numerical Simulation of the Gas Components of Colliding Galaxies // The Astrophysical Journal Supplement Series. 2011. Volume 194 p. 47. doi: 10.1088/0067-0049/194/2/47
19. *Tanaka, K.* The Magellanic Stream and the Interacting Galaxies // Publications of the Astronomical Society of Japan. 1981. Volume 33. P. 247.
20. *Bournaud, F., Duc, P., Amram, P. et al.* Kinematics of tidal tails in interacting galaxies: Tidal dwarf galaxies and projection effects // Astronomy and Astrophysics. Volume. 2004. 425, p. 813. doi: 10.1051/0004-6361:20040394
21. *Grillmair, C.* Detection of a 60°- long Dwarf Galaxy Debris Stream // The Astrophysical Journal. 2006. Volume 645, p. 37. doi: 10.1086/505863
22. *Malhan, K., Yuen, Z., Ibata, R. et al.* Evidence of a dwarf galaxy stream populating the inner Milky Way Halo // astro-ph/2104.09523. 2021.
23. *Odenkirchen, M., Grebel, E.K., Dehnen, W., Rix, H-W. et al.* The Extended Tails of Palomar 5: A 10° Arc of Globular Cluster Tidal Debris // The Astronomical Journal. 2003. Volume 126, 5, pp. 2385-2407. doi: 10.1086/378601
24. *Grillmair, C.J., Freeman, K.C., Irwin, M., Quinn, P.J.* Globular Clusters with Tidal Tails: Deep Two-Color Star Counts // The Astronomical Journal. 1995. Volume 109, p. 2553. doi: 10.1086/117470
25. *Eggen, O.* The Pleiades and alpha Persei Clusters // The Astronomical Journal. 1998. Volume 116, № 4, pp. 1810-1815. doi: 10.1086/300533
26. *van Leeuwen, F.* Parallaxes and proper motions for 20 open clusters as based on the new Hipparcos catalogue // Astronomy and Astrophysics. 2009. Volume 497, p. 1. doi: 10.1051/0004-6361/200811382
27. *Lindegren, L., Hernández, J., Bombrun, A. et al.* Gaia Data Release 2. The astrometric solution // Astronomy and Astrophysics. Volume 616, A2. doi: 10.1051/0004-6361/201832727
28. *van Leeuwen, F., de Bruijne, J.H.J., Arenou, F. et al.* Gaia DR2 documentation. Gaia DR2 documentation, European Space Agency; Gaia Data Processing and Analysis Consortium. Online at https://gea.esac.esa.int/archive/documentation/GDR2/. 2018.
29. *Röser, S., Schilbach, E.* A census of the nearby Pisces-Eridanus stellar stream. Commonalities with and disparities from the Pleiades// Astronomy and Astrophysics. 2020. Volume 638, A9. doi: 10.1051/0004-6361/202037691
30. *Cantat-Gaudin, T., Jordi, C., Vallenari, A., Bragaglia, A., et al.* A Gaia DR2 view of the open cluster population in the Milky Way // Astronomy and Astrophysics. 2018. Volume 618. id.A93, 16 pp.
31. *Galli, P.A.B., Moraux, E., Bouy, H., Bouvier, J., Olivares, J., Teixeira, R.* A revised moving cluster distance to the Pleiades open cluster // Astronomy and Astrophysics. 2017. Volume 598. id.A48, 22 pp.
32. *Gaia Collaboration, Brown, A.G.A., Vallenari, A., Prusti, T., et al.* Gaia Data Release 2. Summary of the contents and survey properties // Astronomy and Astrophysics. 2018. Volume 616. id.A1, 22 pp.
33. *Röser, Siegfried, Schilbach, Elena, Goldman, Bertrand*: Hyades tidal tails revealed by Gaia DR2 // Astronomy and Astrophysics. 2019. Volume 621. id. L2, 5 pp.
34. *Lodieu, N., Pérez-Garrido, A., Smart, R.L., Silvotti, R.* A 5D view of the α Per, Pleiades, and Praesepe clusters // Astronomy and Astrophysics. 2019. Volume 628. id.A66, 26 pp.
35. *Chumak, Ya.O., Rastorguev, A.S.* Analysis of the structure and dynamics of the stellar tails of open star clusters // Astronomy Letters. 2006. Volume 32, № 3, p.157-165.
36. *Johnson, Dean R.H., Soderblom, David R.* Calculating Galactic Space Velocities and Their Uncertainties, with an Application to the Ursa Major Group // The Astronomical Journal. 1987. Volume 93, p.864.
37. VTSS. The Virginia Tech Spectral-Line Survey Arcminute-Resolution Digital Images of

Interstellar H-alpha and [SII] Emission http://www1.phys.vt.edu/~halpha/.

38. *Battersby, C., Bally, J., A. Ginsburg, A. et al.* Characterizing precursors to stellar clusters with Herschel // Astronomy and Astrophysics. 2011. 2011. Volume 535. A128. https://doi.org/10.1051/0004-6361/201116559
39. *Launhardt, R., Stutz, A.M., Schmiedeke, A., et al.* The Earliest Phases of Star Formation (EPoS): a Herschel // Astronomy and Astrophysics. 2013. Volume 551. A98. https://doi.org/10.1051/0004-6361/201220477
40. *Mallick, K.K., Ojha, D.K., Tamura, M., et al.* Study of morphology and stellar content of the Galactic H II region IRAS 16148−5011// Monthly Notices of the Royal Astronomical Society. 2015. Volume 447. № 3. Pp. 2307–2321. https://doi.org/10.1093/mnras/stu2584
41. *Kauffmann, J., Bertoldi, F., Bourke, T.L., Evans, II, N.J., and Lee, C.W.* MAMBO mapping of Spitzer c2d small clouds and cores // Astronomy and Astrophysics. 2008. Volume 487. №. 3. Pp. 993 – 1017 pp. https://doi.org/10.1051/0004-6361:200809481
42. *Vereshchagin, S.V., Tutukov, A.V., Chupina, N.V., Postnikova, E.S. and Sizova, M.D.* Binary Clusters: Theory and Observations // Astronomy Reports. 2022. Volume 66. № 5. pp. 361–386. doi: 10.1134/S1063772922060063
43. UPMASK: Unsupervised Photometric Membership Assignment in Stellar Clusters, Linking: https://CRAN.R-project.org/package=UPMASK
44. *Castro-Ginard, A., Jordi, C., Luri, X., Cantat-Gaudin, T., et al.* Hunting for open clusters in Gaia EDR3: 628 new open clusters found with OCfinder // Astronomy and Astrophysics. 2022. Volume 661. A118.

Postnikova E.S. PhD. Institute of Astronomy of the Russian Academy of Sciences, 48 Pyatnitskaya st. 119017, Moscow, Russia. E-mail: es_p@list.ru

Ryabukhina O.L. Researcher. Institute of Astronomy Russian Academy of Sciences (INASAN), 48 Pyatnitskaya st., 119017, Moscow, Russia. E-mail: ryabukhina@inasan.ru

Tutukov A.V. Professor, Institute of Astronomy of the Russian Academy of Sciences, 48 Pyatnitskaya st. 119017, Moscow, Russia. E-mail: atutukov@inasan.ru

Vereshchagin S.V. Senior Researcher. Institute of Astronomy Russian Academy of Sciences (INASAN), 48 Pyatnitskaya st., 119017, Moscow, Russia. E-mail: svvs@ya.ru (corresponding author)

Chupina N.V. Senior Researcher. Institute of Astronomy Russian Academy of Sciences (INASAN), 48 Pyatnitskaya st., 119017, Moscow, Russia. E-mail: chupina@inasan.ru

Demidov A.P. Software development engineer. Align Technology, Inc., 117105, Varshavskoye shosse, 9, building 1b, Moscow, Russia. E-mail: the-admax@yandex.ru

Astronomical observation planner*

V.Yu. Kim[I,II,III], I.M. Izmailova[I]

[I] Fesenkov Astrophysical Institute, Kazakhstan, Almaty
[II] Higher Shool of Economics, Moscow, Russia
[III] Pulkovo observatory, Saint-Petersburg, Russia

Abstract. One of the tasks of robotization of astronomical observations is the creation of programs for the optimal distribution of time depending on the position of the Sun (efficient use of twilight time), the position and phases of the Moon. An important requirement for this program is the autonomy of its work, independent of external Internet resources. To solve this problem, an autonomous astronomical calendar was developed that makes it possible to estimate the time of sunrise and sunset, the moon (as well as its phases), the onset and end of twilight. This subroutine is the first step in automating the planning of astronomical observations. The next important step is to develop software that will be able to plan observations in an optimal way. The targets for observations are discussed, for these purposes the necessary initial parameters are indicated, which make it possible to form a schedule of observations at telescopes in an automatic mode.

Keywords: *astronomical observation, automation, telescope.*
DOI: 10.14357/20790279230115

Introduction

In 2021, there began the implementation of program to create a Virtual Observatory [1] based on the Fesenkov Astrophysical Institute (FAI). Establishment of a national Virtual Observatory designed to enhance the capabilities of astronomical research and provide a service to external users. Development of methods for processing, storing and analyzing Big Data in astronomy for studying objects of near and far space. Implementing the program will increase the efficiency of observational and numerical studies at the FAI. For external users, a digital portal will be created through which they will be able to take advantage of the results of all innovations under the program, in particular, a) apply for automated observations and the use of computing resources, both for numerical modelling and for processing and analyzing the received data; b) to access the available observational data and the results of astrophysical computer simulations. The process of using observational and computing resources will be fully automated.

One of the program's objectives is to automate the observational process of astronomical sources at the institute's telescopes located at the Assy-Turgen observatory and other observation sites. An important software component in the system being created is the so-called observation planner (OP). This component will allow sorting and creating the most optimal telescope operation plan for a particular night from the resulting list of objects and their coordinates and observation time. The program should take into account the moments of sunrise-sunset, the duration of morning and evening twilight, the moonrise-set and its phases, and the influence of the gradient from the lunar sky illumination. All program calculations should be carried out without external Internet resources.

The interpreted high-level language PHP was chosen as the primary programming language for creating the planner, as it is the most suitable for integrating software into the Internet portal of the Virtual Observatory and for user interaction.

The development of OP is carried out in several parts. In the first part, software was created for numerical calculations of sidereal time, the position of the Sun and Moon, and the calculation of their moments of rise and set and twilight for a specific date and place of observation. The results of the first stage are presented in the form of an Astronomical calendar and posted on the portal of the Astrophysical Institute [2] (Fig. 1). Where any user can get the specified data for any point on the earth's surface (at the specified geographical coordinates).

In the second part of creating the OP, algorithms will be implemented to calculate the brightness of the

* This research is funded by the Science Committee of the Ministry of Education and Science of the Republic of Kazakhstan (Grant No. BR10965141).

sky background depending on the phase and the angular distance between the Moon and the observed source. It will make it possible to exclude from observations dim objects that are close (by angular distance) to the bright Moon. Also, the source sorting algorithm will be implemented at this stage.

1. Basic parameters calculation and algorithms.

To calculate the moments of rise and set of the Sun and the Moon, the following algorithmic steps are implemented (Fig.2):

Step 1. Calculating the equatorial coordinates of the Sun and Moon at a specific point in time:

1.1. Calculation of the position of an object in its orbit at a specified point in time. Since the Sun and Moon are not point sources, here, the coordinates of an object are the coordinates of the center of the solar (or lunar) disk.

1.2. Translation of own orbital coordinates into ecliptic coordinates.

1.3. Calculation of corrections for the change in the inclination of the equatorial plane to the ecliptic plane at a specified point in time. As well as corrections for precession and nutation of the Earth's axis.

1.4. Translation of the object's ecliptic coordinates into equatorial coordinates, considering the corrections specified in 4. At the output, we obtain geocentric equatorial coordinates at the specified epoch.

Step 2. Transformation of the geocentric coordinates of the Sun and the Moon into topocentric coordinates (corresponding to the place of observation):

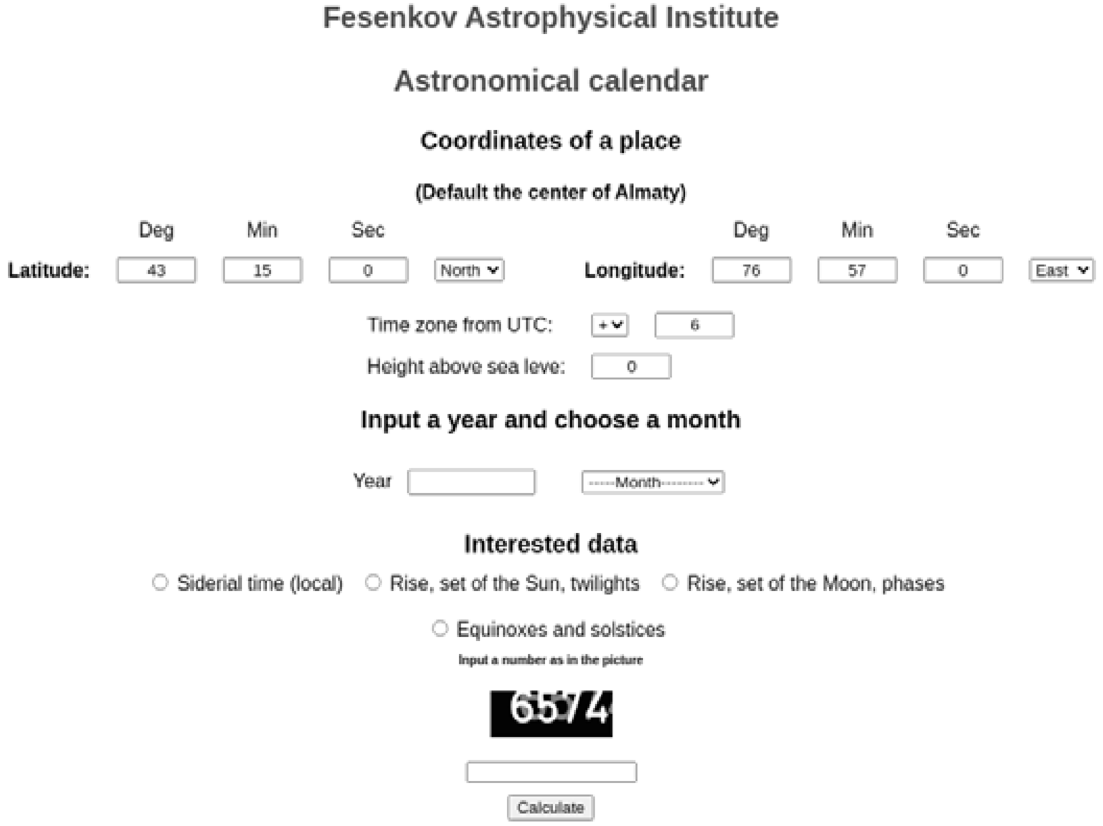
2.1. Calculation of the geocentric parallax of an object, taking into account corrections for latitude and height (above sea level) of the observer's position.

2.2. Transformation with the help of corrections 2.1. from geocentric to topocentric equatorial coordinates.

Step 3. To calculate the moments of rise and set of the Sun and the Moon, the algorithm of successive approximations was used.

3.1. Calculation of topocentric coordinates at the beginning of the day of interest (see steps 1–2).

3.2. Through formulas for transforming coordinates from equatorial to horizontal. We find the local



Fesenkov Astrophysical Institute

Astronomical calendar

Coordinates of a place

(Default the center of Almaty)

	Deg	Min	Sec			Deg	Min	Sec	
Latitude:	43	15	0	North	Longitude:	76	57	0	East

Time zone from UTC: + 6

Height above sea leve: 0

Input a year and choose a month

Year

Interested data

○ Siderial time (local) ○ Rise, set of the Sun, twilights ○ Rise, set of the Moon, phases

○ Equinoxes and solstices

Input a number as in the picture

Fig. 1. Interface of developed on-line Astronomical calendar. Available on the link: https://fai.kz/calendar/calendar_eng.php

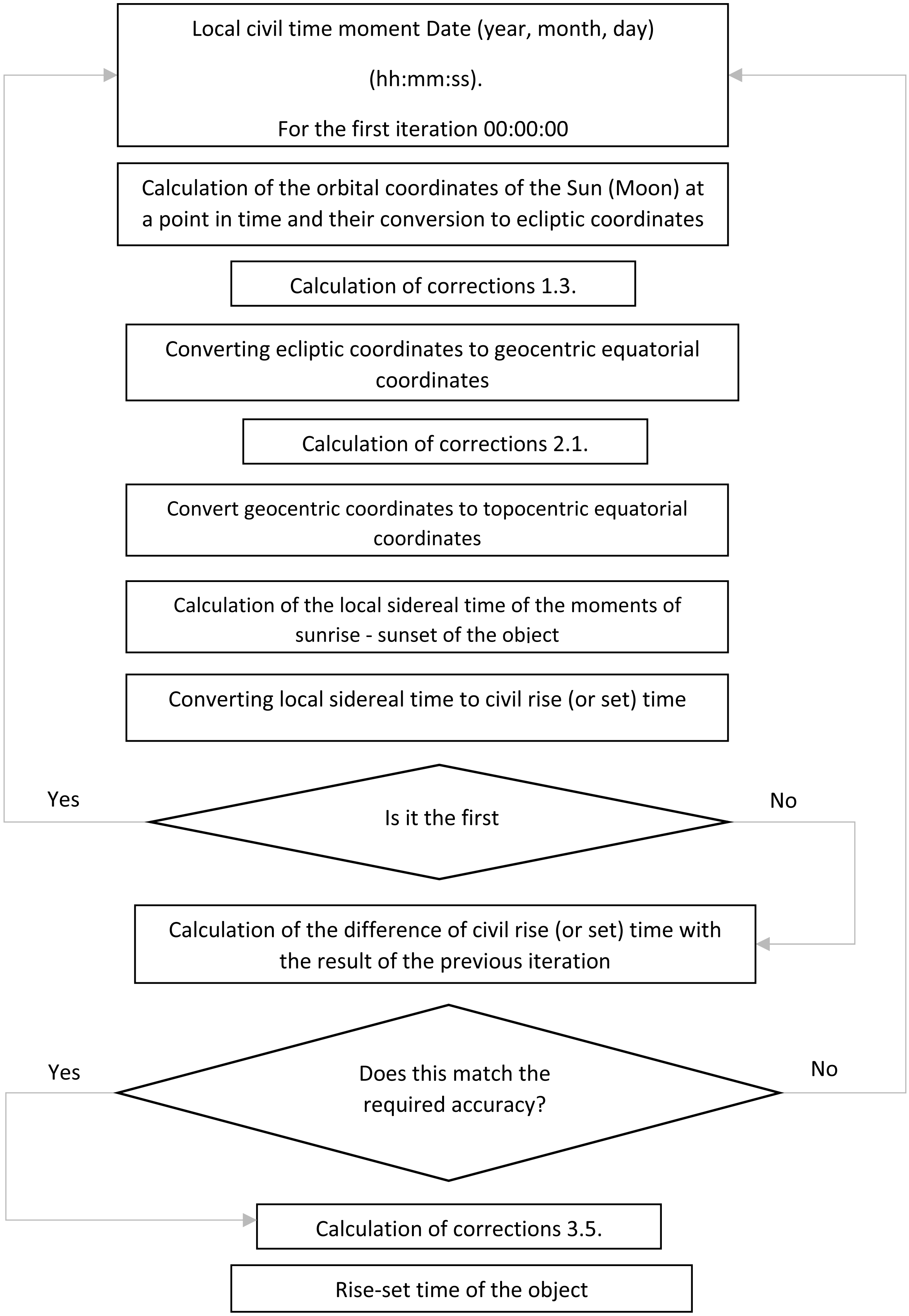

Local civil time moment Date (year, month, day)
(hh:mm:ss).
For the first iteration 00:00:00
Calculation of the orbital coordinates of the Sun (Moon) at a point in time and their conversion to ecliptic coordinates
Calculation of corrections 1.3.
Converting ecliptic coordinates to geocentric equatorial coordinates
Calculation of corrections 2.1.
Convert geocentric coordinates to topocentric equatorial coordinates
Calculation of the local sidereal time of the moments of sunrise - sunset of the object
Converting local sidereal time to civil rise (or set) time
Yes
Is it the first
No
Calculation of the difference of civil rise (or set) time with the result of the previous iteration
Yes
Does this match the required accuracy?
No
Calculation of corrections 3.5.
Rise-set time of the object

Fig. 2. Block diagram for rise-set time calculation for the Sun (the Moon)

sidereal time of the moments of rise and set of the object. That is the fulfilment of the condition that the object's height above the horizon of the place of observation is equal to zero (h = 0).

3.3. Let's convert the local sidereal time of the object's rise and set into local civil time.

Note: The coordinates of the Sun and Moon change significantly even at short intervals due to their relative proximity to the Earth (unlike the coordinates of distant stars). It means that in step 3.3 we will only get an approximate rise/set time. Since, upon reaching time 3.3, the Sun and the Moon will have time to move in their orbit. So, the next step is to repeat some of the points above.

3.4. Let's go to step 3.1. - we calculate the topocentric coordinates again, but for the time points obtained in step 3.3. Then we go to step 3.2. and get a new approximation of the local sidereal time for the moments of rise and set of the object. We again convert to civil time (step 3.3.) and get the updated civil time of rise and set moments. We repeat these iterations until we reach the desired accuracy.

3.5. We calculate corrections for the refraction of the Earth's atmosphere near the horizon and corrections for the spatial disk of the Sun (Moon). Since the moments of rise (set) correspond to the time when the upper (lower) edge of the disk touches the horizon plane. We make corrections to the moments of sunrise and sunset obtained in step 3.4. We get the local civil time of rise-set (Moon).

To calculate the position of the Sun, a numerical solution of the Kepler equation in the two-body problem is used. In this case, to simplify the calculations, we use the relativity of motion, considering the Earth to be stationary and located in the focus of the ellipse and the Sun moving along this ellipse. With known parameters (positions) at a certain point in time (in this case, at the beginning of the epoch of 2000), you can get the position of the Sun at any point in time in the following way:

1) Having the initial positions of the Sun, we calculate the so-called *average anomaly*.
2) By numerically solving the Kepler equation by iteration, we obtain the value of the eccentric anomaly.
3) Through the formula connecting the eccentric and true anomaly, we find the latter, which corresponds to the orbital longitude of the Sun. Here we do not consider the influence of the Moon, planets, etc., since these things can be neglected to calculate the moments of sunrise and sunset.

Calculating the position of the Moon is a rather difficult task due to the significant influence (in addition to the Earth) of the Sun and nearby planets. An algorithm based on Brown's analytical lunar theory was implemented for these calculations. The following important corrections and additions are taken into account:

1. *The equation of the center*. This correction takes into account orbital ellipticity.
2. *Evection*. This lunar inequality considers influence of the Sun on a shape of the lunar orbit.
3. *Variation*. This addition takes into account periodical processes of speed-up or slow-down of the Moon on its orbit due to the Sun.
4. *The annual equation*. This correction considers small periodical changes in lunar orbit, which has maximum in perigee and minimum in apogee.
5. *Parallactic inequality*. It is small correction considering solar parallax.
6. *Corrections for the influence of Venus, Mars and Jupiter, etc.*

To calculate the moments of moonrise and moonset, the horizontal equatorial parallax was also taken into account since the Moon is close to the Earth and moves in an elliptical orbit. As a result, its angular size changes when observed from the Earth.

Comparison of the moments of rises and sets of the Sun and the Moon obtained during the implementation of these algorithms with the data of the Astronomical Yearbooks for 1989, 2012, 2022 [5-7] give a time discrepancy of no more than 30 seconds, which indicates a sufficiently high accuracy.

2. Sorting for objects

The program (Fig. 3) accepts as input a file with a list of objects containing the following information (about each object):

1) Object name
2) Right Ascension (RA)
3) Declination (DEC)
4) Magnitude in V-filter
5) Exposure time of a single snapshot
6) Amount of snapshots

In the first stage, the program rejects objects with incorrectly entered parameters. For example, when the data on the number of frames contains alphabetic characters instead of numbers or when the coordinates (DEC) indicate that the object cannot be observed at this latitude. At the same time, an error log is formed, where similar objects are written with a description of errors.

In the second stage, calculations are carried out according to the time of the culmination of objects and then a preliminary sorting by this parameter. Objects with an earlier culmination time will be observed earlier because they will have earlier set (descending)

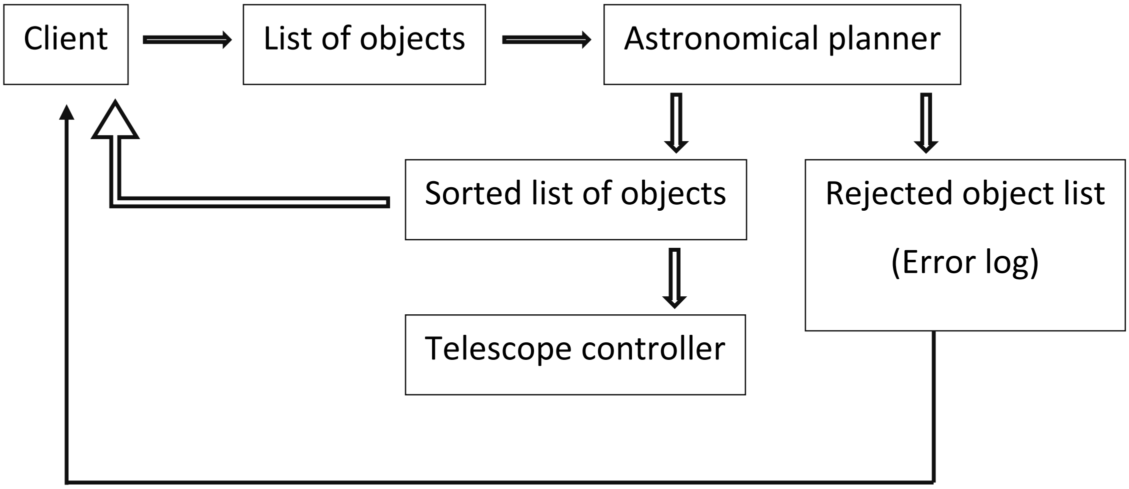


Fig. 3. Scheme of interaction between a client (user), an astronomical planner and telescope

time. For small observation fields (when all objects are situated in one area with each side does not exceed 3 degrees) an algorithm of the nearest neighbor is realized. This method is based on measurement of distances from a one object to others and finding the closest neighbor for next observation.

In the third stage, the moments of sunrise and sunset are calculated, as well as three types of twilight (civil, navigational and astronomical) for the current and next day. Furthermore, the boundaries of the astronomical night are determined, indicating the beginning and end of observations.

In the fourth stage, the heights of objects are calculated at the beginning of the astronomical night. If the objects are descending and their height at the start of observations is below 20 degrees, then they are rejected.

In the fifth stage, the moonrise-set times and its phases are calculated, and the angular distances between the moon and the observed objects are calculated with estimates of the sky background illumination. At this stage, objects whose magnitudes are weaker than the sky background are rejected. For other objects in the list, their angular distance to the center of the lunar disk is displayed. The output is a sorted list of objects and an error log, which is then taken as input by the telescope control program.

3. Other astronomical planners

Nowdays, there are many similar options for astronomical observation planners. The most famous: *AstroPlanner* [8] and *Astro Plan* [9]. The first one works only under MacOS and Windows operating systems. Also, this program is paid and closed source, which does not allow it to be integrated into the Virtual Observatory environment. An alternative free cross-platform program Astro Plan is developed in Python open-source. However, this program (Astro Plan) has some disadvantages:

1. The calculated times of sunrise and sunset differ by more than 3 minutes from the data of the Astronomical Yearbooks. For moonrise and moonset times, discrepancies can be more than 12 minutes. It can be critical for observations of dim objects.
2. Astro Plan does not realize a sorting algorithm by method of nearest neighbor, which can be useful for observations of small fields.
3. For using Astro Plan it is needed Python environment. There is no any ready-to-use GUI and it is difficult to integrate into web-interface that all users would be able to use it.

Conclusions

At the moment, the first part of creating an observation planner has been fully completed – a block for calculating the position of the Sun and Moon, as well as the moments of their rise and set and the time of twilight, working in offline mode. This block also was integrated in web-interface and available for all as an astronomical calendar.

The second block of the OP (a sorter of observation objects) is in the process of completion. Now it is ready to use OP realizing an algorithm of the nearest neighbor. The completed OP program will make it possible to optimally distribute the operating time of astronomical instruments without human intervention. This software product will be an important element, without which full automation of the observational process of astronomical objects is impossible.

References

1. https://fai.kz/projects/virtobs
2. https://fai.kz/calendar/calendar_eng.php
3. Meeus J. *Astronomical algorithms*, Richmond: Willmann-Bell, 2-nd edition, 1998
4. Duffett-Smith P. & Zwart J. *Practical Astronomy with your Calculator or Spreadsheet*, Cambridge University Press, 2011
5. *Astronomical Yearbook – 2022*, St.-Petersburg, The Institute of Applied Astronomy of RAS
6. *Astronomical Yearbook – 2012*, St.-Petersburg, The Institute of Applied Astronomy of RAS
7. *Astronomical Yearbook – 1989*, Leningrad, The Institute of Theoretical Astronomy
8. http://www.astroplanner.net
9. https://buildmedia.readthedocs.org/media/pdf/astroplan/latest/astroplan.pdf

V.Yu. Kim. PhD**,** Fesenkov Astrophysical Institute (FAI), Observatoriya street 23, Almaty, 050020, Kazakhstan, e-mail: kim@fai.kz (corresponding author)

I.M. Izmailova Fesenkov Astrophysical Institute (FAI), Observatoriya street 23, Almaty, 050020, Kazakhstan, e-mail: izmailova@fai.kz

A logical model for integration of heterogeneous experimental data in soil research

N.A. Vasilyeva, A.A. Vladimirov, T.A. Vasiliev

FGBNU Federal Research Center "V.V. Dokuchaev Soil Science Institute", Moscow, Russia

Abstract. The undoubted challenge for science is the extraction of knowledge from fast growing heterogeneous datasets. Particularly, details of experimental setups are insufficiently formalized and cannot be easily inserted into databases. Thus, there is a problem of using these details in the process of data integration and meta-analyses. For this purpose, we developed a scheme of formalization for object descriptions with its origination, protocols for field and laboratory measurements (including instruments and experimental conditions). It allows the integration of larger amounts of data accounting for its specifics of acquisition, for example, by applying adjustments, assigning weights to data sources (based on its reliability, method precision and experimental uncertainty) or directly accounting for experimental conditions in models. This formalization is currently used to develop an electronic laboratory journal for soil research, intended for detailed description of a conducted or planned experiment. The study aims to: increase the re-producibility of scientific research results; allow automatic data processing and error detection, and most importantly; effective soil data mining for decision support systems.

Keywords: *heterogeneous data sources, formalization, standards integration, soils, reproducibility.*

DOI: 10.14357/20790279230116

Introduction

An undoubted challenge for soil science is extracting knowledge and relevant information from the ever-growing, diverse and complex soil data sets [1]. For parameterization and validation of predictive models in analytical systems, it is required to identify the most complete set of relevant data in the database. While the relevance of data is determined by the details of their acquisition, i.e., the setting and conditions of the experiment, these are not formalized enough to be taken into account when choosing data. "Data acquisition" in the current study includes field descriptions of soils, field and laboratory measurements, digitization of archive materials (such as, for example, legacy soil data, maps and thin soil sections).

The task of formalizing a scientific experiment and organizing machine-analyzable data flows coming from various sources, is important for the development of scientific activity in the digital world. Journals, research institutes, universities and manufacturers of laboratory equipment make their own sparse attempts [2-3]. However, it is conducted intensively yet only from the perspective of increasing the reproducibility of the results of scientific research, which is only one of the goals. Currently, more or less formalized templates for describing measurement methods are created for shared use and interlaboratory exchange, such as, *Nature Protocols exchange, OpenWetWare Protocol Categories, Protocol Online: Search Protocols, A secure platform for developing and sharing reproducible methods, A peer-reviewed protocol journal Bio-protocol, Optimized Lab Protocols for Testing Soils, JoVe*. However, typically they represent a set of text descriptions of protocol steps, not suitable for automated processing and comparison. We aim for a formalization of the measurement protocol which would be an unambiguous and machine-readable description of the necessary conditions for performing the measurement, the measurement process itself, the results obtained and their mathematical or algorithmic processing. Formalization of a measurement method together with a detailed protocol of the experiment (which includes a certain instrument, its settings, current calibration, etc.) allows the reproducibility of the results, i.e., increase confidence in the data and make data FAIR [4], i.e., it improves and simplifies the exchange and development of research methods. At the stage of data meta-analysis, this makes it possible to identify and take into account experimental errors. If random errors are detected by repeated measurements, systematic errors (laboratory, operator, instrumental) can only be detected when analyzing large datasets. When systematic errors are found in the conduct of an experiment, one can also see what results they could affect. In some cases, it is possible to recalculate (correct) measurement results (for example, by recalibration). Another important goal is to use exactly the same information about experimental settings in protocol steps descriptions and in data processing scripts to avoid possible errors.

Existing systems for the formal description of experimental protocols – *Electronic Laboratory Journals (ELN) and Laboratory Information Management Systems (LIMS), namely, OSF, Labcollector, Hivebench, SciCloud, Accelrys (BIOVIA), Elabwtf, SciNote, Senaite, Bikalims, Occhiolino (GNU LIMS)* are either paid or shareware (free limited functionality or limited amount of storage, paid technical support), have limited options for embedding calculation functions, export options, and are focused specifically on laboratory analysis of physical samples, having no soil specificity. Abstract field objects, such as "terrain", "surface" and "soil profile" or a "trench", as well as long-term field experiments, for which there may be descriptions and measurements, are not included in the formalization scheme of such standard systems. Even though existing ELNs have convenient constructors for creating formalized protocols of experiments, they are designed only for the convenience of each individual user or group of researchers with their objects. Therefore, when implementing such a product by research institutes, a database collected from a set of ELNs of all employees will not be suitable for further effective joint analysis of the collected data. Moreover, a detailed formalization scheme is needed at least to evaluate existing open-source software suitability as components of a developed information system. Thus, soil research requires a specific implementation of such an information system with at least a soil-specific scripts and models library.

The disadvantage of existing soil and soil-geographic databases is that, firstly, they do not contain the history of the origin of objects and the sequence of actions on them (both in field, for example, technological maps of crop cultivation, and laboratory, for example, various treatments and fractionation of samples). This leads to data fragmentation and lack of relationships (potentially relevant data are lost). Secondly, they do not contain details of data acquisition methodology. Thus, only data obtained by one widely used method are selected, which cuts off all other data obtained by other or similar methods. At the same time, different experimental conditions are unavoidably mixed. Realizing the existing problem of formalizing laboratory measurement details, in the recent years International Soil Data Center in Wageningen has started to request information about methods steps from each laboratory which provided them legacy data, to evaluate datasets considering accuracy and precision [5]. While it is rather difficult to do it in detail for legacy data, it is possible to supply all the newly generated data with formalized data acquisition procedures.

The solution may be to have formalized data acquisition protocols that allow the maximum use of all available related data, for example assigning data sources different weights (calculated based on the accuracy of the method, the reliability of the data source, experimental errors or processing errors), homogenizing data to comparable values, by introducing corrections for experimental conditions (or experimental conditions could be directly used in mathematical models) or in other ways taking into account the differences in obtaining data. An ensemble statistical approach, using the entire available data set and models, while assimilating data coming from heterogeneous sources over time, is considered to be more informative for predictive modeling and estimating its uncertainties than the use of narrow subsampling [6-7]. The presence of links between objects allows combining initially independent experiments to analyze soil properties variability in space and time. For example, to make generalizations to obtain regional/global dependencies necessary for predictive models.

Formalized research protocols are published by a number of authoritative specialized journals *Nature protocols, Springer Protocols, Cell Protocols*. When formalized protocols are used, automatic processing of results and calculation of errors is possible. It becomes possible to transfer the entire database from one classification/description system to another according to established rules. One can identify intersecting sets in soil descriptions, measured properties and experimental parameters, as well as have control over consumables, the state of the instrument base, workload and etc.

The aim of this work was to develop a conceptual scheme for formal description of heterogeneous data and methods for its acquisition to ensure the possibility of organized collection and storage of all soil research results. It gives data reliability estimation for further analyses and generalizations, while providing reproducibility for research studies.

1. Results and discussions

The developed formalization scheme is shown in the Figure 1. This scheme shows logical elements of the database necessary for the coherent collection of complete and formalized information about experimental studies and further analysis of this information. The proposed scheme allows us to formally describe standard and non-standard methods as a sequence of simple actions (method elements). This scheme makes it possible to link descriptions and measurements carried out in the field and in the laboratory according to any formally described methods on such objects as “terrain”, “soil surface”, “transect”, “soil profile”, “soil horizon”, “sample”, “thin section” and etc. a single spatial database with a history of filling the object with data. This allows different researchers to supplement objects in the system with soil studies at any time in an arbitrary order and subsequently carry out meta-analysis on the required spacetime scale.

The database contains four main logical blocks: a block of reference information, a block of data, a block of methods and a user-specific/inventory block (Fig. 1).

1.1. Database structure

Information block

The reference information block contains a single expandable list of soil properties and a table describing various groupings of those properties, as well as a list of measurement units and their conversions. For example, grouping can be according to an object under study (area, profile, sample, etc.), field of knowledge (physical, chemical, etc.), description standard (properties of FAO, EGRPR, WoSIS, WISE etc.). In this case, one property can belong to several groups. Such grouping structure of soil property is universal (compatible with other standards) and is supposed to be extensible by adding new groups. Any legacy data can be imported “as is”, extending existing templates for entry of new data. The idea is not to create another new standard and not create data homogenization in advance but use homogenization scripts at export of data according to user specifications, using the advantage of formalized data acquisition.

Data block

The data block contains information about soil samples and other objects and the results of experiments. Data can be entered in any degree of detail, starting from a simple structure as a table “object-property-value” to a detailed description of the experiment. With a detailed description, all data is stored in the form of a history of accumulated events (creating a soil profile or a plot, description, sampling, incubation, measurement and so on). *The database does not impose strict requirements on the chronology of events, for example, measurements can precede the field description of a profile.* Each event is described by an entry in the e-journal (table “Journal entries”). The entry contains information about:

- described objects (relationship with the “Objects” table);
- created and destroyed objects by the method and relations between objects (table “Objects origin”). For example, a sample taken from a soil profile will be created; the one processed according to a sample preparation protocol will be destroyed and several new ones created: a reduced initial sample and, for example, several fractions;
- values of soil properties and their relations to objects through the table “Object-property-value” and “Values” of different types;
- optional raw data (if a processing script is applied by the method).

Each journal entry is linked to a method block (protocol and protocol steps).

Methods block

The block of data acquisition methods contains formalized methods and specific implementation protocols covering all three stages of data acquisition: preparation/preprocessing, measurement and data processing. The formalized methodology is described by a “Protocol” which consist of “Protocol steps” which are grouped and ordered. Each step represents an application of some method with certain optional settings and parameters. Each method has information on its applicability (types of source and result objects, result property), human-readable description, and optional detailed protocol which is recursively formally described as a sequence of steps (and listed in “Protocols”), i.e. each step itself can be represented by another protocol. Every step can have its processing script (“Model_id” in the “Data processing models” table) which acts on raw data. Journal data entries are allowed to be associated with a certain protocol or just with a single step. When a user changes step parameters a new step is created and is written to the journal, while the original step gets the status “modified”. Protocols can be created from existing base protocol steps (from database or from user’s own journal), while new protocol steps can be created from methods, adding instrumental setting/parameters and a processing model. Parameters depend on the calibration and settings of the instruments and the conditions of the experiment.

Thus, the whole variety of techniques is reduced to a manageable number of basic elements with parameters (such as temperature, duration, rotation speed, reagent, etc.). Each protocol step is associated with information about consumables and equipment that can be used at this step.

The user interface for data entry is created automatically based on the formalized method detail. The concept allows protocols with any degree of detail, assuming continuously increasing formaliza-

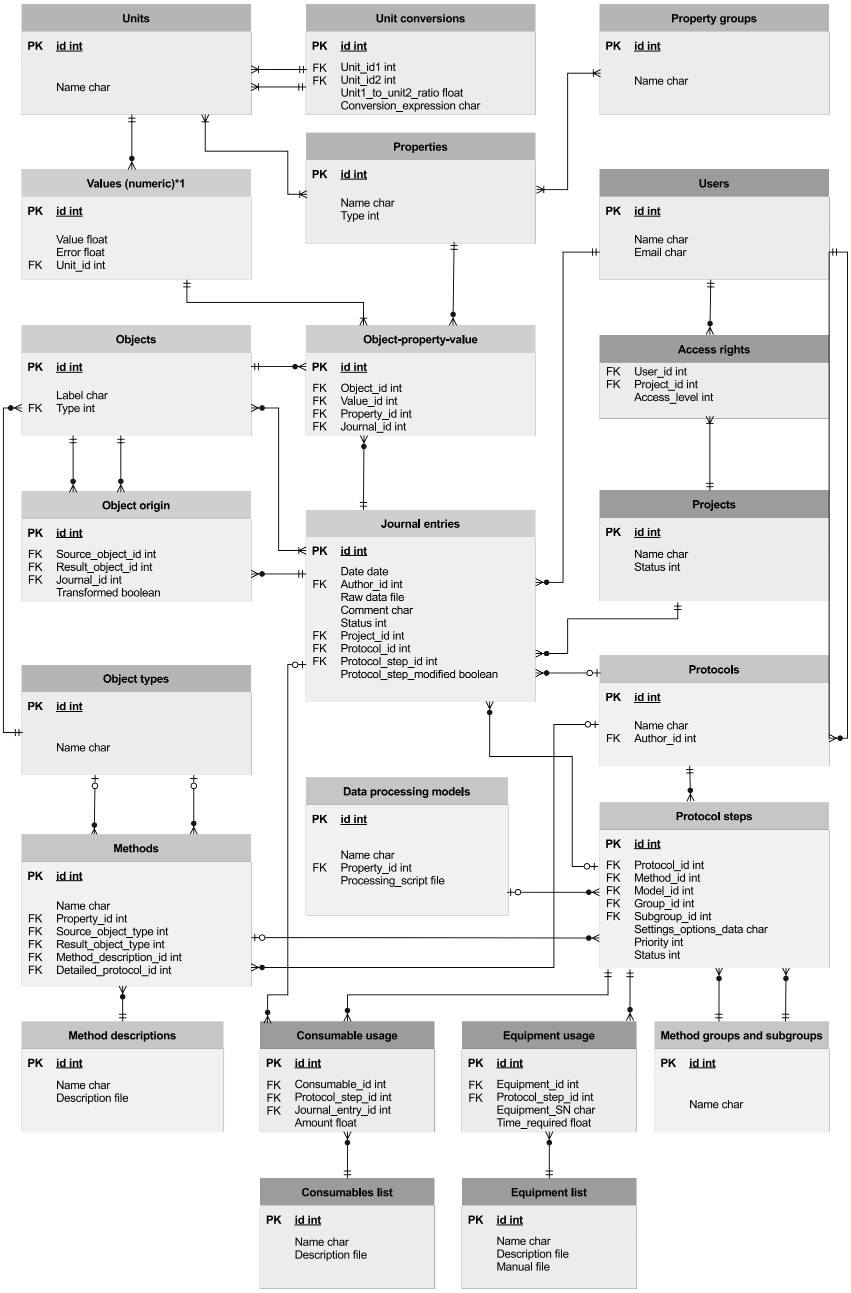


Fig. 1. Logical scheme of formalization for heterogeneous experimental data in Crow's Foot notation[1]

tion down to basic steps. Protocol has a branching structure and data entry can be performed by the user at any level of detail. However, the final estimates of dataset reliability, accuracy and reproducibility in the system are assigned according to the extent to which raw data were supplied. We believe that due to the presence of calculating scripts at each step (when required), overall standardization of data in e-journal (potential bonuses for automation in analyses), protocol reusability, collaborative mode and various "helpers" (statistical quality control, checks and availability of templates) will encourage users to enter raw data into the e-journal.

Along with standardized methods, in many cases of scientific research there is a need to store and use non-standard (author's or temporary), experimental methods and modifications. If for standard methods it

[1] The diagram contains a table "Values (numeric)" for numerical values, at the same time, it is provided that the database contains several similar tables for values of different types. For example, character strings for use in soil descriptions or bibliographic information, geographical information (points, contours) for introducing data from soil maps.

is enough for the user to indicate method name, since their formalization can be entered at any time after, for non-standard methods – formalization is the responsibility of the author.

It is known that in addition to the information described in state standards and other methodological manuals, “many different factors can affect the variability of measurement results performed using the same method, those including: operator; equipment; equipment calibration; environmental parameters; time interval between measurements.” [8]. *The proposed logical model allows to save all the details of the experiment in a formalized form.* In recent years, journals such as Science and Nature, as well as The Transparency and Openness Promotion (TOP) Committee, have been actively urging scientists to make their work transparent so that their experiments can be repeated “at least in theories.” They are developing increasingly stringent criteria for journal publications regarding the provision of formalized methods and detailed experimental protocols to improve the reproducibility of scientific results, and are even promoting testing for a pre-published experimental design (study pre-registration), which reduces the bias towards publishing results with a certain effect detection relative to experiments with a negative result, in which the intended effect was not detected with the corresponding study protocol [9-12].

Inventory block

The inventory block contains general lists of equipment (table “Equipment list”) and consumables (table “Consumables list”), as well as inventory records. Table “Consumable usage” contains information about amount and type of consumables required for each protocol step (when linked to protocol step), as well as amount, actually used in experiment (when linked to journal entry). Information about instrument being used in a given protocol step is stored in the “Equipment usage” table.

1.2. Advantages

This way of presenting data using the developed conceptual scheme differs in that it allows:

- to create and store any type of described object – from plot as a result of field partition to a soil fraction/solution or other objects obtained in complex experimental procedures, retaining the full chain of objects origin and treatments;
- to store a complex data structure: during the experiment, many measurements of the same value for one sample can be made, all of them can be stored in the presented concise scheme. An example of a multivariable dependence, or dependence of several values from several variables, can be the measurement of various gases emissions at changing soil temperature and moisture;
- to save the history of actions on the object, for example, store and update new data in a complex and long-term field experiment: the dynamics of carbon content in soil during changes in vegetation and fertilizers inputs is recorded as a sequence of single actions;
- to estimate reliability and accuracy of the dataset based on raw data provision, methods information and usage of automated data processing;
- to perform various analyses of protocols. For example, comparative analysis of by machine learning approaches to assess the dependence of results on protocol peculiarities, develop protocols, select a protocol for a specific task, taking into account such characteristics as applicability and accuracy.

In the case when the details of some experiment are not available, the database schema also allows to store information just as “object-property-value” as in most existing databases. Thus, the proposed scheme is compatible with known formats, for example, those used in the Soil Geographical Database of Russia (PGBD RF) [13], the Unified State Register of Soil Resources (EGRPR)[14], WISE Soil Property Database [15], the International Soil Carbon Network (ISCN), an intercontinental aggregator and provider of soil data for the Information System (WoSIS) of the International Soil Data Center (ISRIC) [16] and etc., while creating many new opportunities.

The developed scheme for formalization of heterogeneous soil data: 1) increases the reproducibility of scientific research results, 2) allows automatic data processing, and most importantly, 3) allows effective data mining and, thus, is an important base part in creating analytical systems for modeling scenarios and decision making.

1.3. Options for data entry protocols

Data entry is proposed to be carried out in 3 general stages – minimal, extend-ed and detailed (Table 1).

The minimal description allows to quickly make: an inventory of all the objects, involved instruments, a general overview of the methods, and is also used to enter data from literary sources when they do not contain detailed study protocols.

The extended description information already allows to fill up the Unified State Register of Soil Resources. From the point of view of instrumental base, it is already possible at this level to monitor the state and involvement of instruments in specific research protocols. It becomes possible to evaluate the types and volumes of produced data, workload of devices and employees, time ranges and costs of measurements. It allows to optimize work and carry out quality control (errors and systematic errors of personnel and devices with a possibility of its localization and correction).

Table 1

Types of protocols for data entry from external sources

Description\Type	Minimal	Extended	Detailed

Soil profile or sample description (level “data”)	User name;
	geolocation;
	sampling/description date;
	soil name (if description);
	measurement or sampling depth.
		Minimal description; and
		Names and depths of soil horizons
			Minimal description; and
			Any properties of the objects (location, soil profile, horizon, sample)

Methods of data acquisition	Method name;
	property name;
	source object type;
	result object type;
	file with description.
		Minimal description; and
		List of instruments with details (serial number, condition, precision, accuracy, etc.);
		list of consumables with usage
			Extended description; and
			Formalized protocol steps;
			Instruments settings;
			Data processing model with parameters.

A detailed description is produced continuously and is the result of a full-fledged electronic laboratory journal.

2. Case of implementation (work in progress)

The presented scheme of formalization is being implemented for the development of Information system that provides integration and multi-level presentation of legacy and current data (See Fig. 2). The top panel shows user interfaces for data entry, as well as a dashboard with reports and statistics, and the bottom panel shows related components of the database. Interaction between the interfaces and the database occurs through the electronic document management system. User (web) and program (web API) interfaces are divided into interfaces:

a) for data entry (such as field soil description helper, request forms for laboratory analyzes containing

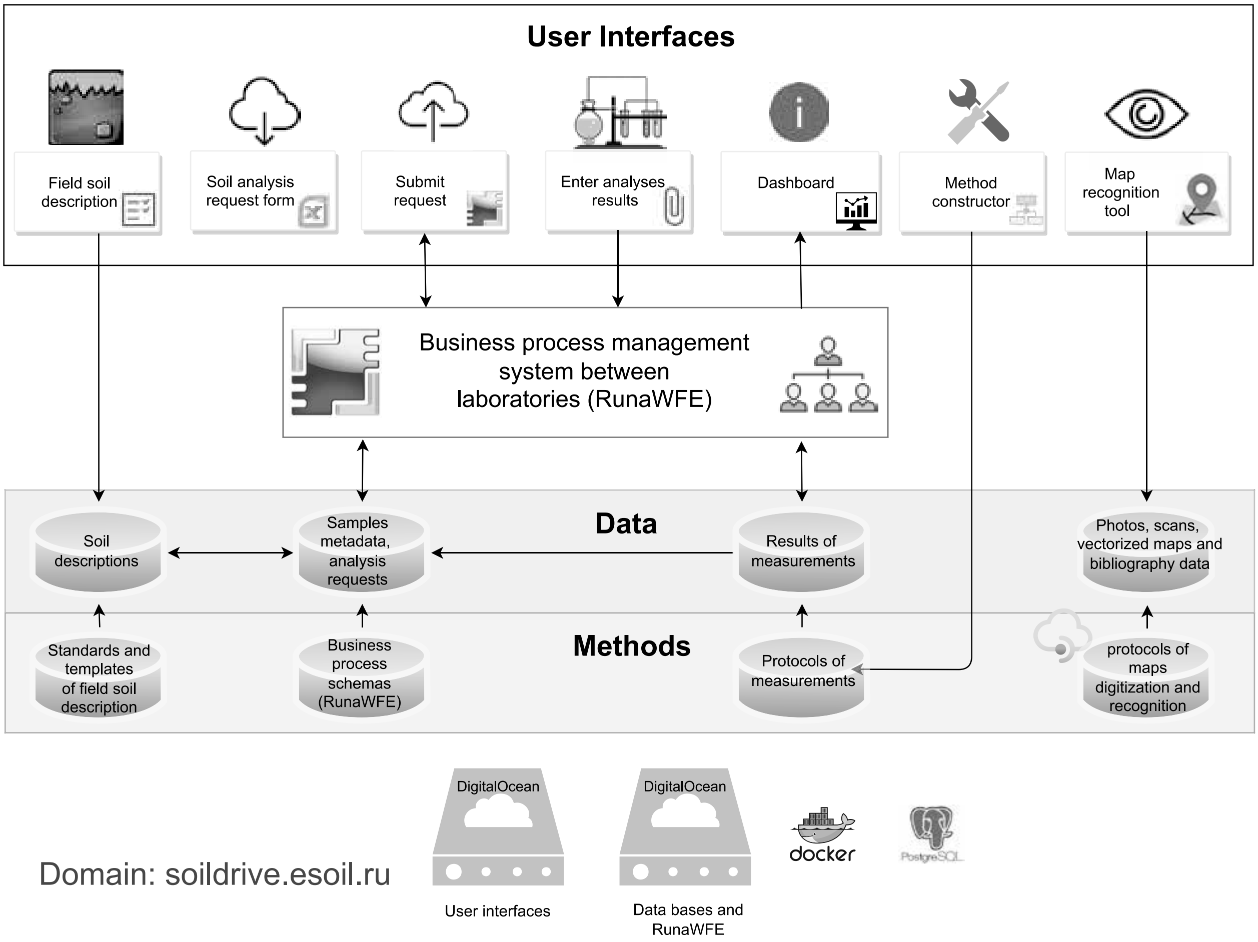

User Interfaces
Field soil description
Soil analysis request form
Submit request
Enter analyses results
Dashboard
Method constructor
Map recognition tool
Business process management system between laboratories (RunaWFE)
Data
Soil descriptions
Samples metadata, analysis requests
Results of measurements
Photos, scans, vectorized maps and bibliography data
Methods
Standards and templates of field soil description
Business process schemas (RunaWFE)
Protocols of measurements
protocols of maps digitization and recognition
Domain: soildrive.esoil.ru
DigitalOcean
DigitalOcean
docker
PostgreSQL
User interfaces
Data bases and RunaWFE

Fig. 2. General scheme of the information system

metadata of soil samples, upload of requests into the system and upload of measurement results by operator, constructor of methods and a subsystem for map recognition),

b) to search and create a data sample, generate statistics and reports, work with models in the information panel (data meta-analysis, generate scenarios, etc.).

The database consists of two levels, the first is "data" level (Fig. 2), which includes field soil descriptions, laboratory requests, metadata of the analyzed samples, the results of measurements, as well as photographs (profiles, thin sections, maps, microscopic, etc.), scans and vectorized maps, maps metadata. The second level is "methods of data acquisition" (Fig. 2), which includes:

– methods/schemas of field descriptions, schemas of electronic document management processes, methods of measurements and data processing models, methods of digitization and image recognition;
– detailed protocols for data acquisition with a description of certain measuring instruments or data processing (characteristics and settings of instruments, calibration curves and model parameters) and software.

Conclusions

The proposed formalization scheme makes it possible to store structured information about soils in various levels of detail. Formalization of data acquisition is the basis for the creation of an electronic laboratory journal containing the un-ambiguous formulation of a conducted or a planned experiment. The scheme provides the ability to search for experimental time series or compile pseudo-time experiments to provide simulation models with a relevant set of data for initialization and parameterization. This makes it possible to further tackle such an important scientific problem as estimating the effects of parametric and structural uncertainties in projections of ecosystem models. This work is part of the scientific rationale for the creation of a multi-level analytical system "Soil and land resources of Russia for agricultural production".

References

1. *Wadoux, A.M.J.-C., M. Roman-Dobarco, and A.B. McBratney.* 2021. Perspectives on data-driven soil research. European Journal of Soil Science 72:1675–1689. doi: 10.1111/ejss.13071.
2. *Giraldo, O, A. Garcia, and O. Corcho.* 2018. A guideline for reporting experimental protocols in life sciences. PeerJ 6:P.e4795. doi: 10.7717/peerj.4795.
3. *Halbritter, A.H, H.J. De Boeck, A.E. Eycott et al.* 2020. The handbook for standardized field and laboratory measurements in terrestrial climate change experiments and observational studies (ClimEx). Methods Ecol Evol. 11:22– 37. doi: 10.1111/2041-210X.13331.
4. *Wilkinson, M.D., M. Dumontier, I.J. Aalbersberg et al.* 2016. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 3(1):160018. doi: 10.1038/sdata.2016.18.
5. *Ribeiro, E., N.H. Batjes, and A.J.M. van Oostrum, eds.* 2020. World Soil Information Service (WoSIS) – Towards the standardization and harmonization of world soil profile data. Procedures manual 2020, Report 2020/01, Wageningen: ISRIC – World Soil Information 145 p. doi: http://doi.org/10.17027/isric-wdc-2020-01.
6. *Niu, S., Y. Luo, M.C. Dietze, T.F. Keenan, Z. Shi, J. Li and III F.S., Chapin.* 2014. The role of data assimilation in predictive ecology. Ecosphere 5(5):1-16. doi: 10.1890/ES13-00273.1.
7. *Martre, P., D. Wallach, S. Asseng, F. Ewert et al.* 2015. Multimodel ensembles of wheat growth: many models are better than one. Glob Change Biol. 21:911-925. doi: 10.1111/gcb.12768.
8. GOST R ISO 5725-1-2002. 2009. Tochnost' (pravil'nost' i pretsizionnost') metodov i rezul'tatov izmereniy. Chast' 1. Osnovnyye polozheniya i opredeleniya [Accuracy (correctness and precision) of measurement methods and results. Part 1. Basic provisions and definitions]. Moscow: StandartinformPubls. 24 p.
9. *Buck S.* 2015. Solving reproducibility. Science 348(6242):1403. doi: 10.1126/science.aac8041.
10. *Alberts, B., R.J. Cicerone, S.E. Fienberg, A. Kamb, M. McNutt, R.M. Nerem et al.* 2015. Self-correction in science at work. Science 348(6242):1420-1422. doi: 10.1126/science.aab3847.
11. *Belyaev, I.* 2015. Kharuko Obokata ne obzhalovala zaklyuchenie o fal'sifikatsii eyu rabot po sozdaniyu STAP-kletok, TASS. Available at: https://nauka.tass.ru/nauka/1685497 (accessed November 18 2022).
12. *Nosek, B.A., G. Alter, G.C. Banks, D. Borsboom, S.D. Bowman, S.J. Breckler et al.* 2015. Promoting an open research culture. Science 348(6242):1422-1425. doi: 10.1126/science.aab2374.
13. *Golozubov, O.M., V.A. Rozhkov, I.O. Alyabina, A.V. Ivanov, V.M. Kolesnikova, S.A. Shoba.* 2015. Technologies and Standards in the Information Systems of the Soil-Geographic Database of Russia. Eurasian Soil Science 48(1):1-10. doi: 10.1134/S1064229315010068.
14. *Alyabina, I.O., V.A. Androkhanov, V.V. Vershinin, S.N. Volkov, N.F. Ganzhara, G.V. Dobrovol'skii,*

A.V. Ivanov, A.L. Ivanov, E.A. Ivanova, L.I. Il'in, M.L. Karpachevskii, A.N. Kashtanov, V.I. Kiryushin et al. 2014. Edinyi gosudarstvennyi reestr pochvennykh resursov Rossii. Versiya 1.0. Available at: http://egrpr.soil.msu.ru/ (accessed November 18 2022).

15. *Batjes, N.H.* 2009. Harmonized soil profile data for applications at global and continental scales: updates to the WISE database. Soil Use and Management 25:124-127. doi: 10.1111/j.1475-2743.2009.00202.x.
16. *Harden, J.W., G. Hugelius, A. Ahlström et al.* 2018. Networking our science to characterize the state, vulnerabilities, and management opportunities of soil organic matter. Glob Change Biol. 24:e705–e718. doi: 10.1111/gcb.13896.

Vasilyeva N.A. PhD, Federal Research Center “V.V. Dokuchaev Soil Science Institute”, Pyzhevsky lane 7/2, Moscow, 119017, Russian, e-mail: nadezda.a.vasilyeva@mail.ru

Vladimirov A.A. PhD, Federal Research Center “V.V. Dokuchaev Soil Science Institute”, Pyzhevsky lane 7/2, Moscow, 119017, Russian, e-mail: artem.a.vladimirov@gmail.com

Vasiliev T.A. Federal Research Center “V.V. Dokuchaev Soil Science Institute”, Pyzhevsky lane 7/2, Moscow, 119017, Russian, e-mail: TarasVasiliev44@gmail.com

Условия результативного применения технологий искусственного интеллекта в агропромышленном комплексе ЕАЭС*

В.И. Будзко, В.И. Меденников

Федеральное государственное учреждение «Федеральный исследовательский центр «Информатика и управление» Российской академии наук», г. Москва, Россия

Аннотация. Рассматриваются решения для снижения экологической опасности в сельском хозяйстве единого агропромышленного пространства ЕАЭС. Предложен механизм формирования такого пространства, позволяющего разрешить возникшие геополитические, экономические, социальные, экологические проблемы. Это единая цифровая платформа управления, включающая возможность облачного построения на основе математического и онтологического моделирования, единых цифровых стандартах (структура подплатформы сбора, хранения и интеграции пооперационной первичной учетной информации всех участников в единой базе данных; структура подплатформы технологического учета; структура подплатформы алгоритмов обработки данных первых двух подплатформ в целях управления производством). При таком подходе применение технологий искусственного интеллекта принесет наибольший эффект и позволит обеспечить максимальную межотраслевую прослеживаемость продукции и будет минимизировано негативное воздействие природных и антропогенных факторов экологической опасности на окружающую среду, на продукцию агропромышленного комплекса и на самого человека.

Ключевые слова: *искусственный интеллект, экологические проблемы, агропромышленный комплекс, ЕАЭС, цифровые стандарты, прослеживаемость продукции.*

DOI: 10.14357/20790279230117

Введение

Высший Евразийский экономический совет провозгласил о планах формирования ряда единых рынков и пространств. Это рынки финансов, энергоресурсов, транспортно-логистического пространства, пространства свободного движения товаров и рабочей силы, а также единого агропромышленного пространства. Последние политические, экономические, социальные события, усиленные пандемией COVID-19 и санкциями, определяют интеграционные процессы в агропромышленном комплексе (АПК) Евроазиатского Экономического Союза (ЕАЭС) в качестве важнейших стратегических задач. Помимо решения проблемы обеспечения населения едой и достижения необходимого уровня продовольственной безопасности требуется решать проблемы нарушения экологии и качества пищи.

АПК при его технологическом развитии, как и транспортная отрасль, энергетический и коммунально-бытовой секторы, стало одним из главных загрязнителей природы, а растениеводство – источником наибольших экологических проблем, что обусловлено широким использованием различных ядохимикатов. Их воздействию подвергаются не только вредители, но и контактирующие с ними полезные организмы (птицы, черви, насекомые-опылители, бактерии и др.). Они либо гибнут, что приводит к нарушениям экосистемы, к эрозии, ухудшению структуры и плодородия почвы, либо накапливают в своих организмах эти ядохимикаты, передавая их по пищевым цепям, вплоть до человека.

По данным департамента мелиорации МСХ в России ежегодно деградирует 1,5-2 млн. га земель, и потери сельхозпродукции в зерновом эквиваленте составляют порядка 3,9 млн. тонн. Только из-за почвенных эрозий ущерб может достигать 25 млрд руб. в год. Потери наиболее богатого гумусом слоя

* Работа поддержана грантом Министерства науки и высшего образования РФ, внутренний номер 00600/2020/51896, договор № 075-15-2022-319.

земли составляют 1,5 млрд. тонн в год, которые включают 75 млн. тонн гумуса, 30 млн. тонн азота, фосфора и калия. Природное плодородие знаменитых чернозёмов в центре России за последние 10 лет уменьшилось в 1,5–2 раза, количество гумуса в почвах сократилось с 8–10% до 3–5% [1].

В республиках Центральной Азии ситуация еще хуже. В докладе Евразийского центра по продовольственной безопасности [2] отмечается, что Западная Европа характеризуется высоким уровнем заботы о почве, Восточная Европа с Россией характеризуются высокими темпами интенсификации сельского хозяйства с чрезмерной эксплуатацией самых плодородных почв и отказом от менее продуктивных земель, а азиатские страны ЕАЭС характеризуются самой высокой степенью и уровнем деградации почв.

В России основная причина экологических проблем – несоблюдение технологий выращивания растений при повсеместном нарушении севооборотных ограничений, норм и правил внесения ядохимикатов, которые попадают в почву, воду, воздух и, наконец, в продукты питания.

Экологические проблемы истощения и засоления плодородной земли, эрозии почв, возрастания гнетущего состояния флоры и фауны все больше привлекают внимание регулирующих органов. А качество пищи вызывает беспокойство у населения многих стран. Возрастающие возможности цифровой трансформации экономики создают условия для формирования единого агропромышленного информационного пространства производства продукции. Цифровизация управления и логистики на всех этапах жизненного цикла производства и доставки будет вынуждать всех участников обеспечивать продукцию надлежащего качества. В статье рассматриваются важные вопросы обеспечения успешного создания таких средств.

1. Прослеживаемость продукции

Цифровизация управления и логистики на всех этапах жизненного цикла производства обеспечивает прослеживаемость продукции и товаров, что позволяет достоверно информировать партнера, контролирующие органы, конечного пользователя об изготовителе, сроках, качестве, цене и других характеристиках товара. В отдельных отраслях предпринимались шаги по реализации такого подхода. Они носят фрагментарный характер, выполняются без полноценного онтологического моделирования предметных областей, не позволяют избежать многократного дублирования вводимой информации. Так, в АПК еще в 2018г. была принята федеральная государственная информационная система (ИС) электронной ветеринарной сертификации «Меркурий», в которой обязаны участвовать все предприятия оборота товаров животного происхождения. В настоящее время в АПК прорабатывается возможность расширения списка продуктов, подпадающих под реализацию механизма прослеживаемости, например, зерна. Однако без разработки цифровых стандартов, исходя из основных принципов цифровой экономики (ЦЭ) [3], которые включают единые онтологическую и концептуальную модель предметной области, словари, ограничения целостности и прочее, в ЕАЭС появится множество несовместимых систем. По этой причине идет медленное внедрение ИС Меркурий в силу отсутствия единых интерфейсов с ИС предприятий, цифровых стандартов на информационные ресурсы, на алгоритмы решения задач.

Единая цифровая платформа управления (ЦПУ) производством АПК, основанная на соответствующей математической модели [4], предлагается в качестве решения проблемы унификации цифрового инструмента прослеживаемости продукции на основе единых цифровых стандартов, на основе детального анализа цифровой экосистемы (ЦЭС) АПК [5] и ее составных частей.

2. Структура цифровой платформы управления АПК в России

С помощью данной модели и ряда технологий проектирования ИС определены несколько цифровых подплатформ, которые формируют конкретные цифровые стандарты, в сумме представляющие единую ЦПУ. Первая – подплатформа сбора и хранения пооперационной первичной учетной информации всех предприятий в единой БД (ЕБДПУ), структура которой включает следующие атрибуты: вид, объем и объект операции, место осуществления, субъект проведения, качественные характеристики, интервал времени проведения, задействованные средства производства, объем и вид потребленного ресурса. Необходимость цифрового стандарта на данные первичного учета обусловлена также прогнозом возрастания количества фиксаций различных действий на цифровизированных предприятиях (к 2050 году до 4,1 млн. в день) [6]. ЕБДПУ может быть реализована как облачная структура. Стандарт распространяется на конечное оборудование, интернет вещей (IoT). Он должен использоваться в межотраслевых взаимоотношениях между производителями, перерабатывающими, логистическими, оптовыми и розничными фирмами. Данный стандарт структуры первичного учета был проверен путем анализа референтных моделей, объединяющих и систематизирующих все знания по отраслевым бизнес-моделям [7, 8] и нашел подтверждение в других отраслях народного хозяйства России, а также в странах ЕАЭС.

Соответствующие атрибуты из смежных отраслей и атрибуты, отражающие бухгалтерскую информацию (накладные, акты), а также данные статистического учета, плановых и оперативных служб должны найти отражение в цифровом стандарте первичного учета в АПК. ЕБДПУ должна строиться в соответствии со стандартными классификаторами, справочниками, словарями и во взаимодействии с единой БД технологического учета (ЕБДТУ) и подплатформой базы знаний (рис. 1).

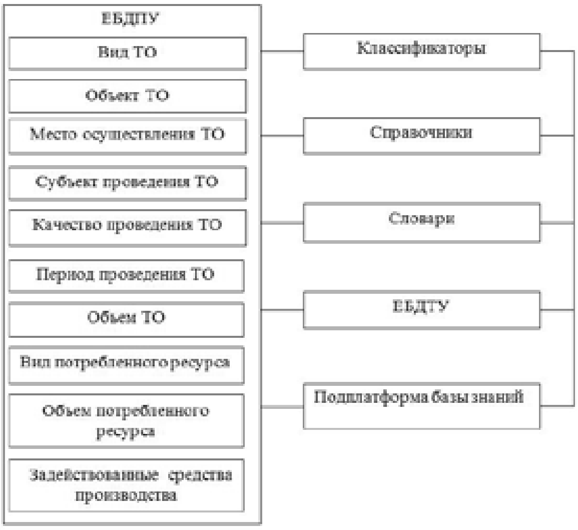


Рис. 1. Цифровой стандарт структуры первичного учета

Заметим, что аналогичный цифровой стандарт в последние 2-3 года начал активно внедряться в США при разработке подплатформ-агрегаторов первичного сбора и накопления сельскохозяйственной информации и прикладных подплатформ (управленческих задач) [9].

Подплатформа ЕБДТУ – единая для всех предприятий уже некоторой отрасли. Так, на рис. 2 представлена укрупненная информационная схема растениеводства единой для всех хозяйств концептуальной информационной модели растениеводства в составе 946 атрибутов. При этом облачная подплатформа ЕБДТУ также должна быть интегрирована с соответствующими классификаторами, справочниками, словарями, ЕЬДПУ и подплатформой базы знаний.

Третья подплатформа – описания алгоритмов управленческих задач (база знаний), единых для всех предприятий также определенной отрасли. Сформулировано около 240 задач для растениеводства.

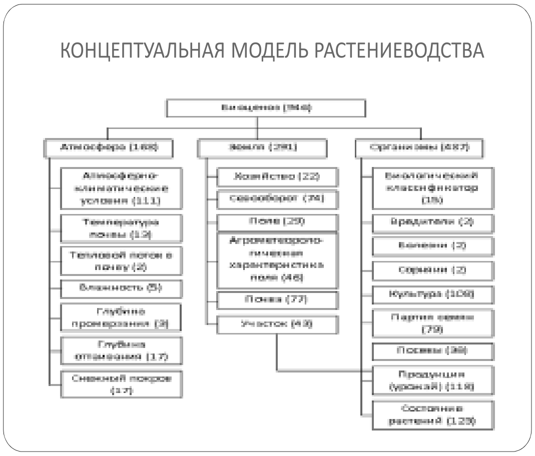


Рис. 2. Укрупненная информационная схема растениеводства

Разработанные стандарты цифровой платформы управления приобретают особенное значение в эпоху ЦЭ, когда технологии дистанционного зондирования земли (ДЗЗ) и геоинформационные системы (ГИС) начинают активно внедряться при реализации точного земледелия (ТЗ). ТЗ все шире внедряется в новые агротехнологии при высокоточном позиционировании на основе технологий ДЗЗ и ГИС, а также дифференцированных высокоэффективных и экологобезопасных агротехнических мероприятий на полях на основе подробной информации химико-физических характеристик каждого участка.

Появление более совершенного электронно-оптического оборудования, устанавливаемого на различных подвижных и стационарных аппаратах, специальных инструментов дешифровки спектральных параметров растений дает возможность рассчитывать различные вегетационные индексы, характеризующие фазы развития и биомассу их во временном разрезе. Такой полученный динамический ряд данных ДЗЗ обеспечивает анализ проведения большинства агротехнических мероприятий с выявлением зараженных болезнями и вредителями угодий с оценкой причиненного им ущерба, а также последствий стихийных природных явлений. В этих условиях учет и мониторинг максимально возможного количества сельскохозяйственных процессов становится основной целью в разработке стратегией цифровизации крупнейших агропромышленных и машиностроительных фирм в мире. Сориентироваться в этом потоке информации самостоятельно практически невозможно [10]. И здесь на помощь должны прийти технологии искусственного интеллекта (ИИ).

А это требует интеграции огромного объема данных для обработки с применением ИИ [11]. Сформулированные цифровые подплатформы определяют рациональную схему интеграции данных технологий в ЦПУ АПК [12].

Такая интеграция с внедрением ТЗ будет способствовать повышению эффективности производства продукции, отвечающей необходимым ценовым, качественным и экологическим требованиям. Интенсивные исследования в совершенствовании этих технологий ведутся во многих странах.

3. Примеры применения технологий искусственного интеллекта в точном земледелии

Многие известные технологии ТЗ, включающие экологическую составляющую, используют ИИ, отметим некоторые из них [10, 13]. В России также есть отдельные разработки, пока узконаправленные, без комплексного подхода. Приведем в качестве примера некоторые наиболее продвинутые зарубежные решения.

3.1. Машинное обучение при мониторинге полей

Израильский продукт Taranis предоставляет точную информацию о состоянии растений на основе показаний полевых датчиков, метеостанций, аэрофотосъёмки, что позволяет своевременно выявлять негативные факторы в виде идентификации болезней и вредителей, дефицита питательных веществ с выработкой рекомендаций по оперативному вмешательству.

Платформа Watson Decision Platform for Agriculture от IBM дает консультации при возникновении рисков поражений кукурузы на основе данных ДЗЗ (индекса HD-NDVI) о дозах, типе пестицидов и оптимальных сроках их внесения. Фермеры получают прогнозы урожайности и др.

Платформа искусственного интеллекта Health Change Maps and Notifications компании Farmers Edge оперативно информирует фермера об эффективности работы техники, состоянии растений, появлении вредителей или болезней, дефиците питательных веществ и др.

Приложение Field Manager от Bayer на мобильном телефоне даёт пользователю информацию о возможных рисках с посевами и рекомендации о способах их предотвращения на основе обработки данных ДЗЗ и большого количества других данных из БД.

Платформа Hummingbird Technologies обеспечивает фермеров информацией о текущем состоянии и объемах растительной массы, наличии сорняков, дефиците у растений азота и др. не только на основе данных ДЗЗ, но и наземных средств мониторинга, снимков БПЛА.

3.2. Технологии ИИ для борьбы с сорняками

Активно развиваются работы по применению ИИ для борьбы с сорняками и вредителями. Так, компаниями Bayer и Bosh разрабатывается технология умного опрыскивания Smart Spraying, которая будет “узнавать” сорняк и определять вид и необходимое количество пестицида. «Убийца сорняков» от компании EcoRobotix способен самостоятельно перемещаться по полю, дифференцированно распознавая и обрабатывая обнаруженные сорняки. Утверждается, что технология позволит в 20 раз сократить объём использования гербицидов.

Автономная система WeedSeeker компании Trimble производит точечное опрыскивание сорной растительности. Система идентифицирует сорняки с помощью светодиодов сканирующих поверхность в красном и инфракрасном диапазоне. Отражённый свет автоматически анализируется, при обнаружении растения сигнал подаётся на форсунку, которая срабатывает точно над ним.

3.3. Технологии ИИ идентификации болезней растений

Современные технологии ИИ помогают фермерам после идентификации заболевания растений выбрать методы их лечения с расчетом экономических показателей. Процесс происходит на основе фотографий поражённой части растения. Мобильное приложение Plantix компании Peat предоставляет фермерам возможность идентификации свыше 60 болезней растений. Приложение содержит огромную БД снимков с идентификацией по сортам растений, видам бактерий, заболеваний и др.

Приложение Scouting Bayer-BASF также помогает диагностировать заболевания, нарушения развития, степень обеспеченности азотом растений на основе обработки фотоснимков.

3.4. Технологии искусственного интеллекта в цифровизации животноводства

Опыт цифровизации животноводства во многих странах показывает, что почти все технологические операции в отрасли поддаются цифровой трансформации с использованием ИИ. Приведем основные направления данной трансформации.

- улучшение качественных условий содержания животных за счет умных систем управления световым режимом, микроклиматом, кормлением, навозоудалением, поскольку комфорт животных влияет на их продуктивность;
- селекция пород. Селекция позволит максимально точно вывести породу по заданным требованиям и свойствам (отсутствие генетической предрасположенности к определенным болезням, мясные и молочные качества, скорость роста и созревания). В настоящее время большие надежды возлагаются именно на методы ИИ, например, на разработку методов анализа геномной информации для оценки племенной ценности животного в раннем возрасте. В настоящее время ведутся исследования по выбору пола животного, молочной продуктивности, толщины отруба для стейка;
- анализ качества молока; диагностика и профилактика заболеваний животных; соблюдение санитарно-гигиенических норм.

3.5. Опыт использования технологий ИИ в сельском хозяйстве России

Вслед за развитыми странами в нашей стране стремительно появляется много компаний, предлагающих различные решения в области цифровизации сельского хозяйства, в частности, технологий ИИ, ведутся интенсивные исследования во многих НИИ.

Например, в 2021г. ФИЦ ИУ РАН совместно с Самарским ГАУ был проведен полевой эксперимент по съемке сельскохозяйственной растительности на основе нейросетевых алгоритмов анализа мультивременных, мультимасштабных гиперспектральных данных, полученных с использованием гиперспектральной камеры на основе схемы Оффнера. Для чего была разработана методология построения вычислительно эффективного и компактного описания характерной (особой) точки изображения. Эта технология предназначена для пространственного сопоставления изображений с помощью аппарата вычисления дискрибирауемых особых точек изображения и последующего поиска гомографии с помощью инфраструктуры RANSAC. Обосновано использование специальных обучаемых дескрипторов характерных (особых) точек.

А также был разработан оригинальный алгоритм быстрого вычисления сверток в нейросетевых моделях на процессорах общего назначения. На сегодняшний день задача сегментации цифровых изображений неразрывно связана с методами детектирования и классификации целевых объектов. Использование глубоких нейронных сетей для задач классификации и особенно детекции ограничивается их большой вычислительной сложностью. В условиях обработки больших объемов данных, ключевым моментом является применение методов оптимизации быстродействия нейросетевых классификаторов. Поэтому был предложен оригинальный алгоритм вычисления свертки с буфером фиксированного размера, который не только вычисляется быстрее на 10-20% относительно классического метода, но и уменьшает объем дополнительной памяти, требуемой для вычисления дискретной свертки.

Однако, в сельском хозяйстве остаётся ряд проблем, мешающих полномасштабной цифровой трансформации отрасли и требующих совместного комплексного решения бизнеса, науки и правительства. Соответственно, эти проблемы касаются и приложений ИИ, которые должны пройти интеграционные преобразования в стандарты предлагаемой единой цифровой платформы управления отраслью. Перечислим некоторые из них.

Большой интерес к цифровым технологиям в развитых странах продиктован все более усложняющимися и дорогими традиционными технологиями повышения эффективности и качества продукции сельского хозяйства, а наличие большого резерва России в совершенствовании таких традиционных направлений повышения эффективности отрасли, как необходимость обновления парка сельхозтехники при высокой стоимости высокотехнологических средств цифровых технологий и сложности их освоения, а также отсутствие квалифицированных кадров в стране не порождают «социального заказа» у большинства хозяйств. При этом фиксируются большие финансовые и продуктовые потери. Комплексное использование цифровых технологий могут начать лишь немногие отечественные хозяйства.

Цифровая трансформация экономики требует замены или доработки производственного оборудования на цифровое, и этот процесс довольно сложен и дорог. Приобретение же дорогостоящей, наукоемкой, цифровой техники и оборудования, для обслуживания которого нужны профессиональные кадры, могут позволить опять же лишь крупные хозяйства.

В стране слабо развито производство приемо-передающих устройств, датчиков, исполнительных механизмов и другой аппаратуры, необходимых для применения технологий автоматического управления технологическими процессами сельскохозяйственной техники.

В отрасли продолжается эпоха «позадачного» фрагментарного проектирования и разработки ИС с формированием в каждом предприятии собственных информационных моделей БД, информационно несовместимых, когда у различных производителей приобретаются отдельные, так называемые «готовые» программные комплексы, не связанные ни функционально, ни информационно. Даже в агрохолдингах созданы оригинальные ИС, разработанные под текущие нужды каждого конкретного предприятия со своим информационным обеспечением, понятийным аппаратом, алгоритмами решения задач и разнородным программным и аппаратным обеспечением. Поэтому оказался невостребованным положительный практический опыт комплексной информатизации крупных агропромышленных предприятий Краснодарского и Ставропольского краев на базе прообраза ЦПУ с внедрением экспертных системы, являющихся прототипом современных методов ИИ для выращивания томатов и сахарной свеклы. За 2 года отдельные подсистемы были внедрены тогда в более, чем 1000 предприятий. При этом в регионах создавались центры внедрения и обучения.

Комплексное применение ИИ возможно лишь при получении реальных данных о каждой отдельной технологической операции, человеке, механизме, животном, а иногда даже растении. А в настоящее время около 90% всех данных записывается на бумаге или вручную вносится в Excel. Использовать такие данные практически невозможно.

Об этом говорят и результаты указанного выше совместного эксперимента ФИЦ ИУ РАН и Самарского ГАУ по использованию технологий ДЗЗ и ИИ в ряде хозяйств Самарской и Ростовской областях. Так, был сделан вывод о том, что руководствуясь лишь данными ДЗЗ, включающими гиперспектральную визуализацию, рекомендовать с высокой долей достоверности о проведении необходимых технологических мероприятиях сложно, так как необходимо учитывать и большое множество других источников данных: погодного мониторинга, солнечной активности, розы ветров, воздействия агрохимикатов, заселенности полей насекомыми, агрохимического состава, влагосодержания и других характеристик почв. Лишь наличие указанной достоверной информации позволяет создавать комплекс математических моделей принятия соответствующих решений.

Поэтому в стране необходимо направить усилия на комплексную отработку на основе ЦПУ самых совершенных отечественных цифровых технологий и компетенций, подобно развитым странам, на нескольких эталонных объектах с оснащением их современными ИКТ, датчиками, приборами, технологическим оборудованием и машинно-тракторным парком, совместимыми как друг с другом, так и приспособленными к различным цифровым технологиям, охватывающими всевозможные направления их развития в мире, с последующим массовым внедрением наиболее эффективных из них по всей стране. На эталонных объектах должна быть проверена на практике потребность в необходимых специалистах для цифровизации сельского хозяйства, полученная с помощью модели формирования ЦПУ в количестве 90000 человек с детальным расчетом по специальностям. На основе опыта комплексной информатизации отрасли в рамках программы электронизации сельского хозяйства внедрение отдельных цифровых технологий могло бы начаться через полгода при наличии организационной структуры генерального конструктора ЦПУ.

4. Единое транспортно-логистическое пространство ЕАЭС

Одним из качественных изменений, связанных с внедрением ЦПУ АПК в ЕАЭС, станет сдвиг в сторону коллективного сознания и кооперативных форм взаимодействия взамен индивидуализма каждой страны. Цепочка производства и сбыта устроена сегодня таким образом, что каждый участник, оценивая свои риски, закладывает их в цену своего продукта. При этом каждый следующий участник цепочки «покупает» риски, заложенные всеми предыдущими участниками цепочки, прибавляя свои в цену продукции. В результате конечный участник, например, магазин собирает все риски и «продаёт» их потребителю. Таким образом, аккумулированные риски оплачивает население. Такое взаимодействие приводит невосприимчивости к цифровой трансформации все звенья цепочки, поскольку каждый заинтересован лишь в своей марже, не вникая в системность всей цепочки.

Цифровая платформа управления при соответствующем расширении ее информационной базы на основе единых стандартов может обеспечить технологический прорыв в ЕАЭС. Логистика производства и доставки продукта до потребителя любого вида продукта в ЕАЭС, а не только сельскохозяйственного, может формироваться на основе математической модели формирования единой ЦПУ в интересах АПК, логично интегрированную в единую производственную ЦПУ. Построение ИС на единой ЦПУ логистикой улучшит управление взаимоотношениями с потребителями, обслуживанием потребителей, спросом, выполнением заказов, производством, поставкой, разработкой продукции и доведением ее до коммерческого использования, реверсивными потоками для всех стран содружества.

Производство большинства товаров в современном мире осуществляется при участии многих предприятий из всего большего количества стран. Дробление производства растет. Даже появился термин – “цепочки добавленной стоимости”. Технология такого дробления требует информационной совместимости потока данных по всей цепочке. Посредники, которые не создают добавленной стоимости, из логистической цепочки в рамках ЦПУ логистики исключаются. Переход на электронную интегрированную логистику на основе технологии распределенных реестров сыграет определяющую роль в достижении постоянного контроля за материальными потоками в реальном масштабе времени в режимах удаленного доступа через ЦПУ, и позволит учитывать потенциальные возможности производства, снабжения, потребления [14, 15].

Как отмечалось выше, аналогичный процесс внедрения ЦПУ в виде ЕБДПУ и базы знаний в последние годы происходит в США. Компания J’son & Partners Consulting, анализируя состояние дел, считает, что использование технологий двух указанных выше платформ в цепочке формирования добавленной стоимости аграрной продукции (оптовые компании, логистика, розничные сети) предоставит возможность перехода к прямым продажам, когда производитель прослеживает конечного потребителя, объем и структуру его спроса. Он производит ровно ту продукцию, которая нужна потребителю, и в нужное ему время при использовании математических моделей, в том числе, предиктивной аналитики. Управление доставкой продукции происходит путем автоматического обмена информацией между участниками цепочки поставок через облачный сервис и с минимизацией использования складской и логистической инфраструктуры оптовых посредников [9]. Такая цифровизация дает возможность исключить из цепочки множество ненужных посредников, на которых сейчас приходится до 80% стоимости от розничной цены товара. Такие сервисы будут доступны, в том числе, для малых хозяйств, что позволит существенно повысить эффективность отрасли и снизить риски деятельности для всех участников цепочки формирования добавленной стоимости: поставщиков ресурсов, потребителей продукции и транспортных фирм.

Центральное звено в цифровизации экономики – цифровая трансформация предприятия, которая требует пересмотра идеологии и технологии управления и их оформления в виде стандартов. Наиболее дальновидные эксперты об этом уже давно предупреждают. Так, директор Института экономики РАН Ленчук Е.Б. считает, что надо сосредоточиться на цифровизации именно реального сектора экономики, где она даст наибольший экономический эффект [16]. Видный экономист Агеев А.И. утверждает также, что, хотя уровень цифровизации банков, связи, государственных услуг будет выше, однако степень вовлеченности промышленности является индикатором цифровой зрелости всей экономики [17]. Для этого требуется единое понятийное поле, единое семантическое пространство за счет создания стандартов и соответствующих систем управления.

ЦПУ и использование технологий ИИ, умных контрактов, входящих в БД знаний платформы могут в корне изменить ситуацию (рис. 3). Современные инструменты позволяют прозрачным и корректным образом оценить и учесть вклад каждого звена цепочки в себестоимость конечного продукта с фиксацией объективного вклада каждого из них.

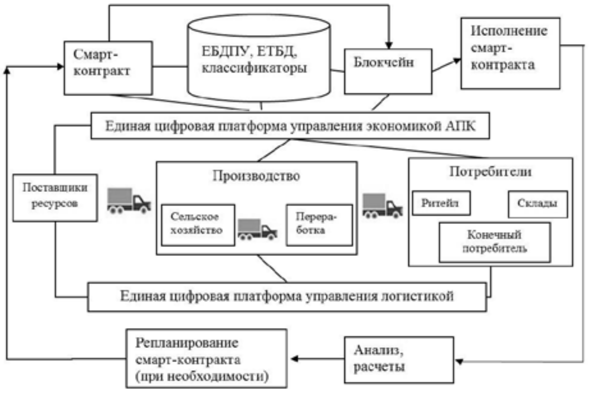

Смарт-контракт
ЕБДПУ, ЕТБД, классификаторы
Блокчейн
Исполнение смарт-контракта
Единая цифровая платформа управления экономикой АПК
Поставщики ресурсов
Производство
Сельское хозяйство
Перера-ботка
Потребители
Ритейл
Склады
Конечный потребитель
Единая цифровая платформа управления логистикой
Репланирование смарт-контракта (при необходимости)
Анализ, расчеты

Рис. 3. Схема прослеживаемости продукции при интеграции ЦПУ АПК и ЦПУ логистики

Таким образом, ЦПУ позволяет проследить весь жизненный цикл продукции и адекватно учесть все транзакции, является основой реализации умных контрактов, становится выгодна всем участникам цепочки, позволяя равномерно распределить риски между всеми участниками, что приводит к снижению издержек и возрастанию инновационной восприимчивости участников с получением существенной экономической выгоды от такой кооперации.

5. ЦПУ АПК и задачи сохранения экологии

Более половины их 946 показателей концептуальной модели растениеводства, укрупненная информационная схема которой представлена на рис. 2, имеют отношение к экологии. Так, в группе «Земля» – 291 показатель, в подгруппе «Севооборот» – 30 показателей. Подгруппа «Участок» группы «Поле» содержит показатели: «Запрещающие условия использования земельного участка», «Геоморфологические характеристики», «Мелиоративная характеристика», «Грунтовые воды», «Засоление», «Почва», «Агрофизическая характеристика», «Гидрофизическая характеристика», «Состояние почвы». Аналогично, в подгруппу «Культура» (108 показателей) входят следующие показатели: «Экологическая группа сорта», «Поражаемость болезнями по видам болезней», «Поражаемость вредителями». Интеграция экологических показателей в DCP AIC позволяет комплексно решать экологические проблемы (рис. 4).

За счет комплексной экологической оценки земель, всего производственного процесса с учетом поступающих ресурсов и продукции на выходе, формирования соответствующих управленческих решений, направленных на предупреждение проявления и минимизацию последствий проявления антропогенных и природных факторов экологической опасности можно значительно снизить экологическую опасность в АПК ЕАЭС.

В упомянутом выше докладе Евразийского центра по продовольственной безопасности [2] предлагаются следующие меры для снижения экологической опасности в Евразии, перекликающиеся с нашими предложениями:

– повышение количества и качества почвенных данных и информации: сбор данных, анализ, проверка, представление, мониторинг и интеграция с другими дисциплинами;
– создание единой базы данных почв Евразии, перевод исторических источников в цифровой формат, разработка единой системы мониторинга почвенного покрова, проведение школ по цифровой почвенной картографии;

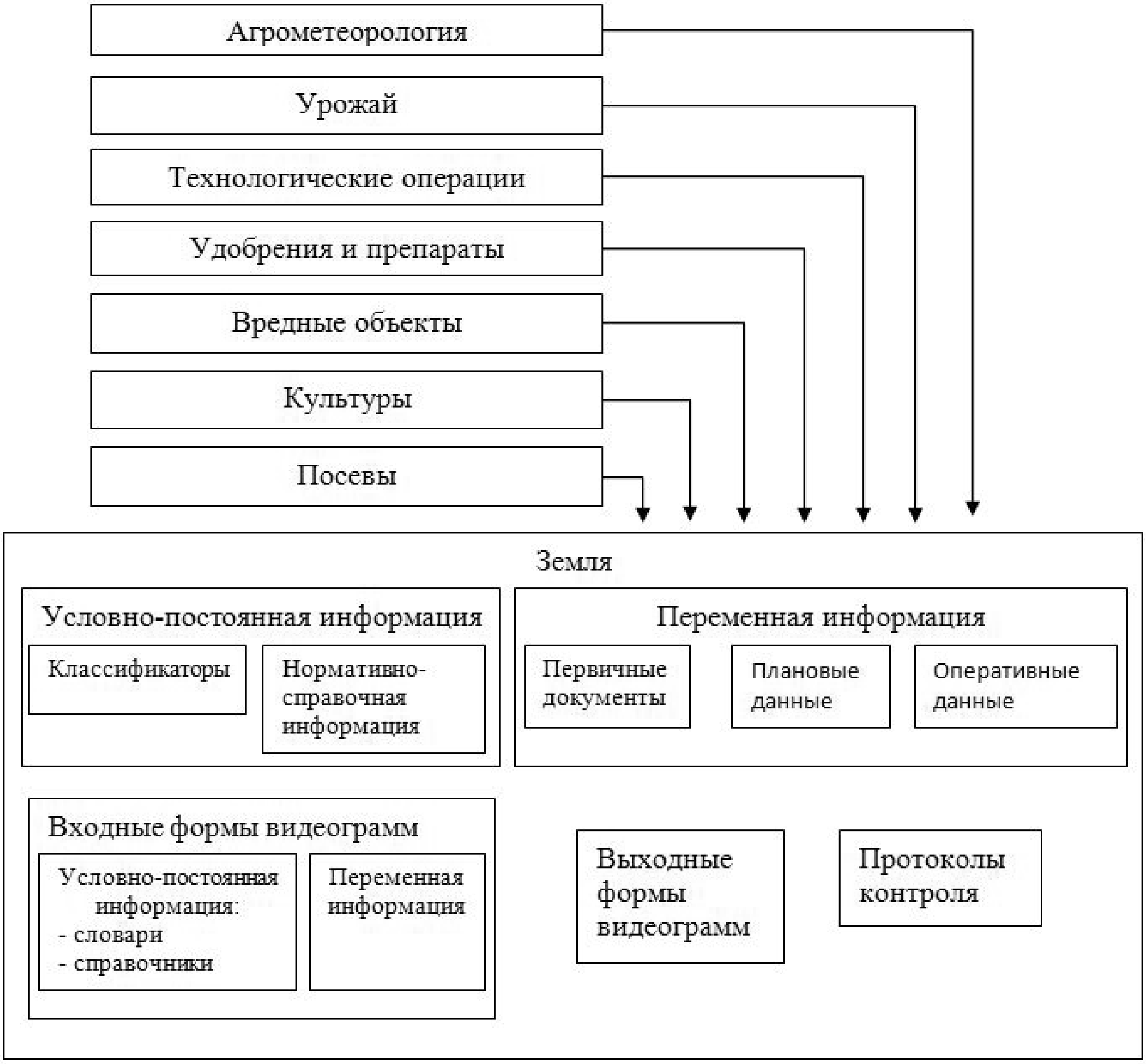


Рис. 4. Решения экологических задач в единой ЦПУ АПК

– внедрение стандартных методов анализов и единиц измерений, гармонизация методов мониторинга, баз данных и картографических материалов.

Заключение

Построение единой системы управления производством и доставки продукции на основе принципов, разработанных для ЦПУ АПК ЕАЭС, позволит снизить себестоимость производимой продукции и конечную цену, которую будет оплачивать потребитель при одновременном значительном снижении экологической опасности в содружестве. Внедрение ЦПУ также обеспечит комплексную экологическую оценку земель, экологический мониторинг их, всего производственного процесса с учетом поступающих ресурсов и продукции на выходе, даст возможность формирования соответствующих управленческих решений, направленных на предупреждение проявления и минимизацию последствий проявления антропогенных и природных факторов экологической опасности.

ЦПУ предоставляет цифровой инструмент прослеживаемости продукции, важное условие результативного применения которого – унификация, основанная на единых цифровых стандартах. Первый стандарт отражает структуру и связи пооперационной первичной учетной информации всех предприятий. Второй – единую для всех предприятий некоторой отрасли структуру и связи технологической информации. Третий стандарт отражает описание алгоритмов управленческих задач (база знаний), единых для всех предприятий определенной отрасли. Эти цифровые стандарты должны быть интегрированы как между собой, так и с соответствующими классификаторами, справочниками, словарями. Только при таком подходе можно обеспечить структуриризацию и заданный уровень надежности необходимого объема данных для эффективного применения технологий ИИ и обеспечить информационную совместимость по всей логистической цепочке перемещений товаров и продукции с возможностью постоянного контроля за материальными потоками в реальном масштабе времени.

Литература

1. Деградация на миллиарды: в России истощены свыше 60% сельхозугодий: URL: https://agroru.com/news/degradatsiya-na-milliardy-v-rossii-istoscheny-svyshe-60-selh-85534.htm (дата обращения 17.06.2022).
2. *Красильников П.В.* Обзор деградации почв и земель в Евразии. Доклад Евразийского центра по продовольственной безопасности 3-5 октября 2017г. URL: https://ecfs.msu.ru/images/documents/ECFS_October_conf/2017/ECFS2017_Krasilnikov_Land-and-soil-degradation.pdf (дата обращения 17.06.2022).
3. *Viktor Medennikov and Alexander Raikov.* Formation of the Digital Platform for Precision Farming with Mathematical Modeling. CEUR Workshop Proceedings 2790: 114–126, (2020), http://ceur-ws.org/Vol-2790/.
4. *Меденников В.И.* Математическое моделирование цифровых платформ и стандартов для управления экономикой страны // Журнал «Информатизация образования и науки», 2020, 3(47), 2020. С. 57-72.
5. *Меденников В.И.* Системный анализ цифровых экосистем производственных отраслей на примере АПК // Цифровая экономика. 2021. № 3. С. 69-74.
6. Искусственный интеллект в сельском хозяйстве. URL: https://agropravda.com/news/novye-technologii/11301-iskusstvennyj-intellekt-v-selskom-hozjajstve (дата обращения 17.06.2022).
7. *Меденников В.И.* Формирование единой цифровой платформы управления сельским хозяйством ЕАЭС // Сборник статей всероссийской научно-практической конференции «Европейский союз в глобальной экономике: агропродовольственный аспект». г. Саранск, 15 марта 2019. Саранск: Изд-во Мордов. ун-та, 2019. С. 134-138.
8. *Меденников В.И., Микулец Ю.И.* Цифровые стандарты – основа интеграции цифровых платформ АПК и других отраслей // Вестник Московского гуманитарно-экономического института. 2021. № 1. С. 208-226. DOI 10.37691/2311-5351-2021-0-1-208-226.
9. Цифровизации сельского хозяйства в России не хватает данных. URL: http://www.iksmedia.ru/news/5533967-Czifrovizacii-selskogo-xozyajstva.html#ixzz6KBD7IYEP (дата обращения 17.06.2022).
10. Как начать внедрять точное земледелие на предприятии. URL: https://smartfarming.ua/ru-blog/kak-nachat-vnedryat-tochnoe-zemledelie-na-predpriyatii (дата обращения 17.06.2022).
11. *Галустьян А.* Пять проблем, которые пока не может решить искусственный интеллект. URL: https://rb.ru/opinion/problemy-ii/ (дата обращения 17.06.2022).
12. *Budzko V. and Medennikov V.* Mathematical modeling of evaluating the effectiveness of using RSD data in precision farming, Procedia

Computer Science : 11th, Natal, Rio Grande do Norte, November 10-15, 2020. – Natal, Rio Grande do Norte: 122-129, (2020), https://doi.org/10.1016/j.procs.2021.06.015.

13. *Абросимов В.К., Райков А.Н.* Интеллектуальные сельскохозяйственные роботы. – М.: Карьера Пресс. 2022. – 512 с.
14. *Толуев Ю.И., Планковский С.И.* Моделирование и симуляция логистических систем. – Киев: «Миллениум», 2009. – 85 с.
15. *Medennikov V. and Raikov A.* (2021) “Optimizing of Product Logistics Digital Transformation with Mathematical Modeling” Journal of Physics: Conference Series : 13, Saint Petersburg, October 06–08, 2020 Saint Petersburg: 012100 (1-9).
16. *Ленчук Е.* Цифровая экономика в России? Секундочку ... URL: https://zen.yandex.ru/media/freeconomy/cifrovaia-ekonomika-v-rossii-sekundochku-5ccc6762a8ac8300b3495949 (дата обращения 17.06.2022).
17. *Агеев А.И.* Насколько Россия подготовлена к вызовам XXI века // НГ-ЭНЕРГИЯ от 16.01.2019.

Будзко Владимир Игоревич. Федеральное государственное учреждение «Федеральный исследовательский центр «Информатика и управление» Российской академии наук», г. Москва, Россия. Главный научный сотрудник, доктор технических наук, академик Академии криптографии РФ. Количество печатных работ: более 150 (в т.ч. 5 монографий). Область научных интересов: системный анализ, управление и обработка информации, вычислительные системы и их элементы, математическое и программное обеспечение вычислительных систем, комплексов и компьютерных сетей, методы и системы защиты информации, информационная безопасность, информатика и информационные процессы. E-mail: vbudzko@ipiran.ru.

Меденников Виктор Иванович. Федеральное государственное учреждение «Федеральный исследовательский центр «Информатика и управление» Российской академии наук», г. Москва, Россия. Ведущий научный сотрудник, доктор технических наук, доцент. Количество печатных работ: более 500 (в т.ч. 11 монографий). Область научных интересов: математическое моделирование, информационные технологии, цифровые технологии, сельское хозяйство. E-mail: dommed@mail.ru (ответственный за переписку).

Conditions for the effective application of artificial intelligence technologies in the agro-industrial complex of the EAEU

V.I. Budzko, V.I. Medennikov

Federal Research Center “Computer Science and Control” of Russian Academy of Sciences, Moscow, Russia

Abstract. The solutions for reducing the environmental hazard in agriculture of the EAEU agro-industrial space formating are considered. The mechanism for such space formatting is proposed. It allows resolving the emerging geopolitical, economic, social, and environmental problems. This is a single digital management platform, which includes the possibility of cloud building based on mathematical and ontological modeling, common digital standards (the structure of the subplatform for collecting, storing and integrating operational primary accounting information of all participants in a single database; the structure of the subplatform for technological accounting; the structure of the subplatform of data processing algorithms of the first two subplatforms for the purpose of production management). The use of artificial intelligence technologies will bring the greatest effect and will ensure maximum cross-industry traceability of products and the negative impact of natural and anthropogenic environmental hazards on the environment, on the products of the agro-industrial complex and on the person himself will be minimized.

Keywords: *artificial intelligence, environmental issues, agro-industrial complex, EAEU, digital standards, product traceability.*

DOI: 10.14357/20790279230117

References

1. Degradation worth billions: over 60% of agricultural land in Russia is depleted. URL: https://agroru.com/news/degradatsiya-na-milliardy-v-rossii-istoscheny-svyshe-60-selh-85534.htm (2022), last accessed 2022/03/27.
2. *Krasilnikov, P.* “Overview of soil and land degradation in Eurasia”. Report of the Eurasian Center for Food Security October 3-5, (2017).
3. *Viktor, Medennikov and Alexander, Raikov.* Formation of the Digital Platform for Precision Farming with Mathematical Modeling. CEUR Workshop Proceedings 2790: 114–126, (2020), http://ceur-ws.org/Vol-2790/.
4. *Medennikov, V.* Mathematical modeling of digital platforms and standards for managing the country’s economy, Journal Informatization of Education and Science 3(47): 57-72, (2020).
5. *Medennikov, V.* System analysis of digital ecosystems of manufacturing industries on the example of the agro-industrial complex. Journal Digital Economy, 3: 69-74, (2021).
6. Artificial intelligence in agriculture. URL: https://agropravda.com/news/novye-tech-nologii/11301-iskusstvennyj-intellekt-v-sels-kom-hozjajstve (2022), last accessed 2022/03/27.
7. *Mikulets, Y. and Medennikov, V.* Digital standards are the basis for the integration of digital platforms for the agro-industrial complex and other industries. Journal Bulletin of the Moscow Humanitarian and Economic Institute 1: 208-226, (2021).
8. *Medennikov, V.* Formation of a unified digital platform for agricultural management of the EAEU” Collection of articles of the All-Russian Scientific and Practical Conference “European Union in the Global Economy: Agro-Food Aspect”. Saransk, March 15, 2019: 134-138, (2019).
9. Digitalization of agriculture in Russia lacks data. URL: http://www.iksmedia.ru/news/5533967-Czifrovizacii-selskogo-xozyajstva.html#ixzz6KBD7I-YEP (2022), last accessed 2022/03/27.
10. How to start implementing precision farming in the enterprise. URL: https://smartfarming.ua/ru-blog/kak-nachat-vnedryat-tochnoe-zemledelie-na-predpriyatii (2022), last accessed 2022/03/27.
11. *Galustyan, A.* Five problems that artificial intelligence cannot solve yet, URL: https://rb.ru/opinion/problemy-ii/ (2021), last accessed 2022/03/27.
12. *Budzko, V. and Medennikov, V.* Mathematical modeling of evaluating the effectiveness of using RSD data in precision farming, Procedia Computer Science : 11th, Natal, Rio Grande do Norte, November 10-15, 2020. – Natal, Rio Grande do Norte: 122-129, (2020), https://doi.org/10.1016/j.procs.2021.06.015.
13. *Abrosimov, V.K. and Raikov, A.N.* 2022. Intelligent agricultural robots. – M.: Career Press. - 512 p.
14. *Toluev, Y. and Plankovsky, S.* Modeling and simulation of logistics systems. Kyiv, Millennium, (2009).
15. *Medennikov, V. and Raikov, A.* Optimizing of Product Logistics Digital Transformation with Mathematical Modeling, Journal of Physics: Conference Series : 13, Saint Petersburg, October 06–08, 2020 Saint Petersburg: 012100 (1-9), (2021), https://iopscience.iop.org/article/10.1088/1742-6596/1864/1/012100.
16. *Lenchuk, E.* Digital economy in Russia? Just a second..., URL: https://zen.yandex.ru/media/freeconomy/cifrovaia-ekonomika-v-rossii-sekundochku-5ccc6762a8ac8300b3495949 (2022), last accessed 2022/03/27.
17. *Ageev, A.* To what extent is Russia prepared for the challenges of the 21st century, NG-ENERGIA, (2019), from 01/16/2019.

Budzko V.I. Doctor of Engineering, Professor, Principal Research Scientist. Federal State Institution «Federal Research Center «Computer Science and Control» of Russian Academy of Sciences», 44/2 Vavilova street, Moscow, 119333, Russia. E-mail: vbudzko@ipiran.ru

Medennikov V.I. Doctor of Engineering, leading researcher. Federal State Institution «Federal Research Center «Computer Science and Control» of Russian Academy of Sciences», 44/2 Vavilova street, Moscow, 119333, Russia. E-mail: dommed@mail.ru

Компьютерный анализ текстов

Методы извлечения биомедицинской информации из патентов и научных публикаций (на примере химических соединений)

Н.А. Колпаков[I], А.И. Молодченков[II,III], А.В. Лукин[III]

[I] Московский физико-технический институт, г. Москва, Россия
[II] Федеральное государственное учреждение «Федеральный исследовательский центр «Информатика и управление» Российской академии наук», г. Москва, Россия
[III] Российский университет дружбы народов, г. Москва, Россия

Аннотация. В данной статье предложен алгоритм для решения задачи извлечения информации из биомедицинских патентов и научных публикаций. Предложенный алгоритм основан на методах машинного обучения. Были проведены эксперименты на патентах из базы USPTO. Эксперименты показали, что лучшее качество извлечения показала модель, построенная на основе BioBERT.

Ключевые слова: *машинное обучение, обработка естественного языка, извлечение именованных сущностей, обработка биомедицинских текстов.*

DOI: 10.14357/20790279230118

Введение

С каждым годом число биомедицинских патентов и научных публикаций значительно увеличивается. Зачастую эти тексты не содержат какие-то описательные метаданные, а это, в свою очередь, приводит к большому объёму неструктурированных данных. Следовательно, увеличивается потребность в инструментах, которые бы могли точно извлекать требуемую информацию из таких текстов.

Для извлечения информации из текстов для дальнейшей её обработки можно использовать как подходы машинного обучения, так и алгоритмы, основанные на регулярных выражениях. В работах [1, 2] ключевую роль играют регулярные выражения, и, напротив, в [3, 4] используются достижения области глубокого машинного обучения, в частности модель условных случайных полей. А в [5] используется нейросетевая модель-трансформер, которая при правильной настройке параметров может достаточно неплохо извлекать биомедицинских данные.

Хотя, были созданы инструменты для анализа и взаимодействия с неструктурированными данными, зачастую эти решения основаны на правилах, которые применимы к конкретным обрабатываемым данным. В этой работе мы предлагаем решение для задачи извлечения биомедицинской информации из патентов с помощью регулярных выражений. Таким образом, полученная структурированная информация может быть использована для обучения сложных нейросетевых моделей, которые позволят корректно извлекать информацию из большего числа текстов.

1. Обзор релевантных работ

Существует не так много решений, которые решают поставленную задачу. Зачастую, существующие алгоритмы разработаны для решения большого спектра задач, поэтому они показывают недостаточно высокие результаты для задачи извлечения определений из биомедицинских патентов и научных публикаций.

Например, Jinhyuk Lee и его коллеги представили BioBERT (Bidirectional Encoder Representations from Transformers for Biomedical Text Mining) [5] – нейросетевую модель-трансформер [6], разработанную для автоматической обработки языка биомедицинской области, которая предварительно обучена на больших биомедицинских текстах. Данная модель способна извлекать биомедицинские именованные сущности, биомедицинские отношения в тексте, а также может выдавать ответы на биомедицинские вопросы. BioBERT инициализируется со значениями весовых функций, которые были получены для BERT [7] (данная модель предварительно обучена на текстах из английской Википедии и BooksCorpus), после чего BioBERT был дообучен на биомедицинских текстах (сюда входят аннотации с PubMed и полнотекстовые статьи PMC).

В статье [3] представлен другой подход для решения задач NLP из области биомедицины. CLAMP (Clinical Language Annotation, Modeling, and Processing) для извлечения информации использует как методы, основанные на машинном обучении, так и методы, основанные на правилах. Данный инструментарий позволяет извлекать именованные сущности, разбивать текст на токены, и многое другое. В своей программе авторы используют 3 типа токенизаторов (на выбор):

1) OpenNLP токенизатор [8] на основе машинного обучения,
2) токенизатор на основе разделения слов по заданным символам,
3) токенизатор на основе правил с различными параметрами конфигурации.

А для задачи извлечения именованных сущностей авторы предлагают использовать:

1) алгоритм условных случайных полей (conditional random fields – CRF) [9],
2) алгоритм на основе словаря с большим количеством биомедицинской лексики, собранной из разных ресурсов, таких как UMLS,
3) алгоритм на основе регулярных выражений для объектов с общими шаблонами.

OSCAR4 (Open-Source Chemistry Analysis Routines) [2] – это открытая система для автоматического извлечения химических терминов из научных статей. В основе данной работы лежит распознавание химических веществ на основе регулярных выражений и распознавание на основе словаря заранее заданных слов. Но для распознавания сложных химических соединений (которые состоят из нескольких токенов) – используется модель Маркова максимальной энтропии.

Ещё, есть работа [1], где авторы используют морфологию для извлечения биомедицинских слов. Система распознавания химических объектов состоит из двух подсистем. Первая извлекает химические объекты и помечает их в нормализованном входном документе с использованием словаря заранее заданных слов и морфологического подхода. Основанный на морфологии подход идентифицирует различные элементы в химическом соединении и объединяет их для создания конечного соединения.

Вторая подсистема – извлекает дополнительные химические элементы и распределяет все распознанные объекты по классам соединений, а также имеет такие возможности как расшифровка аббревиатур и исправление орфографических ошибок. Для того, чтобы определить является ли определённая сущность “химической”, авторы собрали статистическую информацию для каждого уникального объекта. Данная информацию используется как последний этап извлечения именованных сущностей и предназначен для классификации извлеченного объекта (либо объект является химическим, либо нет).

Приведённые методы извлекают информацию из биомедицинских текстов в целом – они не направлены на извлечение структур Маркуша [10] (см. Рис. 1).

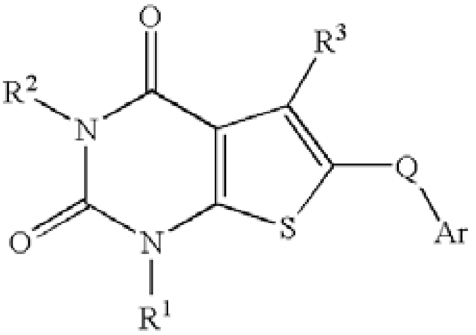
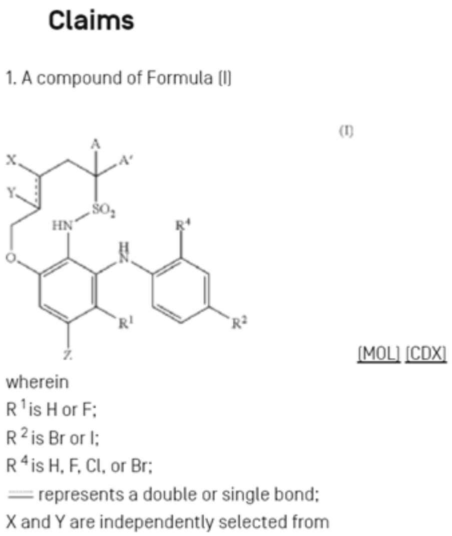


Рис. 1. Пример структуры Маркуша, взят из US Patent 20040171623

2. Постановка задачи

Данные, касающиеся различных биомедицинских патентов, находятся в открытом доступе в различных патентных ведомствах. Патенты обычно имеют четкую структуру, которая включает в себя: название патента, аннотацию, описание, формулы изобретения (Claims) и библиографическую информацию (дата, номер патента, авторы).

Интересующий нас раздел – Claims (см. Рис. 2), содержит описание химических соединений, которые заявлены авторами патента. На это как раз направлена правовая охрана, предоставляемая патентом. Раздел Claims может содержать внутри себя несколько подразделов, которые содержат информацию по разным химическим цепочкам.



Claims

1. A compound of Formula (I)

(I)

[MOL] [CDX]

wherein

R 1 is H or F;

R 2 is Br or I;

R 4 is H, F, Cl, or Br;

== represents a double or single bond;

X and Y are independently selected from

Рис. 2. Пример данных, содержащихся в разделе Claims, взят из US Patent 20120208859

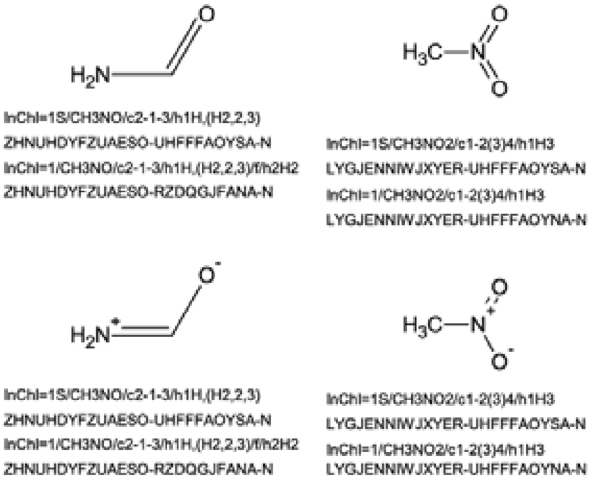
Соединения, представленные в разделе Claims, могут быть описаны с помощью структуры Маркуша [10] (см. Рис. 1). Для того, чтобы найти патенты, у которых структура Маркуша либо такая же, либо схожая, – нужно сравнить эти структуры. Так как структура Маркуша – сетевая модель, то сравнение напрямую таких моделей – очень ресурсоёмкий процесс. Поэтому, зачастую используют так называемые fingerprints, которые отражают в себе информацию, представленную в структурах Маркуша. Но перед этим нужно извлечь информацию, которая входит в такие структуры, на что и направлена данная работа.

Табл. 1.

Примеры химических соединений

Название соединения	Молекулярная формула
nitrogen monoxide	NO
glucose	$C_6H_{12}O_6$
copper (II) sulfate	$CuSO_4$
carbon dioxide	CO_2
dichlorine heptoxide	Cl_2O_7

Задача состоит в извлечении из раздела Claims химических соединений (Табл. 1), названий переменных (вместо которых могут быть подставлены различные значения), химических элементов, формул и InChI кодов [11] (Рис. 3) с целью преобразования данной текстовой информации в некоторую структуру формального представления.



O
H2N
InChI=1S/CH3NO/c2-1-3/h1H,(H2,2,3)
ZHNUHDYFZUAESO-UHFFFAOYSA-N
InChI=1/CH3NO/c2-1-3/h1H,(H2,2,3)/f/h2H2
ZHNUHDYFZUAESO-RZDQGJFANA-N
O
H3C
N
O
InChI=1S/CH3NO2/c1-2(3)4/h1H3
LYGJENNIWJXYER-UHFFFAOYSA-N
InChI=1/CH3NO2/c1-2(3)4/h1H3
LYGJENNIWJXYER-UHFFFAOYNA-N
O-
H2N+
InChI=1S/CH3NO/c2-1-3/h1H,(H2,2,3)
ZHNUHDYFZUAESO-UHFFFAOYSA-N
InChI=1/CH3NO/c2-1-3/h1H,(H2,2,3)/f/h2H2
ZHNUHDYFZUAESO-RZDQGJFANA-N
O
H3C
N+
O-
InChI=1S/CH3NO2/c1-2(3)4/h1H3
LYGJENNIWJXYER-UHFFFAOYSA-N
InChI=1/CH3NO2/c1-2(3)4/h1H3
LYGJENNIWJXYER-UHFFFAOYNA-N

Рис. 3. Примеры InChI кодов [11]

В теоретико-множественной аннотации задачу можно сформулировать следующим образом: имеются патенты и научные публикации X, где каждый элемент x $\in X$ представлен в виде x $= x_1, \ldots, x_n$ ($x_1, \ldots, x_n$ – последовательность слов (токенов)), и задано множество классов $Y = (y_1, \ldots, y_5)$, где:

y_1 – номер Claim,

y_2 – переменная, к которой ищем описание,

y_3 – описание переменной,

y_4 – ссылка на другой Claim,

y_5 – в случаях, если токен не соответствует $y_1, \ldots, y_4$.

Необходимо построить отображение F, которое бы сопоставляло каждому элементу x $\in X$ – соответствующий элемент y $\in Y$.

Задача извлечения информации из текста представляет собой поиск и классификацию именованных сущностей (Named Entity Recognition), имеющихся в неструктурированном тексте, по заранее заданным категориям. Именованная сущность – n-грамма в тексте, для которой определена категория (класс, метка).

3. Описание метода

Алгоритм извлечения информации из текстов можно разбить на следующие шаги:

1) составление набора данных,
2) предварительная обработка входных данных,
3) векторизация данных и извлечение признаков,
4) обучение моделей для извлечения необходимой информации из текстов.

Составление набора данных включает также включает в себя автоматизированную разметку токенов. В предварительную обработку данных входит нормализация и токенизация входных данных. Схема алгоритма приведена на Рис. 4.

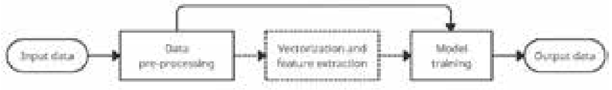


Рис. 4. Схема предложенного алгоритма извлечения информации из текстов

3.1. Составление набора данных

Данные, с которыми мы работали, взяты из базы USPTO [12]. Все данные изначально представлены в XML файлах, которые содержат структурированную информацию о патентах: описание, аннотацию, библиографические данные и Claims. Для разработки алгоритма из файлов берется только раздел Claims.

3.2. Предварительная обработка данных

Первым этапом обработки данных является извлечение из имеющихся данных – раздела Claims. Так как такие данные имеют схожее оформление, то извлечение выполняются с помощью регулярных выражений.

После извлечения Claims – необходимо подготовить данные для дальнейшей работы. Для этого выполняется следующая нормализация строк:

1) Удаляются лишние пробелы в начале и конце строк.
2) Пустые строки тоже удаляются.
3) Каждые строки разбиваются таким образом, чтобы в них содержалось только одно описание переменных. Это выполняется с помощью поиска в каждой строке следующей конструкции: *… variables … definition_verb … definitions … definition_end_symbol*. При этом, учитывается ситуация, когда на этой же строке может быть приведено описание вложенных переменных. Например, “Z is OR3, wherein R3 is C1-C6 alkyl”. В этом случае строка не разбивается.
4) Если в строке не встретился *definition_end_symbol*, то строки объединяются до тех пор, пока не найдется нужный символ.
5) Если строка является начальной для Claim, но номера Claim указаны через тире, то последующее содержимое копируется для каждого номера Claim из указанного промежутка.

Затем полученные строки группируются по Claim. Все описанные выше действия по нормализации строк, тоже выполняются с помощью регулярных выражений.

Применение нормализации позволит обучить модель на небольшой выборке более качественно, а также повысит её точность. А группировка и разбиение строк упростят последующую разметку данных.

Следующим этапом является присвоение каждому токену метки из возможных:

- CLAIM – номер Claim,
- VAR – переменная, к которой ищем описание; описание к этой переменной подставляется только к последнему месту, где она была упомянута до встречи этой самой переменной,
- VAR-ALL – переменная, к которой ищем описание; описание к этой переменной подставляется во все места, где она была упомянута,
- DEF – описание переменной,
- REF – ссылка на другой Claim,
- O – в случае, если токену не присвоена ни одна из вышеперечисленных меток.

Присвоение токенам соответствующей метки выполняется с помощью средств разметки ФИЦ ИУ РАН. Результатом работы этих средств являются данные, содержащие токен, его метку, номер строки, где он был найден и уникальный номер Claim.

3.3. Векторизация данных и извлечение признаков

Так как не все модели классификации в качестве входных данных принимают строковые значения данных, то необходимо векторизовать такие признаки. К ним относятся токены и соответствующие метки.

Если каждой уникальной метке сопоставляется число, то с токенами дело обстоит совсем иначе. Для каждого токена строится вектор размерности 100 с помощью модели Word2Vec [13, 14] для получения векторных представлений слов естественного языка.

Word2Vec была обучена на собранном наборе данных. Обучение происходило 10 эпох, с размером скользящего окна равным 8.

Чтобы в дальнейшем обучить модель машинного обучения – необходимо объединить токены в списки на основе принадлежности к Claims, а затем подать эти списки на вход Word2Vec. Результатом работы такой модели будет сопоставление каждому токену его векторного представления.

Некоторые алгоритмы машинного обучения, например основанные на Conditional Random Fields, будут лучше работать с признаками, содержащими информацию о соседних токенах, относительно рассматриваемого.

Поэтому, ещё одним способом представления данных, подаваемых на вход таким моделям, – является сопоставление каждому токену набора признаков. Этими признаками являются:

- сами токены,

- последние 2–3 символа токена,
- флаг, начинается ли токен с заглавной буквы,
- флаг, является ли токен числом,
- флаг, содержит ли токен только заглавные буквы,
- информация о соседних токенах (соседний токен и 3 флага, как в предыдущих пунктах).

3.4. Обучение моделей

В качестве методов классификации, которые бы на основе размеченных и векторизованных данных присваивали бы ранее неизвестным данным метки, – использовались как стандартные методы машинного обучения (Support Vector Machine [15], Conditional Random Fields [9]), так и модели глубокого обучения (Stanford NER [16], BERT [7], BioBERT [5]), которые уже заранее предобучены.

Дообучение BioBERT, BERT и Stanford NER производилось на данных, полученных в пункте 3.2 – токенах, метках и номерах Claims. Для BioBERT и BERT, чтобы не происходило переобучение, число эпох было выбрано равным 5 (см Рис. 5 и Рис. 6).

Метод опорных векторов (SVM) обучался с нуля на векторизованных данных, а метод условных случайных полей (CRF) – на данных, полученных в пункте 3.3.

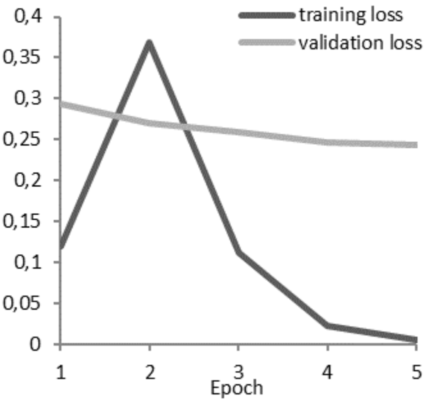


Рис. 5. График *Epoch vs Loss* для модели BioBERT.

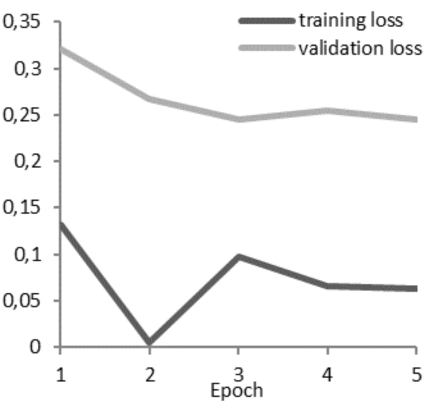


Рис. 6. График *Epoch vs Loss* для модели BERT.

4. Результаты экспериментов

В рамках данной работы была проведена серия экспериментов для решения поставленной задачи классификации. Эксперименты проводились на 100 документах с более чем 1700 Claims. Обучающая выборка состояла из 70 документов, а валидационная – из 30.

Для сравнения результатов использовались стандартные метрики качества: *precision (точность)*, *recall* (полнота) and *F1-score* [17]. Давайте, рассмотрим их более детально.

Для начала рассмотрим, что такое TP, FP и FN:

- TP – число токенов, которым классификатор присвоил правильную меток,
- FP – число токенов, которые имеют метку O, но классификатор присвоил им другую метку,
- FN – число токенов, которые имеют определённую метку (не O), но классификатор отнёс их к другой группе.

Accuracy (точность) – доля токенов, действительно принадлежащих конкретному классу, относительно всех токенов, которым классификатор присвоил такую метку класса. Данная метрика вычисляется по уравнению (1).

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

Recall (полнота) – доля токенов, которым классификатор присвоил конкретную метку класса, относительно всех токенов, имеющих эту метку. Данная метрика вычисляется по уравнению (2).

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

F1-score – среднее гармоническое значение точности и полноты. Данная метрика вычисляется по уравнению (3).

$$F1\text{-}score = \frac{2 * Precision * Recall}{Precision + Recall} = \frac{2 * TP}{2 * TP + FP + FN} \tag{3}$$

Результаты классификации на тестовых данных приведены в Табл. 2.

Табл. 2

Значения метрик для различных методов классификации.

Название модели	Precision	Recall	F1-score
SVM	0.5276	0.6340	0.5675
CRF	0.6701	0.6358	0.6378
Stanford NER	0.7530	0.8488	0.7981
BERT	0.8437	0.8978	0.8699
BioBERT	**0.8467**	**0.9012**	**0.8731**

Из проведённых экспериментов видно, что классические методы машинного обучения показывают результаты намного хуже, чем предварительно обученные модели глубокого обучения, которые, в свою очередь, классифицируют токены на достаточно хорошем уровне.

Заключение

В статье описан метод решения задачи извлечения информации из биомедицинских текстов для дальнейшей ее обработки. Этот метод позволяет извлекать описание химических соединений, которые заявлены авторами патентов. Были обучены модели машинного обучения, такие как SVM, CRF, Stanford NER, BERT и BioBERT, на которых впоследствии проводились эксперименты.

В дальнейшем планируется преобразовать полученные данные в формат InChI кодов и написать fingerprints которые соответствуют структурам Маркуша, заявленным авторами патентов. Также планируется провести ещё серию экспериментов для улучшения качества извлечения информации из текстов.

Литература

1. *Akhondi, S., Rey, H., Schwörer, M., Maier, M., Toomey, J., Nau, H., Ilchmann, G., Sheehan, M., Irmer, M., Bobach, C., Doornenbal, M., Gregory and M., Kors, J.* Automatic identification of relevant chemical compounds from patents. Database: the journal of biological databases and curation. 2019. Vol. 1. P. 1–14.
2. *Jessop, D., Adams, S., Willighagen, E., Hawizy, L. and Murray-Rust, P.* OSCAR4: A flexible architecture for chemical textmining. Journal of cheminformatics. 2011. Vol. 3. No. 1. P. 1–12.
3. *Soysal, E., Wang, J., Jiang, M., Wu, Y., Pakhomov, S., Liu, H. and Qi, W.* CLAMP – a toolkit for efficiently building customized clinical natural language processing pipelines. Journal of the American Medical Informatics Association: JAMIA. 2018. Vol. 25. No. 3. P. 331–336.
4. *Swain, M. and Cole, J.* 2016. ChemDataExtractor: A Toolkit for Automated Extraction of Chemical Information from the Scientific Literature. Journal of Chemical Information and Modeling. 2016. Vol. 56. No. 10. P. 1894–1904.
5. *Jinhyuk, L., Wonjin, Y., Sungdong, K., Donghyeon, K., Sunkyu, K., Chan, H. S. and Jaewoo, K.* BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. 2019. Vol. 36. No. 4. P. 1234–1240.
6. *Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, L. and Polosukhin, I.* Attention Is All You Need. Advances in Neural Information Processing Systems. 2017. Vol. 30. P. 5998–6008.
7. *Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K.* Bert: pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2019. Vol. 1. P. 4171–4186.
8. The OpenNLP Project. Available at: http://opennlp.apache.org (дата обращения 20.02.2022).
9. CRFsuite: a Fast Implementation of Conditional Random Fields (CRFs). Available at: http://www.chokkan.org/software/crfsuite/ (дата обращения 20.02.2022).
10. *Barnard, J.* A comparison of different approaches to Markush structure handling. Journal of Chemical Information and Computer Sciences. 1991. Vol. 31. No. 1. P. 64–68.
11. *Heller, S., McNaught, A., Pletnev, I., Stein, S. and Tchekhovskoi, D.* The IUPAC International Chemical Identifier. Journal of Cheminformatics. 2015. Vol. 7. P. 1–34.
12. USPTO. Available at: https://www.uspto.gov/patents (дата обращения 20.02.2022).
13. *Mikolov, T., Chen, K., Corrado, G. and Dean, J.* Efficient Estimation of Word Representations in Vector Space. Proceedings of Workshop at ICLR. 2013. P. 1–12.
14. *Mikolov, T., Yih, W.-T. and Zweig, G.* Linguistic regularities in continuous space word representations. Proceedings of NAACL-HLT. 2013. P. 746–751.
15. *Cortes, C. and Vapnik, V.* Support-vector networks. Machine Learning. 1995. Vol. 20. No. 3. P. 273–297.
16. *Finkel, J., Grenager, T. and Manning, C.* Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005). 2005. P. 363–370.
17. *Mitchell, T.* Machine Learning. Нью-Йорк: McGraw-Hill, 1997. 432 с.

Колпаков Николай Алексеевич. Федеральное государственное автономное образовательное учреждение высшего образования “Московский физико-технический институт (национальный исследовательский университет)” (МФТИ, Физтех), г. Москва, Россия. Бакалавр. Количество печатных работ: 1. Область научных интересов: машинное обучение, глубокое обучение, извлечение именованных сущностей, обработка естественного языка. E-mail: kolpakov.na@phystech.edu

Молодченков Алексей Игоревич. Федеральное государственное учреждение “Федеральный исследовательский центр “Информатика и управление” Российской академии наук”, г. Москва, Россия. Кандидат технических наук. Количество печатных работ: 96. Область научных интересов: искусственный интеллект, базы знаний, извлечение информации, медицина. E-mail: aim@tesyan.ru (Ответственный за переписку)

Лукин Антон. Федеральное государственное автономное образовательное учреждение высшего образования “Российский университет дружбы народов” (РУДН), г. Москва, Россия. Учёная степень. Количество печатных работ: 10. Область научных интересов: искусственный интеллект, анализ текстов. E-mail: antonvlukin@gmail.com

Methods of extracting biomedical information from patents and scientific publications (on the example of chemical compounds)

N.A. Kolpakov[I], A.I. Molodchenkov[II,III,] A.V. Lukin[III]

[I] Moscow Institute of Physics and Technology, Moscow, Russia

[II] Federal research center “Computer science and control” of Russian Academy of Sciences, Moscow, Russia

[III] Peoples’ Friendship University of Russia, Moscow, Russia

Abstract. This article proposes an algorithm for solving the problem of extracting information from biomedical patents and scientific publications. The introduced algorithm is based on machine learning methods. Experiments were carried out on patents from the USPTO database. Experiments have shown that the best extraction quality was achieved by a model based on BioBERT.

Keywords: *machine learning, natural language processing, named entity recognition, biomedical texts processing.*

DOI: 10.14357/20790279230118

References

1. *Akhondi, S., Rey, H., Schwörer, M., Maier, M., Toomey, J., Nau, H., Ilchmann, G., Sheehan, M., Irmer, M., Bobach, C., Doornenbal, M., Gregory and M., Kors, J.* 2019. Automatic identification of relevant chemical compounds from patents. Database: the journal of biological databases and curation, vol. 1, pp. 1–14.
2. *Jessop, D., Adams, S., Willighagen, E., Hawizy, L. and Murray-Rust, P.* 2011. OSCAR4: A flexible architecture for chemical textmining. Journal of cheminformatics, vol. 3, no. 1, pp. 1–12.
3. *Soysal, E., Wang, J., Jiang, M., Wu, Y., Pakhomov, S., Liu, H. and Qi, W.* 2018. CLAMP – a toolkit for efficiently building customized clinical natural language processing pipelines. Journal of the American Medical Informatics Association: JAMIA, vol. 25, no. 3, pp. 331–336.
4. *Swain, M. and Cole, J.* 2016. ChemDataExtractor: A Toolkit for Automated Extraction of Chemical Information from the Scientific Literature. Journal of Chemical Information and Modeling, vol. 56, no. 10, pp. 1894–1904.
5. *Jinhyuk, L., Wonjin, Y., Sungdong, K., Donghyeon, K., Sunkyu, K., Chan, H. S. and Jaewoo, K.* 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, vol. 36, no. 4, pp. 1234–1240.
6. *Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, L. and Polosukhin, I.* 2017. Attention Is All You Need. Advances in Neural Information Processing Systems, vol. 30, pp. 5998–6008.
7. *Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K.* 2019. Bert: pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, vol. 1, pp. 4171–4186.
8. The OpenNLP Project. Available at: http://opennlp.apache.org (accessed February 20, 2022).

9. CRFsuite: a Fast Implementation of Conditional Random Fields (CRFs). Available at: http://www.chokkan.org/software/crfsuite/ (accessed February 20, 2022).
10. *Barnard, J.* 1991. A comparison of different approaches to Markush structure handling. Journal of Chemical Information and Computer Sciences, vol. 31, no. 1, pp. 64–68.
11. *Heller, S., McNaught, A., Pletnev, I., Stein, S. and Tchekhovskoi, D.* 2015. The IUPAC International Chemical Identifier. Journal of Cheminformatics, vol. 7, pp. 1–34.
12. USPTO. Available at: https://www.uspto.gov/patents (accessed February 20, 2022).
13. *Mikolov, T., Chen, K., Corrado, G. and Dean, J.* 2013. Efficient Estimation of Word Representations in Vector Space. Proceedings of Workshop at ICLR, pp. 1–12.
14. *Mikolov, T., Yih, W.-T. and Zweig, G.* 2013. Linguistic regularities in continuous space word representations. Proceedings of NAACL-HLT, pp. 746–751.
15. *Cortes, C. and Vapnik, V.* 1995. Support-vector networks. Machine Learning, vol. 20, no. 3, pp. 273–297.
16. *Finkel, J., Grenager, T. and Manning, C.* 2005. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005), pp. 363–370.
17. *Mitchell, T.* 1997. Machine Learning. New York: McGraw-Hill. 432 p.

Kolpakov N.A. Moscow Institute of Physics and Technology, 1A, building 1, Kerch str., Moscow, 117303, Russia, e-mail: kolpakov.na@phystech.edu

Molodchenkov A.I. Federal Research Center “Computer Science and Control” of Sciences, 44/2 Vavilova str., Moscow, 119333, Russia, e-mail: aim@tesyan.ru

Lukin A.V. Peoples’ Friendship University of Russia, 6, Miklukho-Maklaya str., Moscow, 117198, Russia, e-mail: antonvlukin@gmail.com

Sentence splitters benchmark

A.P. Zavyalova, P.A. Martynyuk, R.S. Samarev

Bauman Moscow State Technical University, Moscow, Russia

Abstract. There are multiple implementations of text into sentences splitters including open source libraries and tools. But the quality of segmentation and the performance of each segmentation tool are very different. Moreover, it is convenient for NLP developers to have all libraries written in the same programming language, except when using some kind of integration programming language. This paper considers two aspects - building a uniform framework and estimating language features of the modern and popular programming language Julia from one side. And the performance estimation of sentence splitting libraries as is. The paper contains detailed performance results, samples of texts after splitting, and a list of some typical issues related to sentence splitting.

Keywords: *segmentation, sentence, splitting, NLP, Julia language, benchmark, text analysis.*

DOI: 10.14357/20790279230119

Introduction

Segmentation of text or splitting linear text into fragments is one of the fundamental operations required for most applied NLP tasks. It refers to one of the stages of natural language text processing - the pre-syntactic analysis of the text[4].

Despite a tendency to replace traditional methods of NLP with neural-net-based ones in the last years, where document analysis might be done without intermediate stages, traditional methods are still used in some cases. E.g. when we need to calculate statistics of words, parts of speech in a document, do some specific markup, etc. There are multiple implementations of text into sentences splitters including open source libraries and tools. But the quality of segmentation and the performance of each segmentation tool are very different.

Also, we need to take into account a technological stack that is used by developers of NLP applications. In most cases, developers prefer to use libraries with API in the same language which is used in their applications development.

Let's take a look at some of the most commonly used segmentation tools.

1. Segmentation tools review

1.1. NLTK

NLTK is a set of natural language analysis tools written in Python language. It includes a set of libraries for classification, tokenization, word stem finding, tagging, parsing, and semantic text analysis. NLTK supports many languages depending on the specific task. This tool implements the segmentation method suggested by S. Bird, etc. [3] using unsupervised learning (learning algorithm "without a teacher") to build a model of abbreviated words, phrases, and words that begin sentences and then find the supply boundary.

1.2. Stanford CoreNLP

CoreNLP is a set of NLP analysis tools written in the Java language. CoreNLP allows users to obtain linguistic annotations for text, including tokens, sentence boundaries, and many others. Currently supports 6 languages: Arabic, Chinese, English, French, German and Spanish. This software implements a combined segmentation method for text segmentation. A combined method means the combination of machine learning and rule-based methods. Text processing suggested by C. Manning, etc. [11] executes in the form of a pipeline, at each stage in which the user receives linguistic annotations. Each stage fulfills its function. For example, a pipeline can consist of token boundaries, parts of speech, named entities, and sentence segmentation. Each of these stages can be implemented in different ways.

1.3. SpaCy

SpaCy is a set of advanced word processing tools written in Python and Cython programming languages. Unlike NLTK, which is widely used for teaching and research, SpaCy focuses on providing software for production use. With its internal machine learning library “Thinc”, SpaCy supports the connection of statistical models trained by popular machine learning libraries: TensorFlow, PyTorch, and MXNet. SpaCy provides models for part-of-speech tagging, dependency analysis, text segmentation, and named entity recognition. Out-of-the-box statistical models for these tasks are available in 17 languages, including English, Portuguese, Spanish, Russian, and Chinese. There is also a multilingual NER model. Additional tokenization support for over 65 languages allows users to train models on their own datasets. This tool implements two segmentation methods: a method based on heuristic rules and a method based on machine learning, namely “decision trees”. The heuristic rules code is in the SpaCy documentation [9]. The Dependency parser uses a variant of the non-monotonic arc-eager transition system suggested by M. Honnibal, etc. [8], with the addition of a “break” transition to performing the sentence segmentation. The pseudo-projective dependency transformation suggested by J. Nivre, etc. [12] is used so that the parser can predict non-projective parsing.

1.4. Apache OpenNLP

Apache OpenNLP is a machine learning-based toolkit for natural language processing. OpenNLP supports the most common NLP tasks, such as tokenization, sentence segmentation, and others. OpenNLP does not support languages out of the box. The framework can be used to train a model for any language. However, there are adapted models for Danish, German, English, Spanish, Dutch, Portuguese, and Sami. This tool implements a sentence segmentation method based on machine learning, namely unsupervised learning [2].

1.5. WordTokenizers.jl

Apache OpenNLP is a machine learning-based toolkit for natural language processing. OpenNLP supports the most common NLP tasks, such as tokenization, sentence segmentation, and others. OpenNLP does not support languages out of the box. The framework can be used to train a model for any language. However, there are adapted models for Danish, German, English, Spanish, Dutch, Portuguese, and Sami. This tool implements a sentence segmentation method based on machine learning, namely unsupervised learning [14].

1.6. Sentencize.jl

This package is also written in Julia and re-implements the Python-based package sentence-splitter [1]. The sentence-splitter package is a Python implementation of the Lingua :: Sentence module [13] – a Perl-based extension for breaking text paragraphs into sentences. Sentencize.jl supports the following languages: Catalan, Czech, Danish, Dutch, English, Finnish, French, German, Greek, Hungarian, Icelandic, Italian, Latvian, Lithuanian, Norwegian (Bokmål), Polish, Portuguese, Romanian, Russian, Slovak, Slovenian, Spanish, Swedish, Turkish. This tool implements a sentence segmentation method based on heuristic rules suggested by P. Koehn, etc. [10].

1.7. Outcomes of the review

Heuristic methods are still widely used in text processing. They are easy to use and do not require significant memory resources. The advantage of these methods is also the stability and predictability of their work. The supply boundary is determined by matching against a set of rules. However, in texts with a special arrangement of punctuation marks (e-mail and physical addresses), mistakes might occur. The disadvantage of these methods is difficult to modify the rules in case of significant changes. Machine learning methods are more delicate than rule-based methods. They are able to find places with a special arrangement of punctuation marks, as well as various mistakes, and avoid them when splitting the text into sentences. However, when using the wrong training sample, the behavior of the sentence splitter in a given text is poorly predictable. The advantage of modern pre-trained neural networks over all the above is context dependence. However, as with statistical methods, the behavior depends on the training sample and is poorly predictable, and the training costs may not be justified in relation to the quality of the text splitting. Combined methods should take into account the shortcomings of machine learning and rule-based methods. But these combined methods require additional RAM resources, which can slow down the process. Let’s test all the above segmentation tools to find the best one.

2. Benchmark

2.1. Julia as an integration platform

Julia language is available for production use since 2018. Then, the community is continuously growing up [7]. Julia can still be called a new language, but it already has impressive functionality. For example, even a new implementation of old libraries and algorithms to solve the problem of text segmentation is mentioned above.

Julia offers packages designed for high performance from the beginning. In machine learning, and natural language processing an immense advantage of Julia is that most packages with similar functionality were developed after the creation of popular Python, Java, and R libraries. But, at the same time, Julia developed much later compared with other popular language stacks.

For many years, a common approach for development libraries that are applicable for multiple languages use was the way of building C/C++ based binary compiled dynamic libraries with C-style exported names. These libraries might be used in scripting languages with appropriate wrapper code prepared manually or with tools like SWIG.

In the case of Java, there is an interface JNI for calling this kind of library. And there is JSR 223 Java Scripting API with multiple 3-rd party execution modules for exact languages to run a script or code snippet. At the same time, a programmer has to do a lot of additional operations to even activate non-Java code and transfer data to the code e.g. in Python. In the case of Python, as a result of the low performance of baseline implementation of CPython, there is the fragmentation of the language dialects with Cython, PyPy, NIM, etc. And, there are the same issues with additional programmers' work to integrate any other language into a Python project with something other than a binary dynamic library. And, even after 30 years of development, the main issue of Python is low performance and a ponderous toolkit to force a Python code to be a production-applicable application.

Key developers of Julia paid attention to integration features and implemented a way to execute other language scripts with their libraries rather than just giving a binary interface. And, moreover, even Julia code might be integrated into other applications with Julia Embedded API and a system image built on a Julia code.

This allowed us to speak about integration possibilities of other languages and libraries into a Julia code, and take into account all the advantages and disadvantages of all the mentioned above packages, as well as use ready-made solutions or their parts written in other languages. Let's look at some of its integration features.

Julia can initially (without any "glue" code, code generation, or compilation) [6] directly call the C and Fortran libraries (fig. 1).

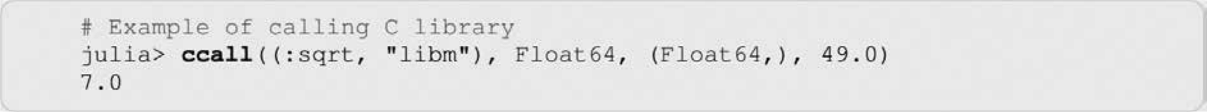
Also, there are special packages to call Python, R, Java code. The following sample uses the package PyCall (fig. 2). It allows us to call Python functions and even to write Python code inside Julia programmers.

JavaCall allows only to call Java packages from within Julia code (fig. 3). First of all, we need to initialize the Java Virtual Machine before we can call any other functions in this module. After that, we get access to all the functions of this package, along with the Java program.

In this benchmark, we take into account all the Julia possibilities making Julia a perfect tool for comparison libraries in different languages under equal conditions.

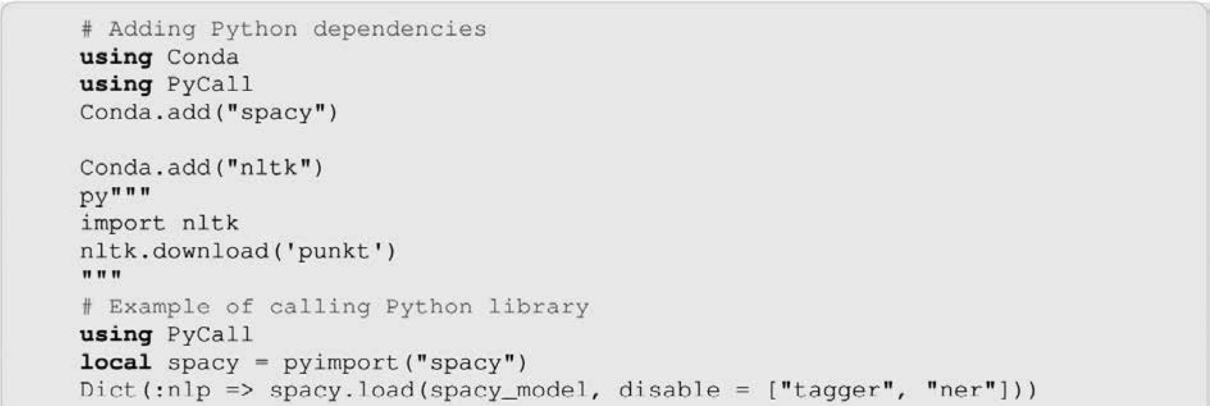
2.2. Benchmark details

The idea of the benchmark is to compare text splitting into sentences with the segmentation tools mentioned above using the same set of text and a reference markup of sentences. We illustrate the principle of operation in figure 4. In the input section, we have a marked-up dataset, on the basis of which we form plain text with and without markup. Next, we split the text without markup into sentences using each segmentation tool. After that, we make a comparison of the obtained sentences with its reference. In the result



```
# Example of calling C library
julia> ccall((:sqrt, "libm"), Float64, (Float64,), 49.0)
7.0
```

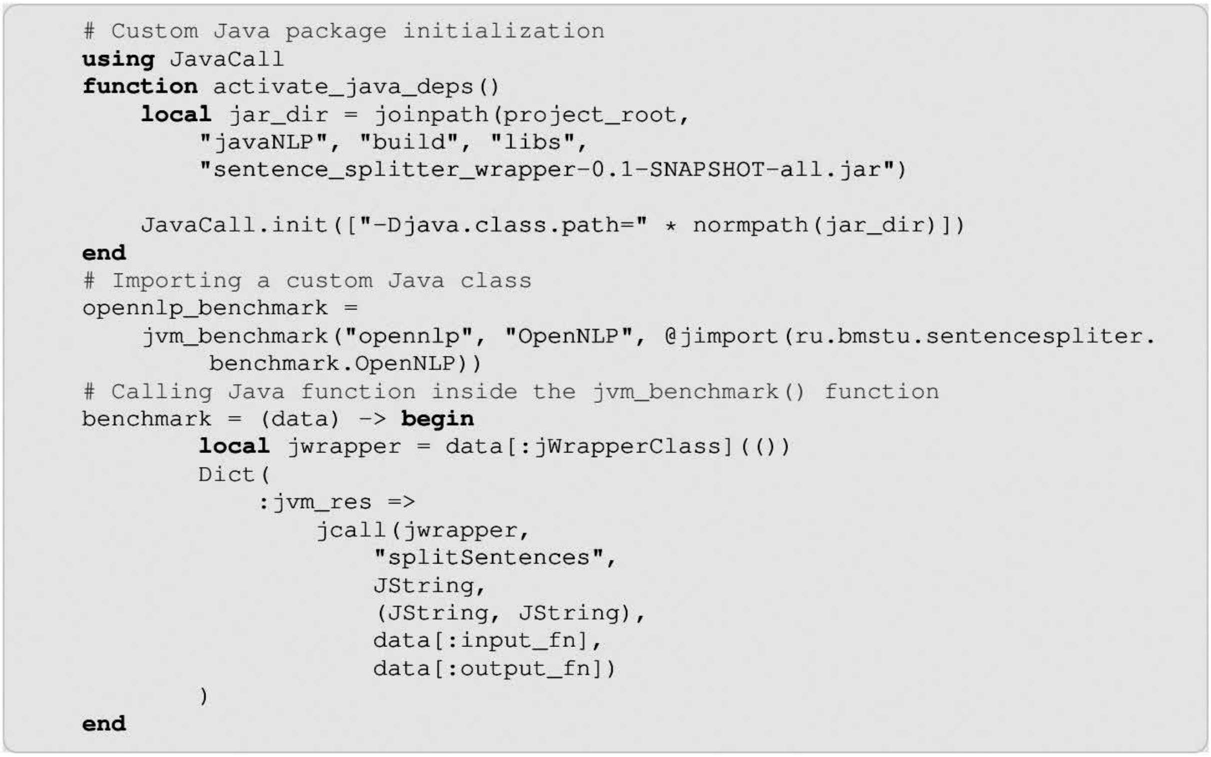
Fig. 1. Example of calling C library



```
# Adding Python dependencies
using Conda
using PyCall
Conda.add("spacy")

Conda.add("nltk")
py"""
import nltk
nltk.download('punkt')
"""
# Example of calling Python library
using PyCall
local spacy = pyimport("spacy")
Dict(:nlp => spacy.load(spacy_model, disable = ["tagger", "ner"]))
```

Fig. 2. Example of calling Python library



```
# Custom Java package initialization
using JavaCall
function activate_java_deps()
    local jar_dir = joinpath(project_root,
        "javaNLP", "build", "libs",
        "sentence_splitter_wrapper-0.1-SNAPSHOT-all.jar")

    JavaCall.init(["-Djava.class.path=" * normpath(jar_dir)])
end
# Importing a custom Java class
opennlp_benchmark =
    jvm_benchmark("opennlp", "OpenNLP", @jimport(ru.bmstu.sentencespliter.
         benchmark.OpenNLP))
# Calling Java function inside the jvm_benchmark() function
benchmark = (data) -> begin
        local jwrapper = data[:jWrapperClass](())
        Dict(
            :jvm_res =>
                jcall(jwrapper,
                    "splitSentences",
                    JString,
                    (JString, JString),
                    data[:input_fn],
                    data[:output_fn])
        )
end
```

Fig. 3. Example of calling Java library

section, we calculate f-measures and the splitting performance time.

We illustrate the benchmark architecture as the package diagram in figure 6. Here you can see how Julia, Python, and Java packages from its libraries connect with each other. SentenceSplitterBenchmark.jl is our package written in Julia language only. We highlight other packages belonging to different programming languages.

The benchmark framework developed in this work gives a simple way to describe a frame for a new library or algorithm for testing. Any specific frame is stored in the structure like this: (fig. 5)

For each segmentation tool, its own object with the described structure is created. That object describes a function for doing some work before starting the benchmark, a function for running the benchmark, and a function for collecting results if these are not available directly. That description is looking like a declarative form.

It was decided to measure the performance (speed of each tool) using BenchmarkTools.jl [5]. This solution will solve problems with launch heterogeneity by averaging measurements.

In the example above (fig. 8), let's pay attention to the fact that the teardown is set in such a way that running the JVM does not affect the result.

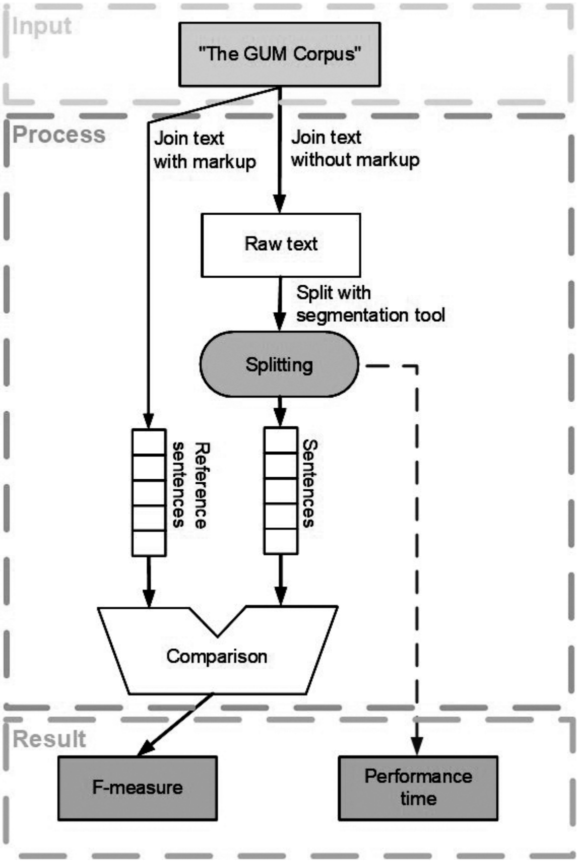
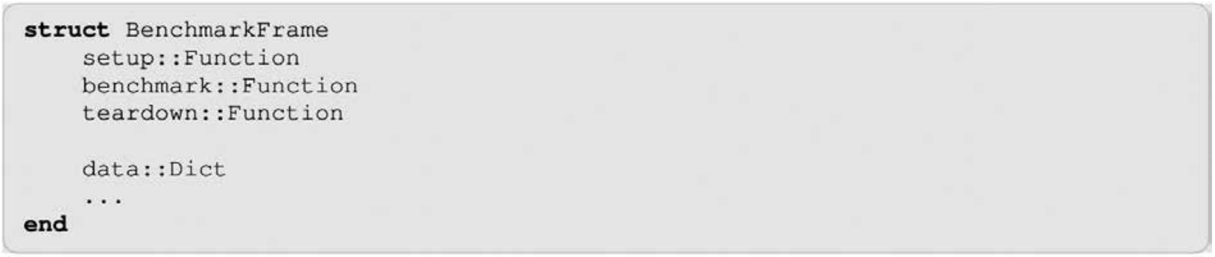


Fig. 4. Principle of benchmark operation



```
struct BenchmarkFrame
    setup::Function
    benchmark::Function
    teardown::Function

    data::Dict
    ...
end
```

Fig. 5. BrenchmarkFrame structure

Then we can call all benchmark objects of segmentation tools sequentially via evaluate_list() with specifying the number of samples and the benchmark duration after that the trial will be terminated(fig.7). Parameters samples = 100, seconds = 40 passed directly into the BenchmarkTools.jl module which is providing running of tests and collecting a stable execution statistics.

2.3. Setup

Let's compare the segmentation tools for the problem of splitting the text into sentences according to the following criteria:

1. Execution performance and required resources.
2. Quality of text segmentation.
3. Errors of text segmentation.

It was decided to measure the performance (speed of each tool) using the BenchmarkTools.jl [5]. Testing will be performed on 5840 sentences from "The GUM Corpus" [16].

GUM stands for Georgetown University Multilayer Corpus, a corpus of English texts with different text types. This corpus consists of interviews, news, travel guides, how-to guides, academic writing, biographies, fiction, online forum discussions, spontaneous face-to-face conversations, political speeches, textbooks, and vlogs. Such a variety of texts allows us to test segmentation close to natural conditions when we don't know what is the input text. For example, one sentence from the GUM Corpus is presented in fig.9

Each token (word, punctuation mark) has its annotation. Our task was to test sentence segmentation, so we use GUM annotation only as a reference for sentence boundaries. Thus, we will combine all the sentences of the corpus into plain text without line breaks (punctuation will remain) and compare the splitting using each of the tools with the reference markup.

We will take average values of the time to obtain the execution performance of tools.

The task of text segmentation is the task of classification (hyphenation might be present or not). The effectiveness of the text segmentation tool can be numerically assessed by the quality of predictions for the

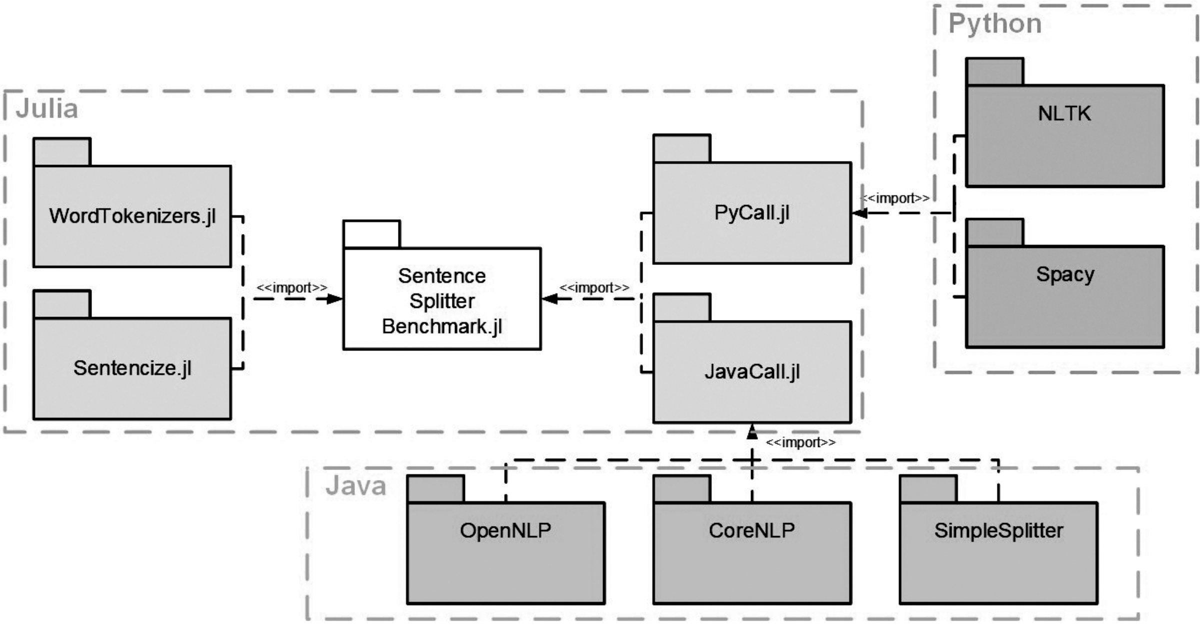
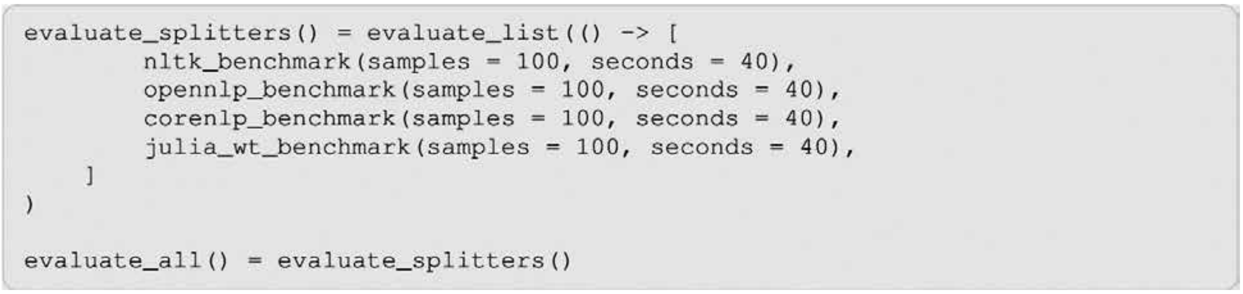


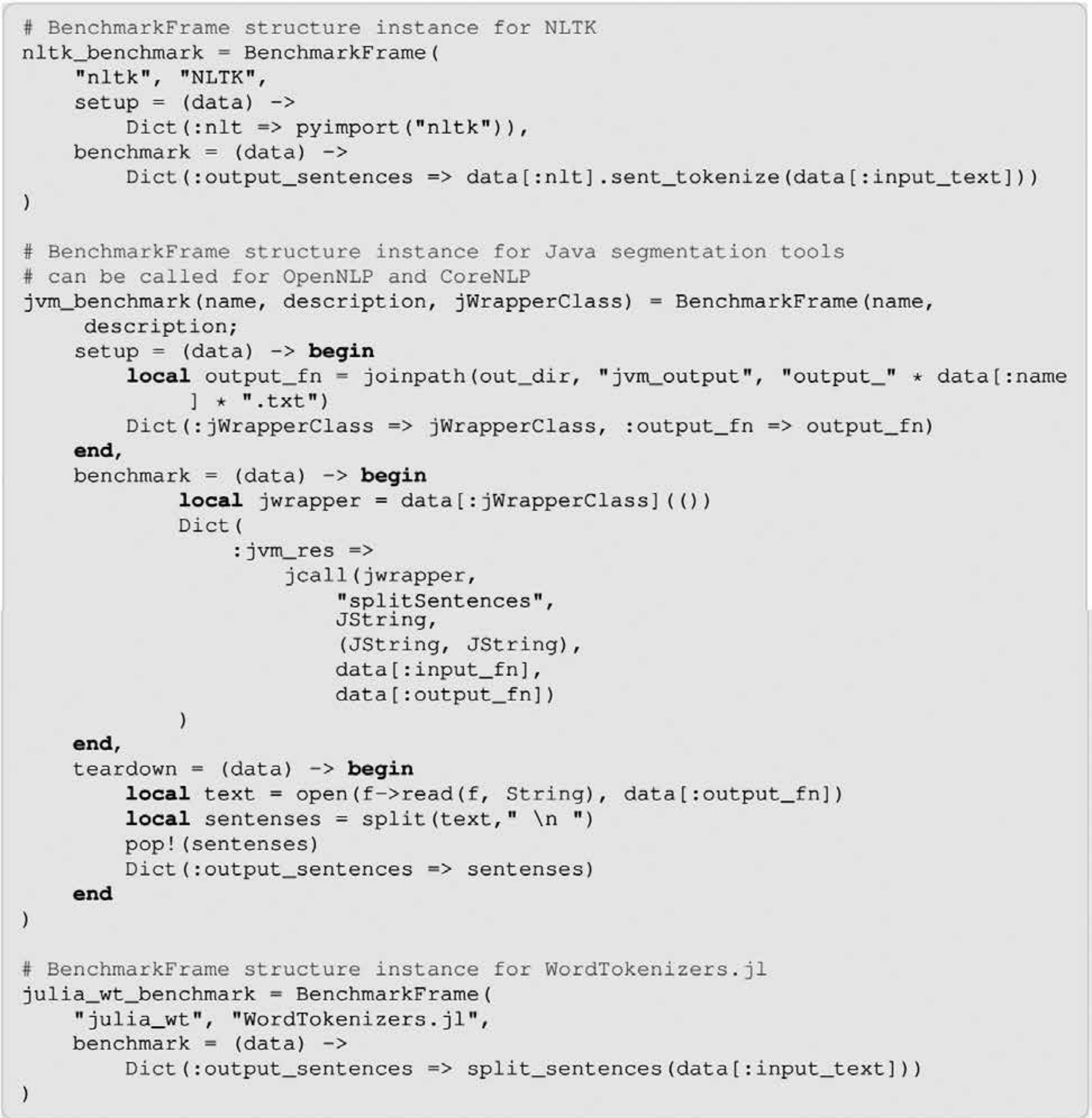
Fig. 6. Package diagram



```
evaluate_splitters() = evaluate_list(() -> [
        nltk_benchmark(samples = 100, seconds = 40),
        opennlp_benchmark(samples = 100, seconds = 40),
        corenlp_benchmark(samples = 100, seconds = 40),
        julia_wt_benchmark(samples = 100, seconds = 40),
    ]
)

evaluate_all() = evaluate_splitters()
```

Fig. 7. Example of calling all benchmark objects

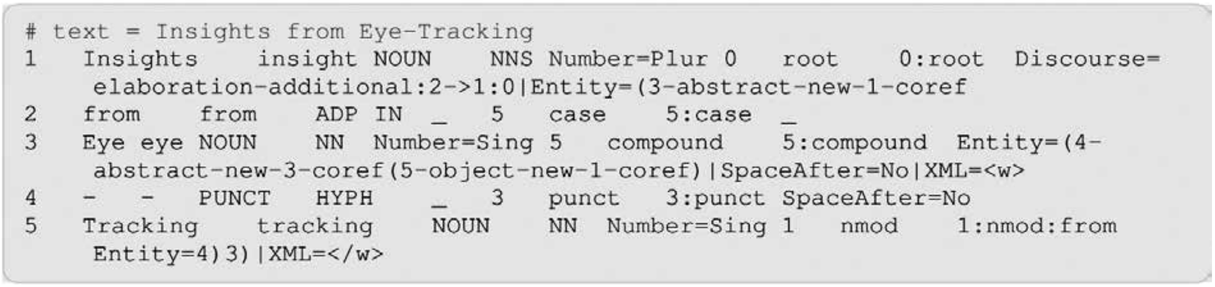


```
# BenchmarkFrame structure instance for NLTK
nltk_benchmark = BenchmarkFrame(
    "nltk", "NLTK",
    setup = (data) ->
        Dict(:nlt => pyimport("nltk")),
    benchmark = (data) ->
        Dict(:output_sentences => data[:nlt].sent_tokenize(data[:input_text]))
)

# BenchmarkFrame structure instance for Java segmentation tools
# can be called for OpenNLP and CoreNLP
jvm_benchmark(name, description, jWrapperClass) = BenchmarkFrame(name,
     description;
    setup = (data) -> begin
        local output_fn = joinpath(out_dir, "jvm_output", "output_" * data[:name
             ] * ".txt")
        Dict(:jWrapperClass => jWrapperClass, :output_fn => output_fn)
    end,
    benchmark = (data) -> begin
            local jwrapper = data[:jWrapperClass](())
            Dict(
                :jvm_res =>
                    jcall(jwrapper,
                        "splitSentences",
                        JString,
                        (JString, JString),
                        data[:input_fn],
                        data[:output_fn])
            )
    end,
    teardown = (data) -> begin
        local text = open(f->read(f, String), data[:output_fn])
        local sentenses = split(text," \n ")
        pop!(sentenses)
        Dict(:output_sentences => sentenses)
    end
)

# BenchmarkFrame structure instance for WordTokenizers.jl
julia_wt_benchmark = BenchmarkFrame(
    "julia_wt", "WordTokenizers.jl",
    benchmark = (data) ->
        Dict(:output_sentences => split_sentences(data[:input_text]))
)
```

Fig. 8. BrenchmarkFrame structure for NLTK and WordTokenizers.jl



```
# text = Insights from Eye-Tracking
1   Insights    insight NOUN    NNS Number=Plur 0   root    0:root  Discourse=
     elaboration-additional:2->1:0|Entity=(3-abstract-new-1-coref
2   from    from    ADP IN  _   5   case    5:case  _
3   Eye eye NOUN    NN  Number=Sing 5   compound    5:compound  Entity=(4-
     abstract-new-3-coref(5-object-new-1-coref)|SpaceAfter=No|XML=<w>
4   -   -   PUNCT   HYPH    _   3   punct   3:punct SpaceAfter=No
5   Tracking    tracking    NOUN    NN  Number=Sing 1   nmod    1:nmod:from
     Entity=4)3)|XML=</w>
```

Fig.9. One sentence from the GUM Corpus

test sample. The forecasts made are considered either positive or negative, and the expected judgments are true or false.

Four classes that include all predictions made by the segmentation tool are shown in Table 1. Predictions must be made for each token (word, punctuation mark) in the sentence. So each token goes to one of the four classes (TP, FN, TN, FN) according to prediction.

Table 1

Confusion matrix

Class	Predict	Result	Explanation
TP, True Positive	1	1	Line break where it should be
FP, False Positive	1	0	Line break NOT where it should be
TN, True Negative	0	1	There is no line break, but there should be
FN, False Negative	0	0	There is no line break, and there shouldn't be

We use the following general indicators [15]:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Error = 1 - Accuracy = \frac{FP + FN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F_{measure} = 2 * \frac{Precision * Recall}{Precision + Recall}$$

2.4. Execution time performance

Table 2 shows the results of measuring the execution time of the text segmentation tools.

The table is sorted in ascending order of average execution time for each tool.

The first three lines are the simplest and the fastest algorithms with the supposedly lowest segmentation quality. Next, the best results are WordTokenizers.jl and OpenNLP. The difference in the average execution time of 100 iterations ranges from 0.1s to 2s at each benchmark run, with WordTokenizers.jl having an advantage. Thus, the WordTokenizers.jl (Julia-based with heuristics) can be said to perform sentence splitting faster than others. At this step, the hypothesis about rule-based methods is only partially confirmed. One of the rule-based tools shows the fastest results, the other two show the slowest. It all depends on the quality of heuristics. The next step is to estimate the quality of the sentence splitting.

Table 2

Performance

Tool Name	Iterations	Time(ms)
Julia split()	100	0,57
Julia split() with file	100	3,58
SimpleSplitter	100	13,77
WordTokenizers.jl	100	30,20
OpenNLP	100	30,73
NLTK	100	216,07
CoreNLP	100	311.99
Spacy (Rule-based)	43	1036.76
Sentencize.jl	7	6476.84
Spacy (Dependency parser)	6	10690.8

2.5. Text segmentation quality

As each segmentation tool has its own tokenizer, the number of tokens (predictions) for each tool might be different. Thus, we can compare the quality of segmentation with relative values (accuracy, error, precision, error, f1).

It is worth clarifying that "Julia split () with file" is omitted in this table, since "Julia split ()" and "Julia split () with file" are one splitting algorithm, therefore, the results of the accuracy estimation will be the same.

The quality of text segmentation is shown in Table 3. As we can see, the best quality is shown by Sentencize.jl - port of a rule-based Perl extension for

Table 3

Comparison metrics results.

Tool Name	tp	fp	tn	fn	accuracy	error	precision	recall	f1
Sentencize.jl	6330	254	107813	1078	0,99	0,01	0,96	0,85	0,905
NLTK	6269	283	107787	1139	0,99	0,01	0,96	0,85	0,898
OpenNLP	6255	276	107791	1153	0,99	0,01	0,96	0,84	0,897
CoreNLP	6278	362	107786	1130	0,99	0,01	0,95	0,85	0,894
WordTokenizers.jl	6140	264	107809	1268	0,99	0,01	0,96	0,83	0,889
Spacy (Dependency parser)	6631	934	107268	777	0,99	0,01	0,88	0,90	0,886
Spacy (Rule-based)	6183	994	107531	1225	0,98	0,02	0,86	0,83	0,848
SimpleSplitter	5760	772	107847	1648	0,98	0,02	0,88	0,78	0,826
Julia split()	5760	878	107780	1648	0,98	0,02	0,87	0,78	0,820

Table 4

Number of 2nd type errors

Tool Name	Errors
Sentencize.jl	0
NLTK	3
OpenNLP	0
CoreNLP	84
WordTokenizers.jl	6
Spacy (Dependency parser)	135
Spacy (Rule-based)	458

sentence splitting. At this step, again the hypothesis is only partially confirmed. The best quality is shown by another heuristic.

2.6. Types of segmentation errors

For a more complete understanding of the work of each segmentation tool, it is necessary to take into account the nuances of their work. This can be done by printing out the markup errors and comparing them to the reference markup. During the analysis, we identified two types of errors: errors in setting the line break and errors in recognizing tokens.

Errors of the 1st type are associated with the absence of punctuation marks in the source text, headings, and enumerations. All segmentation tools make similar errors in the same places. There is no point in showing them.

Errors of the 2nd type are more exotic. They depend on the segmentation tool and are not repeated in the analyzed tools. It is associated with a specific segmentation algorithm (method). These errors consist of incorrect recognition of tokens and occur with consecutive punctuation marks. Consider some examples in Table 5.

Table 5

Error examples

WordTokenizers.jl	
Reference markup	Markup error
- It ‘s a little bit like Achilles and the turtle.\vskip 3pt- ... love story and romance and surprises and tragedies and all this but alsothis structure interested me a lot.	- It ‘s a little bit like Achilles and the turtle.... love story and romance andsurprises and tragedies and all this but also this structure interested me a lot.
NLTK	
Reference markup	Markup error
- A Connecticut Yankee in King Arthur ‘s Court (solo)\vskip 3pt- Ed.: See the LibriVox catalog for a full index.	- A Connecticut Yankee in King Arthur ‘s Court (solo) Ed.\vskip 3pt- : See the LibriVox catalog for a full index.
CoreNLP	
Reference markup	Markup error
- Had they died fast or were they now suffering a fate far worse..?	- Had they died fast or were they now suffering a fate far worse.\vskip 3pt- .\vskip 3pt- ?
Spacy (Dependency parser)	
Reference markup	Markup error
- Finally, our study complements Navarro’s (2016) automatic metrical analyses of Spanish Golden Age sonnets, by covering a wider period and focusing on enjambment.	- Finally, our study complements Navarro\vskip 3pt- s (2016) automatic metrical analyses of Spanish Golden Age sonnets, by covering a wider period and focusing on enjambment.
Spacy (Rule-based)	
Reference markup	Markup error
- The severe concerns underpinning the alleged crisis have several dimensions relating to: (a) the (small) amount of published replication research; (b) the (poor) quality of replication research; and (c) the (lack of) reproducibility, which refers to the extent to which findings can (not) be reproduced in replication attempts that have been undertaken.	- The severe concerns underpinning the alleged crisis have several dimensions relating to: (\vskip 3pt- a) the (small) amount of published replication research; (\vskip 3pt- b) the (poor) quality of replication research;\vskip 3pt- and (c) the (lack of) reproducibility, which refers to the extent to which findings can (not) be reproduced in replication attempts that have been undertaken.

The results in Table 4 shows that CoreNLP and Spacy make the most errors of the 2nd type (in recognizing tokens). Because of this, the share of correct predictions and other indicators were not maximum. It also confirms that Sentencize.jl performs segmentation with the fewest errors (all tokens were recognized correctly).

Conclusion

In this paper, we compared the results of the 8 segmentation tools using comparison metrics and calculating performance. It is worth noting that the best results in terms of performance (WordTokenizers.jl) and quality (Sentencize.jl) belong to Julia tools. The benchmark source code is available at https://bmstu.codes/AnnaZav/sentencesplitterbenchmark.

The novelty from the technical side is developed by our unified benchmarking framework for libraries written in different programming languages which allows connecting the new libraries into a testing pipeline and getting comparison results on both quality and execution performance. And, due to the selected Julia language, these libraries might be written in different languages, including native Julia, Python, R, Java, etc. That work confirms that Julia might be used as an integration platform.

This paper is a part of the research work carried out within the Bauman Deep Analytics project of the Priority 2030 program.

References

1. Text to sentence splitter. https://github.com/mediacloud/sentence-splitter, 2019. Accessed: 2022-01-20.
2. Apache. Opennlp. http://opennlp.apache.org, 2010. Accessed: 2022-01-20.
3. *Bird, S., Klein, E., and Loper, E.* Natural language processing with Python: analyzing text with the natural language toolkit. “ O’Reilly Media, Inc.”, 2009.
4. *Bolshakova, E.I., Peskova, O., Klyshinsky, E., Noskov, A.A., Lande, D., and Yagunova, E.V.* Automatic natural language processing and computational linguistics, 2015.
5. *Chen, J., and Revels, J.* Robust benchmarking in noisy environments. arXiv e-prints (Aug 2016).
6. *Community, T.J.* Calling c and fortran code, may 2022.
7. *Community, T.J.* Why we use julia, 10 years later, february 2022.
8. *Honnibal, M., and Johnson, M.* An improved non-monotonic transition system for dependency parsing. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (Lisbon, Portugal, Sept. 2015), Association for Computational Linguistics, pp. 1373–1378.
9. *Honnibal, M., and Montani, I.* spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. 2017.
10. *Koehn, P., et al.* Europarl: A parallel corpus for statistical machine translation. In MT summit (2005), vol. 5, Citeseer, pp. 79–86.
11. *Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J. R., Bethard, S., and McClosky, D.* The stanford corenlp natural language processing toolkit. In Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations (2014), pp. 55–60.
12. *Nivre, J., and Nilsson, J.* Pseudo-projective dependency parsing. In Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL’05) (Ann Arbor, Michigan, June 2005), Association for Computational Linguistics, pp. 99–106.
13. *Ruopp, A.* Lingua sentence. https://metacpan.org/pod/Lingua::Sentence, 2010. Accessed: 2022-01-20.
14. *Sætre, R., Søvik, H., Amble, T., and Tsuruoka, Y.* Genetuc, genia and google: Natural language understanding in molecular biology literature. In Transactions on Computational Systems Biology V (Berlin, Heidelberg, 2006), C. Priami, X. Hu, Y. Pan, and T. Y. Lin, Eds., Springer Berlin Heidelberg, pp. 68–82.
15. *Soricut, R., and Marcu, D.* Sentence level discourse parsing using syntactic and lexical information. In Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (2003), pp. 228–235.
16. *Zeldes, A.* The GUM corpus: Creating multilayer resources in the classroom. Language Resources and Evaluation 51, 3 (2017), 581–612.

A.P.Zavyalova. Master student. Bauman Moscow State Technical University, ul. Baumanskaya 2-ya, 5, Moscow, 105005, Russia. E-mail: annazav13@gmail.com

P.A.Martynyuk. Master student. Bauman Moscow State Technical University, ul. Baumanskaya 2-ya, 5, Moscow, 105005, Russia. E-mail: martapauline@yandex.ru

R.S.Samarev. Associate Professor. Bauman Moscow State Technical University, ul. Baumanskaya 2-ya, 5, Moscow, 105005, Russia. E-mail: samarev@acm.org

The Conceptual Modeling System Based on Metagraph Approach

N.D. Todosiev, V.I. Yankovskiy, Y.E. Gapanyuk, A.M. Andreev

Bauman Moscow State Technical University, Moscow, Russia

Abstract. The article is devoted to an approach to building a conceptual modeling system, which includes text recognition in a conceptual structure and text generation based on a conceptual structure. The metagraph is used as a conceptual structure. The architecture of the conceptual modeling system is proposed. The metagraph model is considered as a data model for conceptual modeling. The main ideas of the work of the text parsing module and text generation module are considered.

Keywords: *complex Graph Structures, Metagraph, Conceptual Compression, Text Parsing, Text Generation.*

DOI: 10.14357/20790279230120

Introduction

Humanity has accumulated a huge number of text documents. The task of extracting the meaning of a text from a large number of documents is difficult for the user and may require significant time costs.

The task becomes more complicated when the user is a decision-maker and must distinguish the meaning of incoming texts and make decisions in a limited time. One of the common ways to "conceptually compress" text information is to use conceptual models. Such models include mindmap diagrams, concept maps, and, in part, ontological models. Research is currently underway to develop conceptual models presented in complex graph structures, such as hypergraphs, hypernetworks, and metagraphs. The use of complex graph structures provides a significant degree of "conceptual compression". The use of conceptual models in the decision-making task involves the sequential implementation of three enlarged steps:

1. Synthesis of a conceptual model based on a text description.
2. Conceptual modeling, as a result of which new conceptual models are formed.
3. Analysis of the results of modeling, decision-making, and the formation of reports based on the decisions made.

In this regard, the development of methods and algorithms for the synthesis of conceptual models based on a text description, implementing step 1, and generation of text reports based on the models obtained as a result of conceptual modeling in step 3 are urgent tasks without solving which it is impossible to implement conceptual modeling fully.

This article discusses both the architecture of the conceptual modeling system and the principles of implementing the main modules of the system.

1. The Architecture of a Conceptual Modeling System

The Architecture of a Conceptual Modeling System is represented in the Fig. 1.

The system under development contains three large modules:

1. The text parsing module.
2. The text generation module.
3. The metagraph concepts modeling module.

The operation of the system consists of nine main steps:

1. In "Step I", a source text document is read.
2. In "Step II", "the text parsing module" parses the document, extracts concepts and relationships, and creates a metagraph structure.
3. In "Step III", the generated metagraph structure is recorded into "the metagraph concepts storage".
4. In "Step IV", "the metagraph concepts modeling module" receives the sourceconcepts for modeling from "the metagraph concepts storage".
5. In "Step V", the conceptual modeling is performed. The source concepts inthe form of metagraph are translated to the destination concepts.
6. In "Step VI", the results of conceptual modeling are recorded into "themetagraph concepts storage".

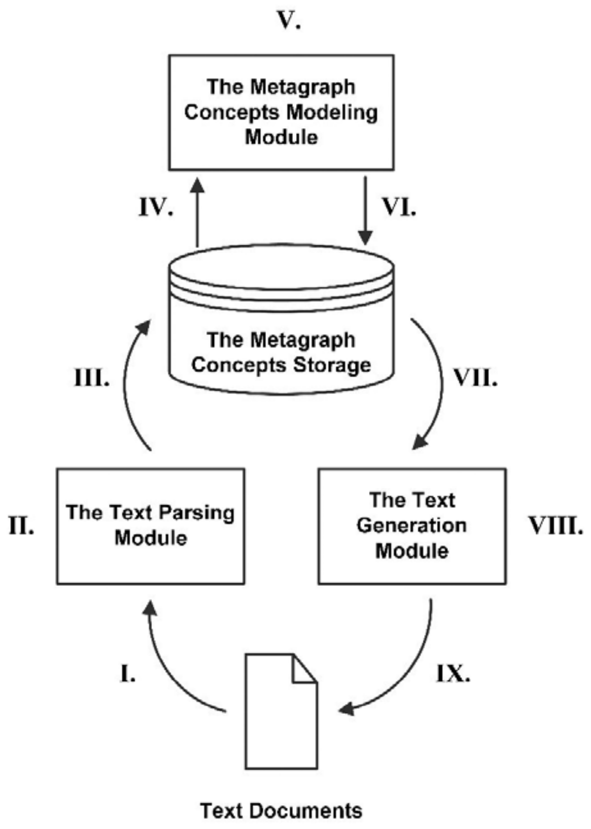

V.
The Metagraph Concepts Modeling Module
IV.
VI.
The Metagraph Concepts Storage
III.
VII.
II.
The Text Parsing Module
The Text Generation Module
VIII.
I.
IX.
Text Documents

Fig.1. The Architecture of a Conceptual Modeling System

7. In “Step VII”, “the text generation module” receives the destination concepts from “the metagraph concepts storage”.
8. In “Step VIII”, “the text generation module” transforms destination concepts into text form.
9. In “Step IX,” the output text document is generated.

The system architecture includes “the metagraph concepts storage”. The detailed structure of the information stored in “the metagraph concepts storage” is the subject of a separate research. The main ideas of storing a metagraph model in relational, document-oriented, and graph databases are discussed in [15].

Various options for metagraph modeling are the subject of a separate study. In this article, we will consider in detail the principles of extracting metagraph structures from text and generating text based on the metagraph structure, which corresponds to modules “The text parsing module” and “The text generation module”. We will also consider the basic principles that underlie “The metagraph concepts modeling module”.

2. Using Metagraphs for Concepts Modeling

In this section, we discuss the use of a metagraph model as a data model for “The metagraph concepts modeling module”. We will also compare flat concept maps (such as MindMaps and CMaps) with a metagraph model.

MindMaps. Probably, the most well-known approach to the representation of conceptual maps in practice is MindMap or “relationship diagram” [3].

The idea of constructing such a diagram is that the main topic (“topic”) is depicted in the center of the sheet, in which hierarchical subtopics are nested. Various views can be used to visualize such a diagram. Hierarchically nested concepts can be displayed in the form of a tree, can be located on concentric circles, can be represented in the form of an Ishikawa diagram [6] (which is also known as a “cause-effect diagram”). The simplest software products support only one version of the representation of the relationship diagram (as a rule, this is the “classic” version, in which the concepts are located on concentric circles). More advanced software products (for example, XMind [1]) support several representation options and automatically convert a structure from one representation to another.

From the point of view of the data model, a relationship diagram is a flat graph whose vertices correspond to concepts and whose edges correspond to connections between concepts. Edges in this model are considered as non-directional and non-annotated (an edge cannot be assigned a label containing auxiliary data). It is the property of non-annotability that allows automatic transformations between views.

Advanced software products (for example, XMind) also allow you to create additional annotated edges (actually turning the link diagram model into a CMap model), but such edges are not subject to automatic transformations between views.

Concept maps (CMaps). The CMap approach was proposed by Professor Joseph D. Novak [10]. The CmapTools system is an automated tool developed based on Novak’s theory, designed for the formalization of subject areas, and is also often used in practice as an automated learning tool.

From the point of view of the data model, a CMap (like a MindMap) is a flat graph whose vertices correspond to concepts, and whose edges correspond to connections between concepts. In contrast to the MindMap diagram, the edges in this model are considered as directed and annotated (an edge can be assigned a label containing auxiliary data). Unlike the MindMap diagram model, it is not possible to perform automatic transformations between views for the CMap model.

Advantages and Disadvantages of Flat Concept Maps. The main advantages of the existing approaches to the presentation of concept maps are:

1. The conceptual map is presented in a graphical

form, which allows the userto form a complete representation of the subject area.

2. This version of the presentation allows you to understand the hierarchicalrelationship between the concepts.
3. In the case of using a concept map as a training tool, it is possible to graduallyadd concepts and connections, which allows the student to form a holistic view of the subject area step by step.
4. The main disadvantages of the existing approaches to the presentation of concept maps are:
5. The MindMap diagram does not allow annotating edges, which seriouslylimits the expressive capabilities of the model. This problem is solved in the CMap approach.
6. Both the MindMap diagram and CMap use flat graphs to represent conceptual maps. This leads to the fact that a concept map with a size of even a few dozen concepts becomes almost unreadable. In particular, this happens due to a violation of Miller's law [13], since at the same time, the user is forced to work with a number of concepts that exceeds the number 7 ± 2.

To solve the second problem, the CMap approach uses the concept of nested nodes. But nested nodes allow you to combine ordinary nodes only once – hierarchical embedding of nested catches into each other is impossible in the CMap approach.

Thus, having analyzed the existing approaches to the representation of conceptual maps, we can conclude that the main problem of the existing approaches is using a flat graph as a model for the representation of a conceptual map. Next, we will consider the possibility of using complex graphs (metagraphs) as a model for conceptual maps.

The Metagraph Model for Concepts Modeling. The metagraph model is a kind of complex graph model. In this article, by metagraph we will understand the following: $MG = \langle V, MV, E \rangle$, where MG – metagraph; V – set of metagraph vertices; MV – set of metagraph metavertices; E – set of metagraph edges.

It should also be noted that in some versions of the metagraph model, there is such an element as a metaedge [14]. But in the proposed approach, metaedges are not used for conceptual modeling, so they are not considered in this article.

Metagraph vertex $v_i = \{atr_k\}, v_i \in V$, where atr_k – attribute. Metagraph edge $e_i = \langle v_S, v_E, \{atr_k\} \rangle, e_i \in E$, where v_S – source vertex (metavertex) of the edge; v_E – destination vertex (metavertex) of the edge; atr_k – attribute.

The metagraph fragment is defined as $MG_i = \{ev_j\}, ev_j \in (V \cup E \cup MV)$, where ev_j – an element that belongs to the union of vertices, edges and metavertices. The metagraph metavertex: $mv_i = \langle \{atr_k\}, MG_f \rangle, mv_i \in MV$, where mv_i – metagraph metavertex; atr_k – attribute, MG_f – metagraph fragment.

The main element of the metagraph model is the metavertex. From the general system theory point of view, metavertex is a special case of manifestation of emergence principle, which means that metavertex with its private attributes and connections became a whole that cannot be separated into its component parts. The example of metagraph representation is given in the Fig. 2.

The example contains three metavertices: mv_1, mv_2 and mv_3. Metavertex mv_1 contains vertices v_1, v_2, v_3 and connecting them edges e_1, e_2, e_3. Metavertex mv_2 contains vertices v_4, v_5 and connecting them edge e_6. Edges e_4, e_5 are examples of edges connecting vertices v_2–v_4 and v_3-v_5 are contained in different metavertices mv_1 and mv_2. Edge e_7 is an example of edge connecting metavertices mv_1 and mv_2. Edge e_8 is an example of edge connecting vertex v_2 and metavertex mv_2. Metavertex mv_3 contains metavertex mv_2, vertices v_2, v_3 and edge e_2 from metavertex mv_1 and also edges e_4,e_5,e_8 showing emergent nature of metagraph structure.

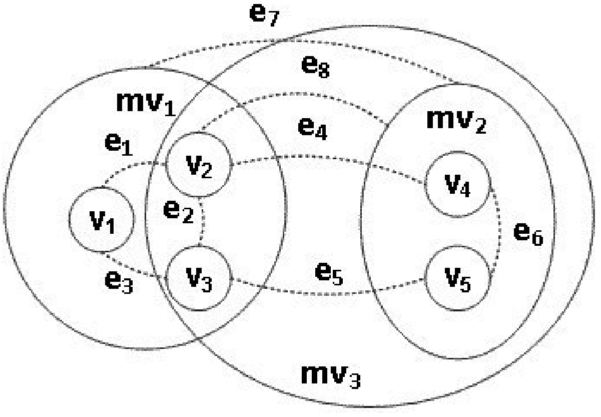


Fig.2. The example of metagraph representation

The metagraph model can be useful for describing conceptual maps. In this case, we can consider "simple" and "complex" concepts. At the same time, "simple" concepts are modeled using ordinary vertices, and "complex" ones are modeled using metavertices.

The use of metavertices to describe conceptual maps allows us to abandon the representation of a conceptual map in the form of a flat graph and switch to a holonic spatial description of a conceptual map in the form of a metagraph.

Using metagraph calculus [16], it is possible to carry out a formal transformation of the metagraph in the process of conceptual modeling.

The process of conceptual modeling is based on the sequential transformation of metagraphs using metagraph agents, which are described in [17]. As a result of modeling, an existing metagraph can either be modified or a new metagraph can be created.

In order to implement "the metagraph concepts storage", the graph model can be transformed into a flat graph model as shown in [15]. Also, in accordance with [15], the database based on the flat graph model or document model or relational model can be used to store a metagraph model.

3. The Text Parsing Module

The module is based on the frame approach. For researching the frame semantic core transformation, the following frame concepts are considered:

- Zhabotinskaya's frames[19]. The author of the article proposes frames of a simpler (lower) level, which are very close in meaning to syntactic constructions (cases, sentence members, etc.) Thus, it is not difficult to find a certain set of rules for converting syntax trees into Zhabotinskaya frames.
- FrameNet [2] database is a great tool providing many frame definitions, semantic roles, and word senses. Frames provide structure information. Frame instances provide data information. Entities in frame instances can be linked, showing that they are the same. Due to the nature of frames, entities inside some frame can also be frames. The resulting data representation is shown in the Fig. 3.
- FrameBank database is a thesaurus for Russian language [8]. Once there were just the Russian version of FrameNet [7], it were transformed and adapted for Russian, which has more morphological and semantic features than English.

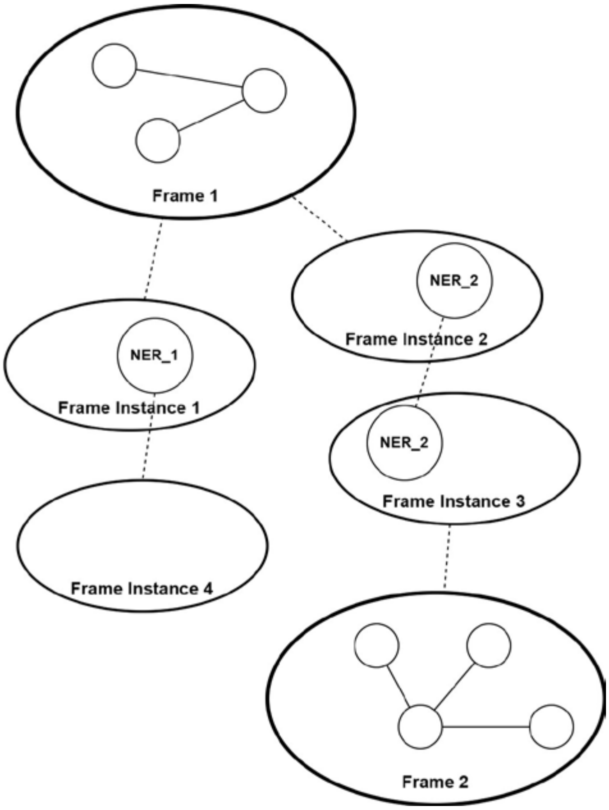


Fig.3. The Data Representation of the Frame Semantic Core

As a result, the FrameNet is considered more suitable because of its more flexible structure and acceptance worldwide. The Russian version of it will be used in further work. In this article, the English version is used.

The general idea of the graph generation algorithm is shown in the Fig. 4.

Initially, the source text is subjected to the resolution of reference links (coreference), which allows you to remove leaves that do not carry useful information, and also make the future graph more connected.

Then the text gets into the module for converting text to frames [12]. After that, the received frames are linked by target words. Thus, a graph is obtained, each node of which is a word. The frame attributes are turned into similar nodes. Links between the main node and the attribute node are marked with the appropriate tag (frame attribute).

Up to this point, we are working with a flat graph in which initially parsed low-level concepts correspond to words and phrases of the source text and relationships between these low-level concepts corresponds to the links both between the members of the sentence and between individual sentences.

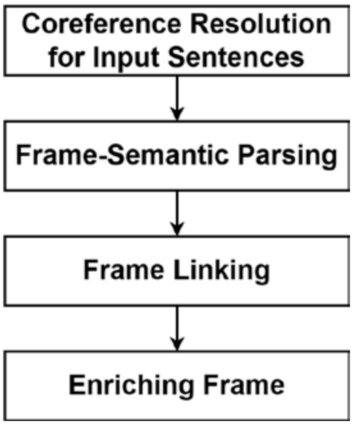


Fig.4. Algorithm for text-to-graph transformation

Further enrichment of frames leads to the formation of high-level concepts based on low-level concepts. But it is impossible within the framework of the flat graph structure, so we turn to the metagraph model.

Enriched frames can be considered as metavertices of a metagraph. This is due to the nested structure of the resulting frames – frames can be atomic or contain other frames. In this model, metavertices may be considered as compositions of low-level vertices.

It should be noted that items of "the metagraph concepts storage" only indirectly corresponds to sentences of the source text. The "the metagraph concepts storage" stores the high-level concepts of the metagraph model that were extracted and enriched based on the source text.

4. The Text Generation Module

The general idea of the text generation algorithm is shown in the Fig. 5.

First of all, the generating text system needs to know the purpose of generating text. This purpose will determine the path of the algorithm. The purpose of generating text is considered of these types of input information:

– User request.
– Meta information i.e., user language, user location, history of user request.

It should be noted that the purpose of generating text can be supplemented by another kind of information, such as another text input from the user explaining how the result text should be like.

This purpose is then analyzed for two main goals: find out which parts of the metagraph will be used to output the answer, and determine in which format output text will be generated. For simplistic purposes, it is supposed that analyzing input information is transforming this input into a semantic representation. That way, the purpose of generating text can be presented as semantic and manipulated as such.

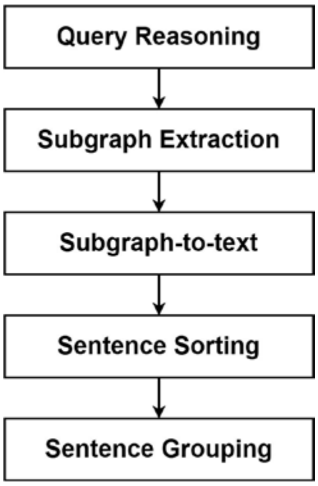


Fig.5. Algorithm for graph-to-text transformation

After analyzing the purpose of the generating text, the algorithm uses this semantic of purpose in semantic representation and highlights the individual parts of the metagraph.

These results are then compared with a graph to find parts of a metagraph. For the general case, it can be done in several approaches:

1. If supposed, that input frame and metagraph are using the same semanticcore, then this task becomes a subgraph isomorphism problem with NPcompleteness, which has too high computational costs. It has been solved recently [9], so this approach can be used for it;
2. In case when input frame and metagraph are not using same semantic core,then the graph comparison methods can be performed: the target frame is matched to search the relevant part of the metagraph;
3. Use unique approaches related to the features of the semantic core.

Highlighted parts of the metagraph will transform into text representation with syntactic and morphological transformations.

In the case of using frame semantic core, this algorithm became rather deterministic. Text of user query transform into semantic frames by the same algorithm, as described earlier, to transform the text into a semantic representation with one caveat: the interrogative words will be marked as "blank words". This helps to split the purpose of generating text into two categories:

1. Search purpose, i.e. the user with search query wants to find an answer inmetagraph. In this case, the model will be like the QnA system.
2. Descriptive purpose, i.e. the user doesn't have a certain question in the inputquery, rather the user wants to learn about something. The algorithm will subsequently focus on this category of purposes.

With the use of frame semantic core, using unique approaches related to the features of the semantic core is preferable because of the strict and descriptive structure of frames in FrameNet.

In the case of search, the algorithm will look for missing values of the vertex in the purpose frame. In the case of requests with an ambiguous answer (i.e. purpose were descriptive), a multi-frame response is obtained, where it is enough to request searches for certain vertices, and then search for all frames associated with this vertex, getting more than one frame. Then all the relevant parts of the metagraph will be picked out for the next step.

In both cases, the result is a subgraph of the metagraph representing the response, which is then converted to text. For text transformation, you can use both algorithmic approaches and deep neural networks (for example, T5 [11]).

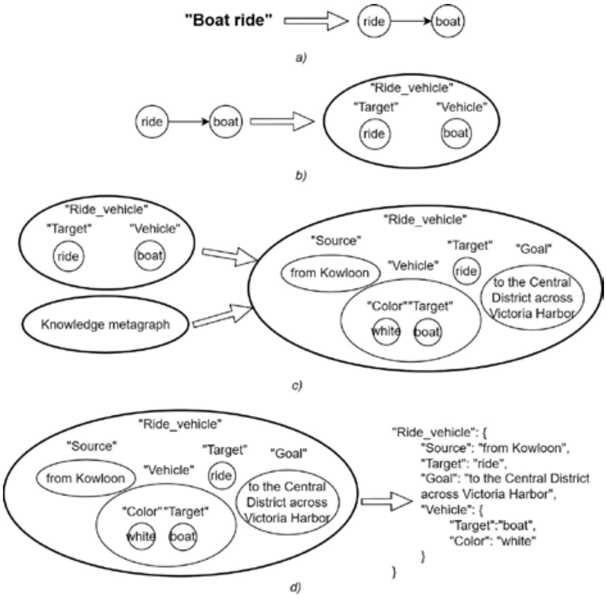
For syntactic transformation, the only thing that is needed is the following:

1. The resulting parts of the metagraph are sorted for generating text. This isdone by ATTOrderNet [4] or similar models [18].
2. The frame chain is split into parts that will represent future sentences. Itcan be done because semantic frames are well-structured, and each one of the frames makes one meaning. Sentences are generated based on the frame description of FrameNet. With the usage of Flesch reading ease scale for natural languages [5] it can be completed without complex logic: $FRE = 206{,}835 - 1{,}015 \times ASL - 84{,}6 \times ASW$; where ASL is an average number of words per sentence, and ASW is the average number of syllables per word. For each language, this formula slightly shifts in coefficients, but the variables stay the same. It is worth to be noted, that this method of analyzing the complexity of text tends to be a poor cause of used mean variables. But in the case of splitting frames into sentences, this approach is enough.
3. Supplementing a chain of frames with punctuation marks. In most cases,this will be a comma between frames, unless a period is included.

The presented algorithm can generate texts based on incoming text queries and constructed metagraph. In subsequent works, it is planned to refine the algorithm in the direction of generating texts in other languages, as well as generating texts with different styles.

5. Experiment

This system will be tested for the tasks of generating text based on data. A prototype was developed to test the performance of the system.



```
"Ride_vehicle": {
        "Source": "from Kowloon",
        "Target": "ride",
        "Goal": "to the Central District
        across Victoria Harbor",
        "Vehicle": {
              "Target":"boat",
              "Color": "white"
        }
}
```

Fig.6. Example of implementation of parsing and generating text

The system receives the user's input query in the form of a text string. According to the rules described in paragraph "The Text Parsing Module" a text string is converted into a graph. After that, the query subgraph is searched in the graph. In case of successful finding of the subgraph, the selection of vertices and links is called the response subgraph.

Next, the selected response subgraph is sent to the text generation module. T5 input must be represented as text, so we convert the graph to a text string according to the following rules:

1. Selected vertices of the response graph are translated into the response dictionary.
2. The root of the dictionary contains the main vertices.
3. The subtrees of the selected vertices turns into nested dictionaries.

Suppose we need to find certain information about a boat ride in the metagraph. In the Fig. 6 a), the input query transforms into syntax tree according to parse module. In the Fig. 6 b), the syntax tree parsed into metagraph that is called query metagraph. In the Fig. 6 c), shows the extraction of the response metagraph from the main metagraph by query metagraph. In the Fig. 6 d), the response metagraph is converted to a dictionary with nested vertex descriptions that is called response dictionary.

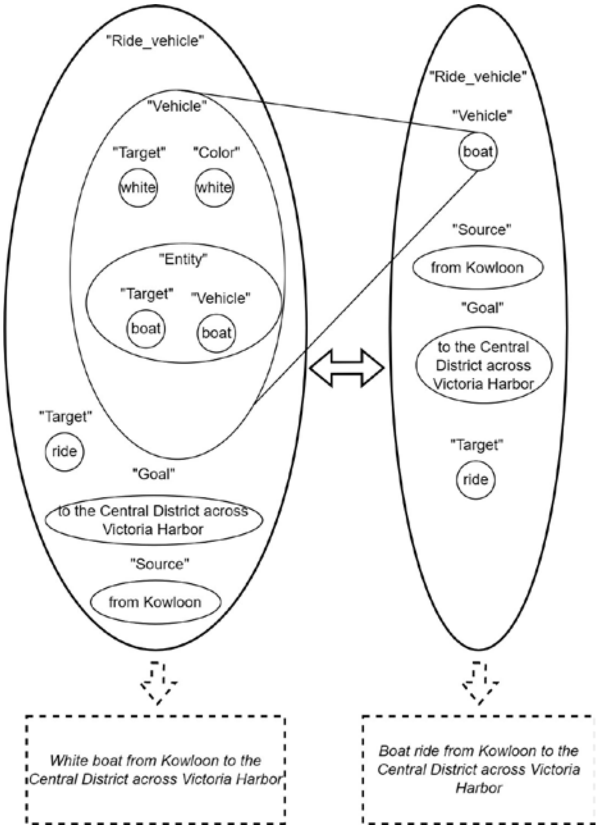

"Ride_vehicle"
"Vehicle"
"Target"
white
"Color"
white
"Entity"
"Target"
boat
"Vehicle"
boat
"Target"
ride
"Goal"
to the Central District across Victoria Harbor
"Source"
from Kowloon
"Ride_vehicle"
"Vehicle"
boat
"Source"
from Kowloon
"Goal"
to the Central District across Victoria Harbor
"Target"
ride
White boat from Kowloon to the Central District across Victoria Harbor
Boat ride from Kowloon to the Central District across Victoria Harbor

Fig.7. Example of Frame Refining and Sentence Representation

The T5 model [11] is used for transforming response dictionary (in text representation) into output text. It is worth noting that T5 can be fine-tuned to generate the desired style of speech, the desired detail of the answer. In this experiment, a FrameNet dataset [2] is used, consisting mainly of news messages. Detailing of answers varies by the number of selected frames.

There were about 5000 frames and their combinations in the original dataset. The dataset was divided into training and test samples in the ratio of 90% and 10%.

The response subgraph selection algorithm requires additional research. However, already now it is possible to vary the number of vertices and links of the response subgraph used to generate the response, get more or less detailed answers, control the subject of the response, the subject area of the response and many other factors of answer generation.

For example, in the Fig. 7 one can see how the graph is refined. So, by expanding the meaning of the word "boat" to "white boat", we change the output text.

Table 1

Result of the Experiments.

Metric Name	Metric Value
BLEU	0.501
METEOR	0.694
Cosine Similarity	0.813

Results of the test dataset validation are shown in Table 1. According to the definition of BLEU and METEOR metrics, result text theme is the same theme as referenced text theme. To confirm that the referenced and result texts are similar, cosine similarity is calculated.

The conducted experiment proves the idea of this system. The proposed system allows generating text based on data. It is worth noting, the FrameNet dataset has a relatively small number of sentences and frames. In the future, it is planned to increase the dataset with other sources, including texts converted into a graph.

6. Related Works and Discussion

The approaches proposed in the article relate to several areas of NLP, such as knowledge graph representation, natural language generation (NLG) and language modelling.

The article [22] proposes a text-enhanced knowledge graph representation model, named BCRL, which utilizes entity description and relation mention to enhance the knowledge representations of a triple. BCRL based on TransE [24] which is an energy-based model that produces knowledge base embeddings. It models relationships by interpreting them as translations operating on the lowdimensional embeddings of the entities.

Article [20] implements a system for generating the end of a story based on graphs. The implementation of the answer selection system is implemented as follows: several nodes of the graph are selected, each of which is weighted in some way, then these nodes fall into the answer. GPT-2 [25] and others were used as technologies.

The article [21] describes a language model based on the transformer architecture BERT [26], but with more details about the location of the token within the text, such as paragraph index, sentence index, and word index. The experimental results demonstrate that this proposed method works on both language models with relative position embeddings and pretrained language models with absolute position embeddings. The F1-score on datasets SQUAD1.1 and SQUAD2.0 [23] is 92.6 and 85.2 respectively.

Based on the results of the study of related works, we see the prospect of modern transformers, so in future work we will consider the option of combining metagraph concepts and transformers (e.g., BERT or GPT-2). Also, in future work it is planned to divide the system into three modules for a deeper comparison with other models.

Conclusions

The article proposes the conceptual modeling system based on a metagraph model that includes three main steps: synthesis of a conceptual model based on a text description; conceptual modeling, as a result of which new conceptual models are formed; analysis of the results of modeling, decision-making, and the formation of reports based on the decisions made.

The system architecture includes "the text parsing module", "the text generation module", and also "the metagraph concepts modeling module".

Having analyzed the existing approaches to the representation of conceptual maps, we can conclude that the main problem of the existing approaches is the use of a flat graph as a model for the representation of a conceptual map.

The "text parsing module" contains the transformation of the input text using various already existing techniques into a graph using frames. The resulting graph is enriched with additional data sources for a better description of the subject area.

The "Text generation module" is a straight forward algorithm in case of frame semantic core. The T5 model was used as a model for the experiment. Various query subgraphs are fed to the input, the output is a text.

The metagraph model can be useful for describing conceptual maps. In this case, we can consider "simple" and "complex" concepts. At the same time, “simple” concepts are modeled using ordinary vertices, and “complex” ones are modeled using metavertices. The use of metavertices to describe conceptual maps allows us to abandon the representation of a conceptual map in the form of a flat graph and switch to a holonic spatial description of a conceptual map in the form of a metagraph.

References

1. The XMind homepage. Available at: https://www.xmind.net/ (accessed August 30, 2022)
2. *Baker, C.F., C.J. Fillmore and J.B. Lowe.* 1998. The Berkeley FrameNet Project, In Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 1, Association for Computational Linguistics, pp: 86–90.
3. *Buzan, T.* 2018. Mind Map Mastery: The Complete Guide to Learning and Using the Most Powerful Thinking Tool in the Universe. Watkins Media.
4. *Cui, B., Y. Li, M. Chen and Z. Zhang*. 2018. Deep attentive sentence ordering network, In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, , pp: 4340–4349.
5. *Flesch, R.* 1948. A new readability yardstick. The Journal of Applied Psychology, 32(3): 221–233.
6. *Ishikawa, K.* 1986. Guide to Quality Control. Asian Productivity Organization.
7. *Lyashevskaya, O.N. and J.L. Kuznetsova*. 2009. Russian FrameNet: constructing a corpus--based dictionary of constructions [Russkij Frejmnet: k zadache sozdanija korpusnogo slovarja konstruktsij], In Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialog”[Komp’juternaja Lingvistika I Intelleltual’nye Tehnologii: Po Materialam Ezhegodnoj Mezhdunarodnoj Konferentsii “Dialog”], , pp: 306–312.
8. *Lyashevskaya, O. and E. Kashkin*. 2015. FrameBank: A Database of Russian Lexical Constructions, In Analysis of Images, Social Networks and Texts, Springer International Publishing, pp: 350–360.
9. *McCreesh, C., P. Prosser and J. Trimble*. 2020. The Glasgow Subgraph Solver: Using Constraint Programming to Tackle Hard Subgraph Isomorphism Problem Variants. Graph Transformation, 316–324.
10. *Novak, J. and A.J. Cañas*. 2006. The Origins of the Concept Mapping Tool and the Continuing Evolution of the Tool. Information Visualization, 5(3): 175–184. https://doi.org/10.1057/palgrave.ivs.9500126
11. *Raffel, C., N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li and P.J. Liu.* 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. arXiv [cs.LG]. https://doi.org/10.48550/arxiv.1910.10683
12. *Swayamdipta, S., S. Thomson, C. Dyer and N.A. Smith*. 2017. Frame-Semantic Parsing with Softmax-Margin Segmental RNNs and a Syntactic Scaffold. arXiv [cs.CL]. https://doi.org/10.48550/arxiv.1706.09528
13. *Talvitie, V.* 2018. The Foundations of Psychoanalytic Theories.
14. *Gapanyuk, Y.* 2021. The development of the metagraph data and knowledge model. In Selected Contributions to the 10th International Conference on “Integrated Models and Soft Computing in Artificial Intelligence (IMSC-2021)”. pp: 1–7
15. *Chernenkiy, V.M., Y.E. Gapanyuk, Y. Kaganov and I. Dunin*. 2018. Storing Metagraph Model in Relational, Document-Oriented, and Graph Databases. DAMDID/RCDL.
16. *Tarassov, V., Y. Kaganov and Y. Gapanyuk*. 2021. The Metagraph Model for Complex Networks: Definition, Calculus, and Granulation Issues, In Artificial Intelligence, Springer International Publishing, pp: 135–151. https://doi.org/10.23919/FRUCT48808.2020.9087470
17. *Chernenkiy, V., Y. Gapanyuk, A. Nardid and N. Todosiev*. 2020. The Implementation of Metagraph Agents Based on Functional Reactive Programming, In 2020 26th Conference of Open Innovations Association (FRUCT), pp: 1–8. https://doi.org/10.23919/FRUCT48808.2020.9087470
18. *Yin, Y., L. Song, J. Su, J. Zeng, C. Zhou and J. Luo*. 2019. Graph-based Neural Sentence Ordering. arXiv [cs.CL].
19. *Zhabotynska, S.A.* 2010. Principles of building conceptual models for thesaurus dictionaries. Cognition, Communication, Discourse, 1: 75–92.
20. *Ji, H., P. Ke, S. Huang, F. Wei, X. Zhu and M. Huang*. 2020. Language Generation with Multi-Hop Reasoning on Commonsense Knowledge Graph. arXiv [cs.CL]. https://doi.org/10.18653/v1/2020.emnlp-main.54
21. *Bai, H., P. Shi, J. Lin, Y. Xie, L. Tan, K. Xiong, W. Gao and M. Li*. 2021. Segatron: Segment-Aware Transformer for Language Modeling and Understanding. Proceedings of the AAAI Conference on Artificial Intelligence, 35(14): 12526–12534.
22. *Wu, G., W. Wu, L. Li, G. Zhao, D. Han and B. Qiao*. 2020. BCRL: Long Text Friendly Knowledge Graph Representation Learning, In The Semantic Web – ISWC 2020, Springer International Publishing, pp: 636–653.
23. *Rajpurkar, P., J. Zhang, K. Lopyrev and P. Liang*. 2016. SQuAD: 100,000+ Questions for Machine

Comprehension of Text. arXiv [cs.CL]. https://doi.org/10.18653/v1/D16-1264

24. *Bordes, A., N. Usunier and A. Garcia-Duran*. 2013. Translating embeddings for modeling multi-relational data. Advances in Neural Information Processing Systems.
25. *Radford, A., J. Wu, R. Child, D. Luan, D. Amodei and I. Sutskever*. 2019. Language models are unsupervised multitask learners. OpenAI Blog, 1(8): 9.
26. *Devlin, J., M.-W. Chang, K. Lee and K. Toutanova*. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv [cs.CL].

Todosiev N.D. Graduate student, Federal state budgetary institution of higher professional education «Bauman Moscow State Technical University», Moscow, Russia. Number of publications: 8. Research interests: information technology, natural language processing. E-mail: todosievnik@gmail.com.

Yankovsky V.I. Graduate student, Federal state budgetary institution of higher professional education «Bauman Moscow State Technical University», Moscow, Russia. Number of publications: 4. Research interests: information technology, natural language processing. E-mail: lucker1005000@gmail.com.

Gapanyuk Yu.E. Associate professor, Federal state budgetary institution of higher professional education «Bauman Moscow State Technical University», Moscow, Russia. Number of publications: about 100. Research interests: designing of automated systems, designing of hybrid intelligent information systems, complex graph models. E-mail: gapyu@bmstu.ru

Andreev A.M. Associate professor, Federal state budgetary institution of higher professional education «Bauman Moscow State Technical University», Moscow, Russia. Number of publications: about 100 (2 monographs). Research interests: designing of automated systems, designing of hybrid intelligent information systems, complex graph models. E-mail: arkandreev@gmail.com

Trudy Instituta sistemnogo analiza Rossiyskoy akademii nauk (ISA RAN)
(Proceedings of the Institute for Systems Analysis Russian Academy of Sciences (ISA RAS))

http://www.isa.ru/proceedings/

The journal was founded in 2002
The journal is published quarterly

Founder
Federal Research Center "Computer Science and Control" of Russian Academy of Sciences (FRC CSC RAS)
44/2, Vavilova str., Moscow, 119333, Russia, tel: +7(499)135-62-60,
e-mail: frccsc@frccsc.ru, http://www.frccsc.ru

Editor-in-Chief
Popkov Y.S. Academician of RAS, Professor,
9, Prospekt 60-letiya Oktyabrya, Moscow, 117312, Russia,
tel: +7(499)135-42-22, e-mail: popkov@isa.ru

Editorial address
9, Prospect 60-letiya Oktyabrya, Moscow, 117312, Russia
E-mail: tisa@isa.ru, tel. +7 (499) 135-24-33, +7 (926) 490-87-87

Scientific journal “Trudy Instituta sistemnogo analiza Rossiyskoy akademii nauk (ISA RAN)” publishes materials on a wide range of fundamental problems of developing systems analysis methodology and its applying to solving various problems in the field of science and practice. The journal is destined for scientists, engineers and researches working within the framework of these problems, as well as for politicians, employees of state and municipal departments, specialists of enterprises and representatives of social organizations. The rules for articles submission to the journal, as well as for their review are given at the journal’s site

The journal is registered in Federal service of observing and controlling legislation
in the area of mass media and protection of cultural heritage in 20 of May 2016. Digital certificate - ПИ № ФС77-65718.
Journal Subscription Index in the Ural-Press e-catalog (http://ural-press.ru) is 36089.

The journal is supported by the Department of Information Technologies and Computing Systems of the Russian Academy of Sciences.
The journal is included in the list of leading peer-reviewed scientific journals recommended by the Higher Attestation Commission.

Vol. 73
Issue 1

ISSN 2079-0279
2023

TRUDY
INSTITUTA SISTEMNOGO ANALIZA ROSSIYSKOY AKADEMII NAUK (ISA RAN)
(PROCEEDINGS OF THE INSTITUTE FOR SYSTEMS ANALYSIS RUSSIAN ACADEMY OF SCIENCES (ISA RAS))

ISBN 978-5-6048784-1-5
9 785604 878415

Отпечатано в ООО «Поли Принт Сервис»
127015, г. Москва, ул. Бутырская, д. 86
127247, г. Москва, Дмитровское шоссе, д. 100, стр. 2
Формат 60x84/8. Зак. № В-140
Тираж 100 экз.
Тел.: +7 (495) 797-35-59