Kazan (Volga Region) Federal University
Federal Research Center “Computer Science
and Control” of the Russian Academy of Sciences
ACM SIGMOD Chapter

Data Analytics and Management in Data Intensive Domains

Proceedings of the XXI International Conference DAMDID / RCDL'2019

October 15–18, 2019
Kazan, Russia

Edited by Alexander Elizarov, Boris Novikov,
Sergey Stupnikov

Kazan 2019

УДК [004.6:004.8]
ББК [32.973]

Data Analytics and Management in Data Intensive Domains: XXI International Conference DAMDID/RCDL' 2019 (October 15–18, 2019, Kazan, Russia): Conference Proceedings. Edited by Alexander Elizarov, Boris Novikov, Sergey Stupnikov. – Kazan: Kazan Federal University, 2019. – 496 p.

ISBN 978-5-00130-214-8

The “Data Analytics and Management in Data Intensive Domains” conference (DAMDID) is held as a multidisciplinary forum of researchers and practitioners from various domains of science and research, promoting cooperation and exchange of ideas in the area of data analysis and management in domains driven by data-intensive research. Approaches to data analysis and management being developed in specific data-intensive domains (DID) of X-informatics (such as X=astro, bio, chemo, geo, med, neuro, physics, chemistry, material science etc.), social sciences, as well as in various branches of informatics, industry, new technologies, finance and business contribute to the conference content. DAMDID conference was formed in 2015 as a result of transformation of the RCDL conference (“Digital libraries: advanced methods and technologies, digital collections”, http://rcdl.ru) so that the continuity with RCDL has been preserved after many years of its successful work.

The proceedings were funded by the subsidy allocated to Kazan Federal University for the state assignment in the sphere of scientific activities, project No. 1.13556.2019/13.1.

ISBN 978-5-00130-214-8

Preface

In 2019 XXI International Conference "Data Analytics and Management in Data Intensive Domains" (DAMDID/RCDL'2019) takes place on October 15–18 at the Kazan Federal University.

The "Data Analytics and Management in Data Intensive Domains" conference (DAMDID) is held as a multidisciplinary forum of researchers and practitioners from various domains of science and research, promoting cooperation and exchange of ideas in the area of data analysis and management in domains driven by data-intensive research. Approaches to data analysis and management being developed in specific data-intensive domains (DID) of X-informatics (such as X=astro, bio, chemo, geo, med, neuro, physics, chemistry, material science etc.), social sciences, as well as in various branches of informatics, industry, new technologies, finance and business contribute to the conference content.

Traditionally the program of the DAMDID/RCDL'2019 is oriented data science and data-intensive analytics as well as on data management topics. The program of this year includes keynotes and invited talks covering a broad range of conference topics.

Keynote by Bernhard Thalheim (full professor at Christian-Albrechts-University at Kiel) is devoted to models as one of the universal instruments of humans. Four specific utilisation scenarios for models are considered as well as methodologies for the development of proper and well-applicable models. Anton Osokin (leading research fellow and deputy head of the Centre of Deep Learning and Bayesian Methods at National Research University Higher School of Economics) gives a talk on ways for combining neural networks and powerful algorithms already developed in various domains. Experiences in creating B. Sc. and M. Sc. study programs in Data Science are considered in the keynote by Ivan Luković (full professor at the Faculty of Technical Sciences, University of Novi Sad). Invited talk by Mikhail Zymbler (head of Department for Data Mining and Virtualization at South Ural State University) considers an approach for the integration of data mining methods with open-source relational DBMS enhanced by small-scale modifications of the original codes to encapsulate parallelism.

Workshop on Data and Computation for Materials Science and Innovation (DACOMSIN) hosted by the National University of Science and Technology MISiS in Moscow constitutes the first day of the conference on October 15. The workshop is aimed to address communication gap across communities in the domains of materials data infrastructures, materials data analysis, and materials in silico experiment. The workshop brings together professionals from across research and innovation to share their experience and perspectives of using information technology and computer science for materials data management, analysis and simulation.

The conference Program Committee has reviewed 34 submissions for the conference, 18 submissions for DACOMSIN Workshop and 5 submissions for the Ph.D. workshop. For the conference, 17 submissions were accepted as full papers, 10 as short papers, whereas 7 submissions were rejected. For the DACOMSIN Workshop, 3 submissions were accepted as full papers, 3 submissions were accepted as short papers, 5 submissions were accepted as demos, 7 submissions were accepted as posters. For the Ph.D. workshop, 4 papers were accepted, and 1 was rejected.

According to the conference and workshops program, these 42 oral presentations are structured into 11 sessions including Advanced Data Analysis Methods, Digital Libraries and Data Infrastructures, Data Integration, Ontologies and Applications, Data and Knowledge Management, Data Analysis in Astronomy, Information Extraction from Text, Data Formats, Metadata and Ontologies in Support of Materials Research, Materials Databases, Materials Data Infrastructures and Data Services, IT Applications and IT Platforms for Materials Design and Simulation.

Though most of the presentations are dedicated to the results of researches conducted in the research organizations located on the territory of the Russian Federation including Chernogolovka, Dubna, Ekaterinburg, Kazan, Moscow, Novosibirsk, Obninsk, Tomsk, Samara, St.-Petersburg, the conference confirms features of internationalization witnessed by 7 talks prepared by the foreign researchers from such countries as Armenia (Yerevan), China (Shenzhen, Beijing), France (Clermont-Ferrand), Great Britain (Harwell), Japan (Tokyo), Switzerland (Lausanne).

The chairs of Program Committee expresses their gratitude to the PC members for carrying out the reviewing of the submissions and selection of the papers for presentation, to the authors of the submissions as well as to the host organizers from Kazan Federal University. The Program Committee of the conference appreciates the possibility of using the Conference Management Toolkit (CMT) sponsored by Microsoft Research, which provided great support during various phases of the paper submission and reviewing process.

October 2019

Alexander Elizarov
Boris Novikov
Sergey Stupnikov

Organization

General Chair

Alexander Elizarov	Kazan Federal University, Russia

Program Committee Co-chairs

Boris Novikov	National Research University Higher School of Economics, Saint Petersburg, Russia
Sergey Stupnikov	Federal Research Center “Computer Science and Control” of RAS, Russia

DACOMSIN Workshop Co-chairs

Nadezhda Kiselyova	IMET RAS, Russia
Vasily Bunakov	Science and Technology Facilities Council, Harwell, UK

PhD Workshop Chair

Mikhail Zymbler	South Ural State University, Russia

Organizing Committee Chair

Ayrat Khasyanov	Kazan Federal University, Russia

Organizing Committee Deputy Chair

Denis Zuev	Kazan Federal University, Russia

Organizing Committee

Iurii Dedenev	Kazan Federal University, Russia
Elena Tutubanilna	Kazan Federal University, Russia
Evgeny Lipachev	Kazan Federal University, Russia
Nikolay Skvortsov	Federal Research Center “Computer Science and Control” of RAS, Russia

Victor Zakharov	Federal Research Center "Computer Science and Control" of RAS, Russia
Natalya Zaitseva	Kazan Federal University, Russia

Supporters

Kazan Federal University
Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences (FRC CSC RAS), Moscow, Russia
Moscow ACM SIGMOD Chapter

Coordinating Committee

Igor Sokolov, Federal Research Center "Computer Science and Control" of RAS, Russia (Co-Chair)
Nikolay Kolchanov, Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia (Co-Chair)
Sergey Stupnikov, Federal Research Center "Computer Science and Control" of RAS, Russia (Deputy Chair)

Arkady Avramenko	Pushchino Radio Astronomy Observatory, RAS, Russia
Pavel Braslavsky	Ural Federal University, SKB Kontur, Russia
Vasily Bunakov	Science and Technology Facilities Council, Harwell, Oxfordshire, UK
Alexander Elizarov	Kazan (Volga Region) Federal University, Russia
Alexander Fazliev	Institute of Atmospheric Optics, RAS, Siberian Branch, Russia
Alexei Klimentov	Brookhaven National Laboratory, USA
Mikhail Kogalovsky	Market Economy Institute, RAS, Russia
Vladimir Korenkov	JINR, Dubna, Russia
Mikhail Kuzminski	Institute of Organic Chemistry, RAS, Russia
Sergey Kuznetsov	Institute for System Programming, RAS, Russia
Vladimir Litvine	Evogh Inc., California, USA
Archil Maysuradze	Moscow State University, Russia
Oleg Malkov	Institute of Astronomy, RAS, Russia
Alexander Marchuk	Institute of Informatics Systems, RAS, Siberian Branch, Russia
Igor Nekrestjanov	Verizon Corporation, USA
Boris Novikov	St.-Petersburg State University, Russia
Nikolay Podkolodny	ICaG, SB RAS, Novosibirsk, Russia
Aleksey Pozanenko	Space Research Institute, RAS, Russia

Vladimir Serebryakov	Computing Center of RAS, Russia
Yury Smetanin	Russian Foundation for Basic Research, Moscow
Vladimir Smirnov	Yaroslavl State University, Russia
Konstantin Vorontsov	Moscow State University, Russia
Viacheslav Wolfengagen	National Research Nuclear University “MEPhI”, Russia
Victor Zakharov	Federal Research Center “Computer Science and Control” of RAS, Russia

Program Committee

Alexander Afanasyev	Institute for Information Transmission Problems, RAS, Russia
Arkady Avramenko	Pushchino Observatory, Russia
Ladjel Bellatreche	Laboratory of Computer Science and Automatic Control for Systems, National Engineering School for Mechanics and Aerotechnics, Poitiers, France
Pavel Braslavski	Ural Federal University, Yekaterinburg, Russia
Vasily Bunakov	Science and Technology Facilities Council, Harwell, UK
Evgeny Burnaev	Skolkovo Institute of Science and Technology, Russia
Yuri Demchenko	University of Amsterdam, Netherlands
Jerome Darmont	ERIC - Université Lumière Lyon 2
Boris Dobrov	Research Computing Center of MSU, Russia
Alexander Elizarov	Kazan Federal University, Russia
Alexander Fazliev	Institute of Atmospheric Optics, SB RAS, Russia
Yuriy Gapanyuk	Bauman Moscow State Technical University, Russia
Vladimir Golenkov	Belarusian State University of Informatics and Radioelectronics, Belarus
Vladimir Golovko	Brest State Technical University, Belarus
Olga Gorchinskaya	FORS Group, Moscow, Russia
Evgeny Gordov	Institute of Monitoring of Climatic and Ecological Systems, SB RAS, Russia
Valeriya Gribova	Institute of Automation and Control Processes, FEB RAS, Far Eastern Federal University, Russia
Maxim Gubin	Google Inc., USA
Natalia Guliakina	Belarusian State University of Informatics and Radioelectronics, Belarus
Sergio Ilarri	University of Zaragoza, Spain
Mirjana Ivanovic	University of Novi Sad, Serbia
Nadezhda Kiselyova	IMET RAS, Russia
Mikhail Kogalovsky	Market Economy Institute, RAS, Russia
Sergey Kuznetsov	Institute for System Programming, RAS, Russia

Dmitry Lande	Institute for Information Recording, NASU, Ukraine
Evgeny Lipachev	Kazan Federal University, Russia
Giuseppe Longo	University of Naples Federico II, Italy
Natalia Loukachevitch	Lomonosov Moscow State University, Russia
Ivan Lukovic	University of Novi Sad, Serbia
Oleg Malkov	Institute of Astronomy, RAS, Russia
Yannis Manolopoulos	Open University of Cyprus
Archil Maysuradze	Lomonosov Moscow State University, Russia
Manuel Mazzara	Innopolis University, Russia
Alexey Mitsyuk	National Research University Higher School of Economics, Russia
Xenia Naidenova	S. M. Kirov Military Medical Academy, Russia
Dmitry Namiot	Lomonosov Moscow State University, Russia
Igor Nekrestyanov	Verizon Corporation, USA
Panos Pardalos	Department of Industrial and Systems Engineering, University of Florida, USA
Nikolay Podkolodny	Institute of Cytology and Genetics SB RAS, Russia
Jaroslav Pokorny	Charles University in Prague, Czech Republic
Natalia Ponomareva	Scientific Center of Neurology of RAMS, Russia
Alexey Pozanenko	Space Research Institute, RAS, Russia
Andreas Rauber	Vienna Technical University, Austria
Timos Sellis	Swinburne University of Technology, Australia
Vladimir Serebryakov	Computing Centre of RAS, Russia
Nikolay Skvortsov	Federal Research Center “Computer Science and Control” of RAS, Russia, Russia
Manfred Sneps-Sneppe	AbavaNet
Leonid Sokolinskiy	South Ural State University, Russia
Valery Sokolov	Yaroslavl State University, Russia
Alexander Sychev	Voronezh State University, Russia
Bernhard Thalheim	University of Kiel, Germany
Alexey Ushakov	University of California, Santa Barbara, USA
Dmitry Ustalov	University of Mannheim, Germany
Pavel Velikhov	Finstar Financial Group, Russia
Alexey Vovchenko	Federal Research Center “Computer Science and Control” of RAS, Russia, Moscow, Russia
Peter Wittenburg	Max Planck Computing and Data Facility, Garching/Munich, Germany
Anna Yarygina	Saint Petersburg State University, Russia
Vladimir Zadorozhny	University of Pittsburgh, USA
Yury Zagorulko	Institute of Informatics Systems, SB RAS, Russia
Victor Zakharov	Federal Research Center “Computer Science and Control” of RAS, Russia, Russia
Sergey Znamensky	Institute of Program Systems, RAS, Russia

Abbreviations

SB RAS	Siberian Branch of the Russian Academy of Sciences
FRCCSC RAS	Federal Research Center “Computer Science and Control“ of the Russian Academy of Sciences
SAI MSU	Lomonosov Moscow State University, Sternberg Astronomical Institute
NRU HSE	National Research University Higher School of Economics
SRI RAS	Space Research Institute of the Russian Academy of Sciences
MIPT	Moscow Institute of Physics and Technology
INASAN	Institute of Astronomy of the Russian Academy of Sciences
KFU	Kazan Federal University
IMM UB RAS	N. N. Krasovskii Institute of Mathematics and Mechanics of the Ural Branch of the Russian Academy of Sciences
RUDN	Peoples’ Friendship University of Russia
KIAM PSU	Keldysh Institute of Applied Mathematics, Petrozavodsk State University
IPMU	Kavli Institute for the Physics and Mathematics of the Universe, The University of Tokyo
ITEP	Institute of Theoretical and Experimental Physics
ISTP	Institute of Solar Terrestrial Physics, Irkutsk
FAPHI	Fesenkov Astrophysical Institute, Almaty
CrAO	Crimean Astrophysical Observatory, Nauchny, Crimea
SAAO	South African Astronomical Observatory, Cape Town
AbAO	Kharadze Abastumani Astrophysical Observatory, Ilia State University, Tbilisi
JIHT RAS	Joint Institute for High Temperatures of the Russian Academy of Sciences
STFC	Science and Technology Facilities Council, UK
EPFL	École Polytechnique Fédérale de Lausanne
IMET RAS	A. A. Baikov Institute of Metallurgy and Materials Science of Russian Academy of Sciences

Contents

Keynotes and Invited Talks

Advanced Data Analysis Methods

Digital Libraries and Data Infrastructures

Data Integration, Ontologies and Applications

Data and Knowledge Management

Data Analysis in Astronomy

PhD Workshop

Information Extraction from Text I

Information Extraction from Text II

DACOMSIN Workshop

Data Formats, Metadata and Ontologies in Support of Materials Research

Materials Databases, Materials Data Infrastructures and Data Services

IT Applications and IT Platforms for Materials Design and Simulation

Posters, Demonstrations and Networking

KEYNOTES AND INVITED TALKS

How to Put Algorithms into Neural Networks

Anton Osokin[1] $^{[0000-0002-8807-5132]}$

National Research University Higher School of Economics, Russia
aosokin@hse.ru

1 Introduction

Recently, neural networks have achieved remarkable success in many fields [4]. Many practical systems for fundamental tasks are built with neural networks. For example, in computer vision, it is image classification,, object detection and image segmentation; in natural language processing, it is language modeling and automatic translation; in audio processing, both speech recognition and synthesis. Many approaches have become an industrial standard, and companies around the world are building products based on this technology.

Successful algorithms for various tasks are very different from each other and required years of research to arrive at the current level of performance. Constructing a good algorithm for a new task is often a non-trivial challenge. It also turns out that networks can not just learn from data without exploiting some domain knowledge. This knowledge is usually encoded at least in the architecture itself. For example, convolutional neural networks [2] exploit intuition that translation of the object does not change the object itself, i.e., a cat does not stop being a cat if moved left.

At the same time, in many domains we already have powerful algorithms that do a decent job. It is a very natural idea to exploit those to construct better networks. We can look at this from two sides. From one side, this means constructing new layers or blocks of layers for networks. From another side, this means making trainable algorithms. In any case, the attempt is to take best of both worlds. This direction has been around since 90s [1,3,5], but for long time was not getting significant attention (together with neural networks).

In this talk, we will review three ways to combine algorithms and networks (see Fig. 1):

1. structured pooling: an algorithm is used to select active features (similarly to max pooling);
2. unrolling iterations into layers: an algorithm simply becomes a part of the network;
3. analytical derivative w.r.t. the algorithm input, i.e., building a layer with a special backward operator.

To illustrate all the approaches, we will use a running example of a simplified task of handwriting recognition: recognize a word given a sequence of images where each image shows exactly one letter.

Acknowledgments

Supported by RSF project 19-71-30020.

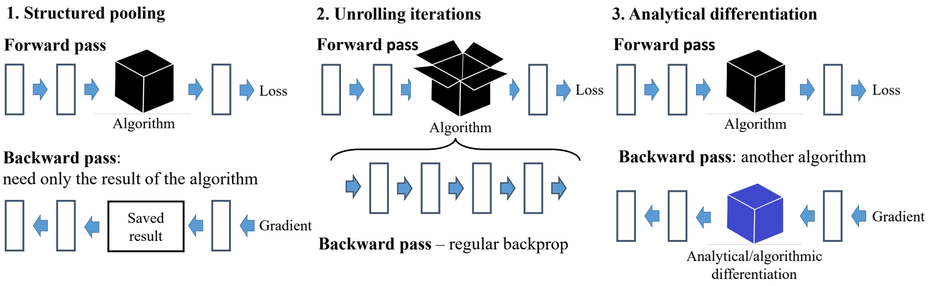

1. Structured pooling
Forward pass
Algorithm
Loss
Backward pass:
need only the result of the algorithm
Saved result
Gradient
2. Unrolling iterations
Forward pass
Algorithm
Loss
Backward pass – regular backprop
3. Analytical differentiation
Forward pass
Algorithm
Loss
Backward pass: another algorithm
Gradient
Analytical/algorithmic differentiation

Fig. 1. Thee approaches to combine an algorithm and a neural network.

References

1. Bottou, L., Le Cun, Y., Bengio, and Y.: Global training of document processing systems using graph transformer networks. In: Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR) (1997).
2. Goodfellow, I., Bengio, Y., and Courville, A.: Deep Learning. MIT Press (2016), http://www.deeplearningbook.org
3. Le Cun, Y., Bottou, L., Bengio, Y., and Haffner, P.: Gradient based learning applied to document recognition. Proceedings of IEEE **86** (11), 2278–2324 (1998).
4. LeCun, Y., Bengio, Y., and Hinton, G.: Deep learning. Nature **521** (7553), 436–444 (2015).
5. LeCun, Y., Chopra, S., Hadsell, R., Ranzato, M., and Huang, F.: A tutorial on energy-based learning. Predicting structured data **1** (0) (2006).

Models for Communication, Understanding, Search, and Analysis

Bernhard Thalheim$^{[0000-0002-7909-7786]}$

Christian-Albrechts University at Kiel, Dept. of Computer Science, D-24098 Kiel, Germany thalheim@is.informatik.uni-kiel.de
http://www.is.informatik.uni-kiel.de/~thalheim

Abstract. Models are one of the universal instruments of humans. They are equally important as languages. Models often use languages for their representation. Other models are conscious, subconscious or preconscious and have no proper language representation. The wide use in all kinds of human activities allows to distinguish different kinds of models in dependence on their utilisation scenarios. In this keynote we consider only four specific utilisation scenarios for models. We show that these scenarios can be properly supported by a number of model construction conceptions. The development of proper and well-applicable models can be governed by various methodologies in dependence on the specific objectives and aims of model utilisation.

1 Introduction

Models are widely used in life, technology and sciences. Their development is still a mastership of an artisan and not yet systematically guided and managed. The main advantage of model-based reasoning is based on two properties of models: they are focused on the issue under consideration and are thus far simpler than the application world and they are reliable instruments since both the problem and the solution to the problem can be expressed by means of the model due to its dependability. Models must be sufficiently comprehensive for the representation of the domain under consideration, efficient for the solution computation of problems, accurate at least within the scope, and must function within an application scenario.

The Notion of Model

Let us first briefly repeat our approach to the notion of model:

A **model** *is a well-formed, adequate, and dependable instrument that represents origins and that functions in utilisation scenarios* [6, 24, 25].

Its criteria of well-formedness, adequacy, and dependability must be commonly accepted by its community of practice within some context and correspond to the functions that a model fulfills in utilisation scenarios.

The model should be well-formed according to some well-formedness criterion. As an instrument or more specifically an artifact a model comes with its *background*, e.g. paradigms, assumptions, postulates, language, thought community, etc. The background its often given only in an implicit form. The background is often implicit and hidden.

A well-formed instrument is *adequate* for a collection of origins if it is *analogous* to the origins to be represented according to some analogy criterion, it is more *focused* (e.g. simpler, truncated, more abstract or reduced) than the origins being modelled, and it sufficiently satisfies its *purpose*.

Well-formedness enables an instrument to be *justified* by an empirical corroboration according to its objectives, by rational coherence and conformity explicitly stated through conformity formulas or statements, by falsifiability or validation, and by stability and plasticity within a collection of origins.

The instrument is *sufficient* by its *quality* characterisation for internal quality, external quality and quality in use or through quality characteristics [23] such as correctness, generality, usefulness, comprehensibility, parsimony, robustness, novelty etc. Sufficiency is typically combined with some assurance evaluation (tolerance, modality, confidence, and restrictions).

A well-formed instrument is called *dependable* if it is sufficient and is justified for some of the justification properties and some of the sufficiency characteristics.

Model Deployment Scenarios are Multi-Facetted

The model notion can be seen as an initialisation for more concrete notions. We observe that model utilisation follows mainly four different kinds of scenarios (see Figure 1). The four scenarios do not occur in its pure and undiffused form they are interleaved. We can however distinguish between:

Problem solving scenarios: Problem solving is a well investigated and well organised scenario (see, for instance, [1, 8]). It is based on (1) a problem space that allows to specify some problem in an application in an *invariant* form and (2) a solution space that *faithfully* allows to back-propagate the solution to the application. We may distinguish three specific scenarios: perception & utilisation; understanding & sense-making, and making your own.

Engineering scenarios: Models are widely used in engineering. They are also one of the main instruments in software and information systems development, especially for system construction scenario. We may distinguish three specific scenarios depending on the level of sophistication: direct application:, managed application, and application according to well-understood technology.

Science scenarios: Sciences have developed a number the distinctive form in which a scenario is organised. Sciences make wide use of mathematical modelling. The methodology of often based on specific moulds that are commonly accepted in the disciplinary community of practice, e.g. [1]. We may distinguish three specific scenarios: comprehension, computation and automatic detection for instance in data science, and intellectual adsorption.

Social scenarios: Social scenarios are less investigated although cognitive linguistics, visualisation approaches, and communication research have contributed a lot. Social models might be used for the development of an understanding of the environment, for agreement on behavioural and cultural pattern, for consensus development, and for social education.

We may distinguish three specific scenarios: development of social acceptance, internalisation & emotional organisation, and concordance & judgement.

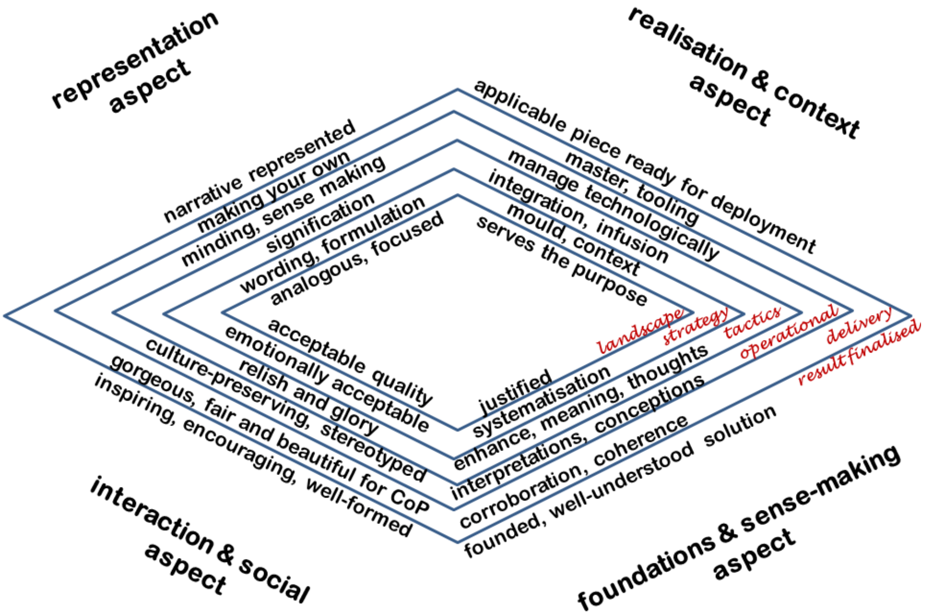


Fig. 1. The four aspects of model scenarios: problem solving, engineering, science, social scenarios. Each of the scenarios combines initialisation by exploring the landscape, strategy, tactics, operational, and delivery layers. Each of these layers adds quality characteristics and specific activities to the previous layer in dependence on the aspects considered

The notion of model mainly reflects the initialisation or landscape layer. Depending on the needs and demands to model utilisation we may distinguish various layers from initialisation towards delivery. The strategy, tactics, operational, and delivery layers are essentially refinements and extensions of the initialisation. The dependability and especially the sufficiency are based on other criteria while the landscape layer is permanent for all models due to the consideration of the concern, the issue, and the specific adaptation to the community of practice, . The strategy layer is governed by the context (e.g. the discipline) and the mould for model utilisation, and the matrix (including methodologies and commonly accepted approaches to modelling). The tactics layer depends on the settlement of the strategy and initialisation layers considers the well-acknowledged experience (e.g. generic approaches), the school of thought or more generally the background, and the framing of the modelling. Which origin(al)s are reflected and which are of less importance is determined in the operational layer that orients on the design and on mastering the modelling process. Finally the model is delivered and form for its application in scenarios that are considered. We thus observe various specific quality characteristics for each of these aspects and layers.

Do We Need a Science of Models and Modelling?

Since everybody is using models and has developed a specific approach to models and modelling within the tasks to be solved, it seems that the answer is "no". From the other side we deeply depend on decisions and understandings that are based on models. We thus might ask a number of questions ourselves. Can models be misleading, wrong, or indoctrinating? Astrophysics uses a Standard Model that has not been essentially changed during the last half century. Shall we revise this model? When? What was really wrong with the previous models? Many sciences use modelling languages in religious manner, e.g. think about UML and other language wars. What is the potential and capacity of a modelling languages? What not? What are their restrictions and hidden assumptions? Why climate models have been deeply changing and gave opposite results compared to the previous ones? Why we should limit our research on impacts of substances to a singleton substance? What is the impact of engineering in this case? What has been wrong with the two models on post-evolution of open cool mines after deployment in Germany which led to the decision that revegetation is far better than water flooding? Why was iron manuring a disaster decision for the Humboldt stream ocean engineering? What will be the impact of the IPCC/NGO/EDF/TWAS proposal for Solar Radiation Management (SRM) for substantial stratosphere obscuration for some centuries on the basis of reflection aerosols (on silver, sulfate, photophoretic etc. basis)? Why reasoning on metaphors as annotations to models may mislead? Are "all models wrong"[1]?

Developing a science of models and modelling would allow us to answer questions like the following one: What is a model in which science under which conditions for whom for which usage at a given time frame? What are necessary and sufficient criteria for an artefact to become a model? What is the difference between models and not-yet-models or pre-models? What is not yet a model? How are models definable in sciences, engineering, culture, ...? Under which conditions we can rely on and believe in models? Logical reasoning: which calculus? Similarity, regularity, fruitfulness, simplicity, what else (Carnap)? Treatment, development, deployment of models: is there something general in common? Models should be useful! What does it mean? Is there any handling of usage,

[1] "All models are wrong. ... Obtain a 'correct' one. ... Alert to what is importantly wrong." [2] We claim: *Models might be 'wrong'. But they are **useful**.*

usefulness, and utility? What is the difference between an object, a model, and a pre-model? What might be then wrong with mathematical models? What is the problem in digging results through data mining methods?

The Storyline of this Paper

Models are the first reasoning and comprehension instruments of humans. Later other instruments are developed. The main one is language. Models then often become language-based if they have to be used for collaboration. Others will remain to be conscious, preconscious or subconscious. Based on the clarification of the given notion of model and a clarification of the model-being we explore in this paper what are the constituents of models, how models are composed, and what are conceptions for model constructions. Since models are used in scenarios and should function sufficiently well in these scenarios we start with an exploration of specific nature of models in four scenarios. We are not presenting all details for a theory of models[2].

2 Case Study on some Scenarios for Model Utilisation

Models are used in various *utilisation scenarios* such as construction of systems, verification, optimization, explanation, and documentation. *In these scenarios* they *function* as *instruments* and thus satisfy a number of properties [7, 26–28].

Models for Communication

The model is used for exchange of meanings through a common understanding of notations, signs and symbols within an application area. It can also be used in a back-and-forth process in which interested parties with different interests find a way to reconcile or compromise to come up with an agreement.

The model has several functions in this scenario: (personal/public/group) *recorder of settled or arranged issues*, *transmitter of information*, *dialogue service*, and *pre-binding*. Users act in the speaker, hearer, or digest mode.

The communication act is composed of six sub-activities: derive for communication, transfer, receive, recognise and filter against knowledge and experience, understand, and integrate. We may distinguish two models at the speaker side and six models at the hearer side: speaker's extracted model for transfer, transferred model for both, hearer's received model, hearer's understanding and recognition model, hearer's filtered model, hearer's understood model, and hearer's integration model. These models form some kind of a model ensemble. Some are extensions or detailing ones; others are zooming ones. Communication is based on some common understanding or at least on transformation of one model to another one.

[2] Collections of papers wich are used as background for this paper is downloadable via Research Gate. Notions and definitions we used can be fetched there.

Models for Understanding

Models may be used for understanding the conceptions behind. For instance, conceptualisation is typically shuffled with discovery of phenomena of interest, analysis of main constructs and focus on relevant aspects within the application area. The specification incorporates concepts injected from the application domain.

The function of a model within these scenario is *semantification* or *meaning association* by means of concepts or conceptions. The model becomes enhanced what allows to regard the meaning in the concept.

Models tacitly integrate knowledge and culture of design, of well-forming and well-underpinning of such models and of experience gained so far, e.g. meta-artifacts, pattern and reference models. This experience and knowledge is continuously enhanced during development and after evaluation of constructs.

Models are functioning for *elaboration, exploration, detection*, and *acquisition* of tacit knowledge behind the origins which might be products, theories, or engineering activities. They allow to understand what is behind drawn curtain.

Models for Search

Users often face the problem that their mental model and their fact space are insufficient to answer more complex questions [12]. Therefore, they seek information in their environment, e.g. from systems that are available. Information is data that have been shaped into a form that is meaningful and useful for human beings. Information consists of data that are represented in form that is useful and significant for a group of humans. This information search is based on their on the *information need*, i.e. a perceived lack of some information that is desirable or useful. The information is used to derive the current *information demand*, i.e. information that is missing, unknown, necessary for task completion, and directly requested. Is is thus related to the task portfolio under consideration and to the intents.

Search is one of the most common facilities in daily life, engineering, and science. It requires to examine the data and information on hand and to carefully look at or through or into the data and the information.

There is a large variety of information search [5] such as:

1. querying data sets (by providing query expressions in the informed search approach),
2. seeking for information on data (by browsing, understanding and compiling),
3. questing data formally (by providing appropriate search terms during stepwise refinement),
4. ferreting out necessary data (by discovering the information requested by searching out or browsing through the data),
5. searching by associations and drilling down (by appropriate refinement of the search terms),
6. casting about and digging into the data (with a transformation of the query and the data to a common form), and
7. zapping through data sets (by jumping through provided data, e.g., by partially uninformed search).

Models for Analysis

Data analysis, data mining or general analysis combines engineering and (systematic) mathematical problem solving [20]. The model development process combines problem specification and setting with formulation of the analysis tasks by means of macro-models, integration of generic models, selection of the analysis strategy and tactics based on methodology models, models for preparation of the analysis space, and model combination approaches for development of the final model society as the analysis result [16, 14]. The typical process model that governs the analysis process is based on a layering approach, e.g. initial setting, strategy, tactics with generic (or general parameterised models), analysis initialisation, puzzling the analysis results, and final compilation. It is similar to experiment planning in Natural Sciences. The analysis puzzling may follow a number of specific scenarios such as pipe scenarios [19].

3 Model Conceptions for These Scenarios

It seems that these scenarios require completely different kinds of models. This is however often not the case. We can develop stereotypes which are going to be refined to pattern and later to templates as the basis for model development. We demonstrate for the four scenarios (communication, understanding, search, analysis) how models can be composed in a specific form and which kind of support we need for model-backed collaboration.

Deep Models

A typical model consists of a normal (or surface) sub-model and of deep (implicit, supplanted) sub-models which represent the disciplinary assumptions, the background, and the context. The deep models are the intrinsic components of the model. Conceptualisation might be four-dimensional: sign, social embedding, context, and meaning spaces. The deep model is relatively stable. In science and engineering it forms the disciplinary background. It is often assumed without mentioning it. For instance, database modelling uses the paradigms, postulates, assumptions, commonsense, restrictions, theories, culture, foundations, practices, and languages as carrier within the given thought community and thought style, methodology, pattern, and routines. This background is assumed as being unquestionable given. The normal model mainly represents those origins that are really of interest.

The deep model combines the unchangeable part of a model and is determined by (i) the grounding for modelling (paradigms, postulates, restrictions, theories, culture, foundations, conventions, authorities), (ii) the outer directives (context and community of practice), and (iii) the basis (assumptions, general concept space, practices, language as carrier, thought community and thought style, methodology, pattern, routines, commonsense) of modelling. The deep model can be dependent on mould principles such as the conceptualisation principle [9].

A typical set of deep models are (the models and) foundations behind the origins which are inherited by the models of those origins. Also modelling languages have there specific deep parts. As well as methodologies or more generally moulds of model utilisation stories.

Model Capsules

Model capsules follow a global-as-design approach (see Figure 2). A model has a number of sub-models that can be used for exchange in collaboration or communication scenarios. A model capsule consists of a main model and exchange sub-models. Model capsules are stored and managed by their owners. Exchange sub-models are either derived from the main model in dependence on the viewpoint, on foci and scales, on scope, on aspects and on purposes of partners or are sub-models provided by partners and transformed according to the main model. A sub-model might be used as an export sub-model (e.g. $A_{4,E}$) that is delivered to the partner on the basis of the import sub-model (e.g. $B_{4,I}$). The sub-models received are typically transformed. We thus use the E(xtract)T(ransform)L(oad) paradigm where extraction and loading is dependent on the language of the sending or receiving model and where transformation allows adaptation of the export sub-model to the import sub-model.

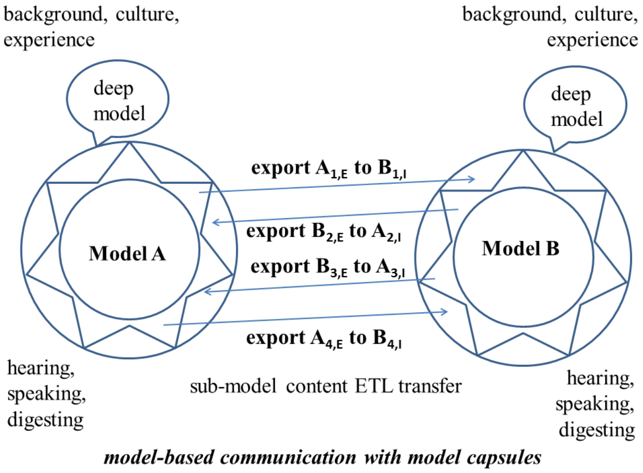

background, culture, experience
background, culture, experience
deep model
deep model
export $A_{1,E}$ to $B_{1,I}$
export $B_{2,E}$ to $A_{2,I}$
export $B_{3,E}$ to $A_{3,I}$
export $A_{4,E}$ to $B_{4,I}$
Model A
Model B
hearing, speaking, digesting
sub-model content ETL transfer
hearing, speaking, digesting
model-based communication with model capsules

Fig. 2. Exchange on the basis of model capsules with sub-models in model-based ETL-oriented communication scenarios

Model Suites

Most disciplines simultaneously integrate a variety of models or a *society of models*, e.g. [3, 11]. The four aspects in Figure 1 are often given in a separate form as an integrated society of models. Models developed vary in their scopes, aspects and facets they represent and their abstraction.

A typical case are the four aspects that might coexist within a complex model. For instance, models in Egyptology [4] [3] can be considered have four aspects where each of the aspects has its specific model. The entire model is an integrated combination of (1,2) signs in textual representation and an extending it hieroglyph form (both as representation), (3) interpretation pattern (as the foundation and integration into the thoughts), (4) social determination (as the social aspect), and (5) a context or realisation models into which the model is embedded. The co-design framework for information systems development (integrated design of structuring, functionality, interaction, and distribution) uses four different interrelated and interoperating modelling languages. These modelling languages are at the same level of abstraction and may be combined with additional orientation on usage (as a social component, e.g. represented by storyboards [21]). In this case, the foundational aspect is hidden within the modelling language and within the origins of the models, for instance in the conceptualisation. Following the four aspects in Figure 1, we derive now models that consider one, two, three, or all four aspects (Figure 3).

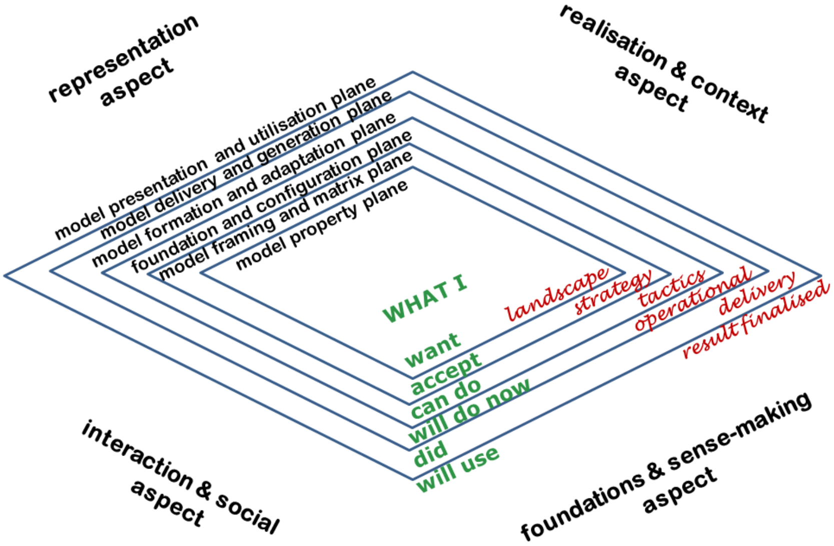

representation aspect
realisation & context aspect
model presentation and utilisation plane
model delivery and generation plane
model formation and adaptation plane
foundation and configuration plane
model framing and matrix plane
model property plane
WHAT I
landscape
strategy
tactics
operational
delivery
result finalised
want
accept
can do
will do now
did
will use
interaction & social aspect
foundations & sense-making aspect

Fig. 3. The four aspect model suite and the corresponding planes for the layers within a model. Activities are governed at each plane by the WHAT I <actually_consider> as main activity.

A *model suite* consists of set of models $\{M_1, ..., M_n\}$, of an association or collaboration schema among the models, of controllers that maintain consistency or coherence of the model suite, of application schemata for explicit maintenance

[3] The rich body of knowledge resulted in [22] or the encyclopedia with [17].

and evolution of the model suite, and of tracers for the establishment of the coherence.

Model suites typically follow a local-as-design paradigm of modelling, i.e. there must not exist a global model which combines all models. In some cases we might however construct the global model as a model that is derived from the models in a model suite. The two approaches to model-based exchange can be combined. A model capsule can be horizontally bound to another capsule within a horizontal model suite or vertically associated to other model capsules. Model capsules are handled locally by members in a team. For instance, model capsules are based on models A and B that use corresponding scientific disciplines and corresponding theories as a part of their background. The models have three derived exchange sub-models that are exported to the other capsule and that are integrated into the model in such a way that the imported sub-model can be reflected by the model of the capsule.

Model Scenes

Model scenes for the development process may be specified in a similar way as storyboarding [21]. A scene is used by members of the community of practice, follows a certain modelling mould, considers a typical ensemble of origin(al)s, inherits certain stereotypes and pattern, is embedded into a context and the tasks, and uses the deep model as the background for model development. These parameters govern and thus control the scene. The developer or modeller is involved into this scene. The input for the scene is the current model, the specific properties of the ensemble of origin(al)s, and especially the experience gained so far. This experience may be collected in a library or generalised to generic models. The output is an enhanced model. We notice that model utilisation scenes can be specified in a similar way.

Figure 4 displays the embedding of a model scene into the model mould or more specifically into the methodology as a macro-model for development.

A model scene is an element of a model story. We imagine that the story can be represented as a graph. A model scene considers an actual or normal model and at the same time the desired embedding into the deep model. The scene is relevant for the community of practice. The model should be accepted by this community. The model scene also embeds the deep model. The scene has its cargo [18], i.e. its mission, determination, meaning, and specific identity. The cargo allows to determine the utility that the model gained so far.

Model Stories

Model development and utilisation can be described as a graph of scenes. Let us consider the model development for search scenarios [12, 13] in Figure 5. This story can be used for derivation of a waterfall-like approach in Figure 6. We start with initialisation of the search landscape. The result is a search guideline (or search activity meta-model). The information demand is transferred to a search question. The search strategy is configured out of the seven kinds of search. The

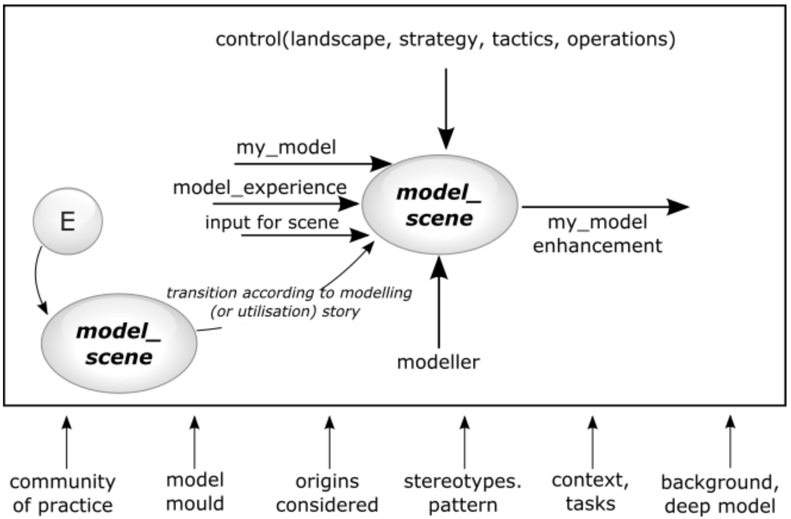


Fig. 4. Model scenes for development (similar to for utilisation)

result is a macro-search model. Selection of the search pattern depends on system information and on the data that is available. The result is a search meso-model, for instance, question-answer forms. Finally we may derive a model on the basis of the data. We might also reconsider the intermediate results and preview or prefetch the potential solutions.

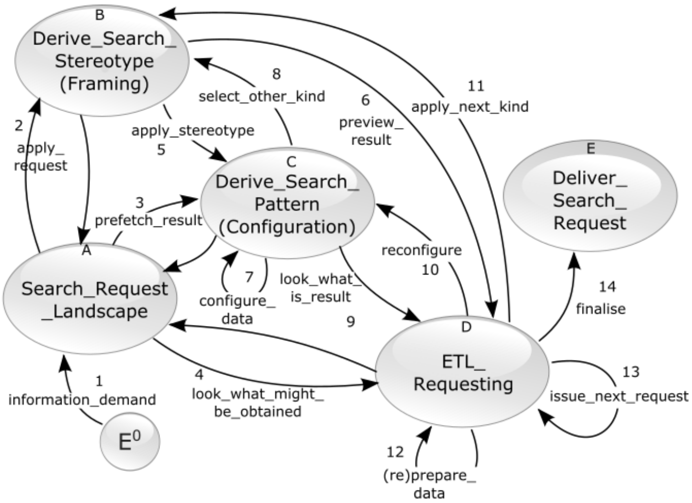


Fig. 5. The layered search story that is used for general search

This story is similar to data mining stories [13, 15]. Data mining uses macro-models as methodological foundation. Frameworks for data mining start with problem specification and setting, continue with formulation of data mining tasks by means of macro-models, reuse generic models according to required adequacy and dependability, next then select appropriate algorithms according to the capacity and potential of algorithms, prepare furthermore the data mining as a process, and finally apply this process. The data mining mould can be supported by controllers and selectors.

Spaces for Models

Figures 1 and 3 use six planes for detailing models. Each of the planes has it specific quality requirements, support tools, and tasks. At the landscape layer we determine the orientation of the model that should be developed, its problem space, its focus and scope, its integration into the value chain of the application (domain), and its stakeholder from the community of practice with their specific interests and their responsibilities, We rely on mental and codified concepts which are often provided by the world of the origins that a model should properly reflect. The strategic layer adds to this the ‘normal way’ of development for utilisation of models as methodologies or mould, the embedding into the context and especially the infrastructure, the disciplinary school of thought or more generally the background of the model. The tactics plane embeds the foundations into the modelling process, for instance, by deep model incorporation. It also allows to sketch and to configure the model. The operational plane orients on the formation of the model and the adaptation to the relevant origin(al)s that are going to represented by the model(s). The main issue for the delivery plane is the design of the model(s). The last plane orients utilisation of the model(s) that have been developed. This outer plane might also be structured according to the added value that a model has for the utilisation scenario. Each plane allows to evaluate the model according to quality characteristics used in the sufficiency portfolio.

The model planes have their own workspace and workplaces which are part of the infrastructure for modelling and utilisation.

4 Model Development

Model Development Story

The modelling story consists of the development story and of the utilisation story. The model development story integrates activities like

1. a selection and construction of an appropriate model according to the function of the model and depending on the task and on the properties we are targeting as well as depending on the context of the intended outcome and thus of the language appropriate for the outcome,
2. a workmanship on the model for detection of additional information about the original and of improved model,
3. an analogy conclusion or other derivations on the model and its relationship to the application world, and

4. a preparation of the model for its use in systems, for future evolution, and for change.

Model utilisation additionally uses assured elementary deployment that includes testing and model detailing and improvement. It may be extended to paradigmatic and systematic recapitulation due to deficiencies from rational and empirical perspectives by the way(s) incommensurability to be resolved. Model deployment also orients on the added value in dependence on the model function in given scenarios. A typical model mould is the mathematics approach to modelling based on (1) exploration of the problem situation, (2) development of an adequate and dependable model, (3) transformation of the first model to a mathematical one that is invariant for the problem formulation and is faithful for the solution inverse mapping to the problem domain, (4) mathematical problem solution, (5) mathematical verification of the solution and validation in the problem domain, and (6) evaluation of the solution in the problem domain [1].

Greenfield Development

Although development from scratch is rather seldom in practice nd daily life we will start with the activities for model development. These activities can be organised in an explorative, iterative, or sequential order in the way depicted in Figure 3. We can separate activities into[4]:

(1) Exploration of the origin(al)s what results in a well-understood domain-situation and perception models: The origin(al)s will be disassembled into a collection of units. We ensemble (or monstrate) and manifest the insight gained so far in a domain-situation model and develop nominal or perception models for the community of practice. It is based on a plausible model proposition, on a selection of appropriate language and of theories, on generic models, and on commonsense structuring.

(2) Model amalgamation and adduction is going to result in a plausible model proposition according to the selected aspects of the four aspects. Amalgamation and adduction are based on an appropriate empirical investigation on origin(al)s, on agreed consensus in the school of thought within the community of practice, on hypothetical reasoning, and on investigative design.

(3) Final model formulation results in an adequate and dependable model that will properly function in the given scenarios. We use appropriate depictions for a viable but incomplete model formulation, extend it by corroborated refinements and modifications, and rationally extrapolate the model in dependence on the given ensemble of origin(al)s. In order to guarantee sufficiency of the model, we assess by elementary and prototypical deployment for proper structuring and dependability, within the application domain, within the boundaries of the background, and within the meta-model or mould for model organisation.

[4] As a generalisation, reconsideration of [10]

A number of moulds can be used for refinement of this development meta-model such as agile or experience-backed methodologies Modelling experience knowledge development might be collected in a later rigor cycle (see design science, for instance, [29]). Model development is an engineering activity and thus tolerates insufficiencies and deficiencies outside the quality requirements. A model must not be true. It must only be sufficient and justified. It can be imperfect.

The result of development can also be a model suite or a model capsule. For instance, information system modelling results in a conceptual structure model, a conceptual functionality model, a logical structure and functionality model, and a physical structure and functionality model. It starts with a business data and process viewpoint model.

Model development can be based on a strictly layered approach in Figure 6 that follows the mould in Figure 5 based on planes in Figure 3.

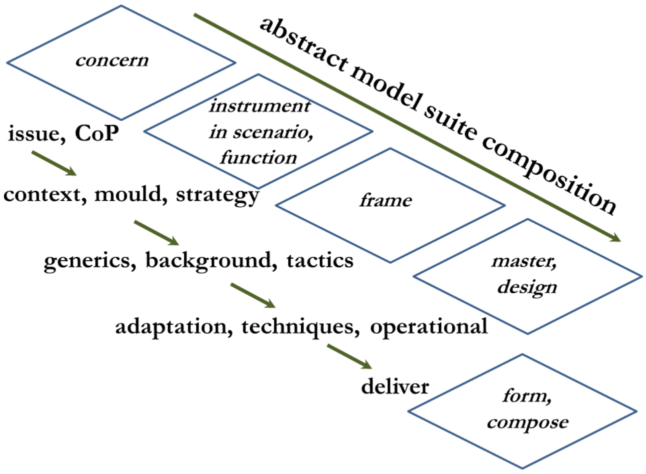

concern
abstract model suite composition
instrument in scenario, function
issue, CoP
context, mould, strategy
frame
generics, background, tactics
master, design
adaptation, techniques, operational
deliver
form, compose

Fig. 6. Model suite development mould for some of the four aspects in Figure 1

Brownfield Development

Modelling by starting from scratch ('greenfield') must be extended by methods for 'brownfield' development that reuses and re-engineers models for legacy systems and within modernisation, evolution, and migration strategies The corresponding model already exists and must be revised. It may also need a revision of its deep sub-model, its basis and grounding, and its ensemble of origin(al)s. All activities used for greenfield development might be reconsidered and revised.

5 Conclusion

Models are widely used and therefore many-facetted, many-functioning, many-dimensional in their deployment, and. Based on a notion of model developed at Kiel university in a group of more than 40 chairs from almost all faculties, we explore now the ingredients of models. The model-being has at least four dimensions which can be grouped into four aspects: *representation* of origins and their specific properties, providing essential *foundations* and thus *sense-making* of origins, relishing and glorifying models as things for *interaction and social collaboration*, and blueprint for *realisation* and constructions within a *context*. This four-aspect consideration directly governs us during introduction of model suites as a model or model capsules. The utilisation scenario and the function of a given model (suite) determine which of the four aspects are represented by a normal model and which aspects are entirely encapsulated in the deep model .

Models are embedded into their life, disciplinary, and technical environment, and their culture. They reuse intentionally or edified (or enlightened) existing sub-models, pre-model, reference model, or generic models. A model typically combines an intrinsic sub-model and an extrinsic extrinsic sub-model. The first sub-model forms the deep model. For instance, database modelling is based on a good number of hidden postulates, paradigms, and assumptions.

The model-being is thus dependent on the scenarios in which models should function properly. We considered here four central scenarios in which models are widely used: communication, understanding, search, and analysis. These four utilisation scenarios can be supported by specific stereotypes of models which model assembling and construction allows a layered mastering of models. The mastering studio has its workspace and its workplace, i.e. in general space for models.

References

1. Berghammer, R. and Thalheim, B.: Wissenschaft und Kunst der Modellierung: Modelle, Modellieren, Modellierung, chap. Methodenbasierte mathematische Modellierung mit Relationenalgebren, pp. 67–106. De Gryuter, Boston (2015).
2. Box, G.E.P.: Science and statistics. Journal of the American Statistical Association **71** (356), 791–799 (1976).
3. Coleman, A.: Scientific models as works. Cataloging & Classification Quarterly, Special Issue: Works as Entities for Information Retrieval **33**, (3-4) (2006).
4. Deicher, S.: The language of objects. BMBF Project KunstModell in Egyptology (2018). https://www.bmbf.de/files/Kurztexte_SdOIII.pdf
5. Düsterhöft, A. and Thalheim, B.: Linguistic based search facilities in snowflake-like database schemes. Data and Knowledge Engineering **48**, 177–198 (2004).
6. Embley, D. and Thalheim, B. (eds.): The Handbook of Conceptual Modeling: Its Usage and Its Challenges. Springer (2011).
7. Feyer, T. and Thalheim, B.: E/R based scenario modeling for rapid prototyping of web information services. In: Proc. ER'99 Workshops. pp. 253–263. LNCS 1727, Springer (1999).
8. Greefrath, G., Kaiser, G., Blum, W., and Ferri, R.B.: Mathematisches Modellieren für Schule und Hochschule, chap. Mathematisches Modellieren – Eine Einführung in theoretische und didaktische Hintergründe, pp. 11–37. Springer (2013).
9. Griethuysen, J.J.V.: The Orange report ISO TR9007 (1982–1987) Grandparent of the business rules approach and sbvr part 2 – The seven very fundamental princi-

ples. https://www.brcommunity.com/articles.php?id=b479 (May 2009), accessed Sept. 21, 2017.

10. Halloun, I.: Modeling Theory in Science Education. Springer, Berlin (2006).
11. Hunter, P.J., Li, W.W., McCulloch, A.D., and Noble, D.: Multiscale modeling: Physiome project standards, tools, and databases. IEEE Computer **39** (11), 48–54 (2006).
12. Jaakkola, H. and Thalheim, B.: Supporting culture-aware information search. In: Information Modelling and Knowledge Bases XXVIII. pp. 161–181. Frontiers in Artificial Intelligence and Applications, 280, IOS Press (2017).
13. Jannaschk, K.: Infrastruktur für ein Data Mining Design Framework. Ph.D. thesis, Christian-Albrechts University, Kiel (2017).
14. Kiyoki, Y. and Thalheim, B.: Analysis-driven data collection, integration and preparation for visualisation. In: Information Modelling and Knowledge Bases. vol. XXIV, pp. 142–160. IOS Press (2013).
15. Kropp, Y. and Thalheim, B.: Data mining design and systematic modelling. In: Proc. DAMDID/RCDL'17. pp. 349–356. FRC CSC RAS, Moscow (2017).
16. Kropp, Y.O. and Thalheim, B.: Deep model guided data analysis. In: DAMDID/RCDL 2017, Revised Selected Papers. Communications in Computer and Information Science, vol. 822, pp. 3–18. Springer (2018).
17. Liepsner, T.: Lexikon der Ägyptologie, vol. IV, chap. Modelle, pp. 168–180. Otto Harrassowitz, Wiesbaden (1982).
18. Mahr, B.: Visuelle Modelle, chap. Cargo. Zum Verhältnis von Bild und Modell, pp. 17–40. Wilhelm Fink Verlag, München (2008).
19. Nissen, I.: Wissenschaft und Kunst der Modellierung: Modelle, Modellieren, Modellierung, chap. Hydroakustische Modellierung, pp. 391–406. De Gryuter, Boston (2015).
20. Podkolsin, A.: Computer-based modelling of solution processes for mathematical tasks (in Russian). ZPI at Mech-Mat MGU, Moscow (2001).
21. Schewe, K.D. and Thalheim, B.: Design and development of web information systems. Springer, Chur (2019).
22. Teeter, E.: Religion and ritual in Ancient Egypt. Cambridge University Press (2011).
23. Thalheim, B.: Towards a theory of conceptual modelling. Journal of Universal Computer Science **16** (20), 3102–3137 (2010) http://www.jucs.org/jucs_16_20/towards_a_theory_of
24. Thalheim, B.: The conceptual model ≡ an adequate and dependable artifact enhanced by concepts. In: Information Modelling and Knowledge Bases. Frontiers in Artificial Intelligence and Applications, 260, vol. XXV, pp. 241–254. IOS Press (2014).
25. Thalheim, B.: Conceptual modeling foundations: The notion of a model in conceptual modeling. In: Encyclopedia of Database Systems. Springer US (2019).
26. Thalheim, B. and Jaakkola, H.: Models and their functions. In: Proc. 29'th EJC. p. 150. LUT, Finland, Lappeenranta, Finland (2019).
27. Thalheim, B. and Nissen, I. (eds.): Wissenschaft und Kunst der Modellierung: Modelle, Modellieren, Modellierung. De Gruyter, Boston (2015).
28. Thalheim, B. and Tropmann-Frick, M.: Wherefore models are used and accepted? The model functions as a quality instrument in utilisation scenarios. In: Comyn-Wattiau, I., du Mouza, C., Prat, N. (eds.) Ingénierie Management des Systèmes d'Information. pp. 131–143. Cépaduès (2016).
29. Wieringa, R.J.: Design Science Methodology for Information Systems and Software Engineering. Springer (2014).

Formal Education in Data Science – Recent Experiences from Faculty of Technical Sciences of University of Novi Sad

Ivan Luković[1][0000-0003-1319-488X]

[1] University of Novi Sad, Faculty of Technical Sciences, Novi Sad, Serbia
ivan@uns.ac.rs

Abstract. In recent years, Data Science has become an emerging education and research discipline all over the world. Software industry shows an increasing and even quite intensive interest for academic education in this area. In this extended abstract, we announce main motivation factors for creating a new study program in Data Science at Faculty of Technical Sciences of University of Novi Sad, and why it is important to nurture the culture of interdisciplinary orientation of such program from early beginning of B.Sc. studies. Also, we announce how we structured the new study program and addressed the main issues that come from evident industry requirements. The program was initiated in 2017, both B.Sc. and M.Sc. studies, and we collect the new experiences.

Keywords: Academic Education, Data Science, Information Engineering.

In 2015, the three study programs in Data Science were accredited at the University of Novi Sad, Faculty of Technical Sciences in Novi Sad. One is a 4-year B.Sc. program in Information Engineering, and the two are master-level study programs: a) 1-year M.Sc. in Information Engineering, and b) 1,5-year M.Sc. in Information and Analytics Engineering. All the programs are officially accredited in the category of interdisciplinary and multidisciplinary programs, in the two main areas: Electrical Engineering and Computing, and Engineering Management. In practice, the programs cover in deep the disciplines in Data Science, as a completely new combination of courses predominantly coming from Computer Science, Software Engineering, Mathematics, Telecommunications and Signals, Finances, and Engineering Management.

Execution of the two of these study programs has been initiated in 2017. Those are B.Sc. in Information Engineering, and M.Sc. in Information Engineering. By this, now we have two active generations of students at both levels. Our first experiences with these generations of students are quite positive.

Design of the aforementioned study programs were motivated mainly by the idea to profile specific study programs in the scope of Computer Science, Informatics, or Software Engineering (CSI&SE) disciplines, so as to nurture the appropriate level of interdisciplinarity and contribute to resolving the following two paradoxes: (P1) More interdisciplinary oriented experts, capable of covering a wide range of tasks, knowledge and skills are always significantly better positioned in the software industry Human Resource (HR) market, while academic institutions offer study programs that are rather self-contained, i.e. oriented to a narrower knowledge scope. (P2) Students or young software engineers believe that they will be better positioned in software industry HR market just as they are good IT experts, i.e. programmers, while employers rather expect experts capable of recognizing and resolving their interdisciplinary oriented and complex requirements.

In our current academic education, we can identify study programs of the three categories, covering in some extent disciplines of CSI&SE, as a basis to provide Data Science education. Those are: (1) Specific study programs in CSI&SE; (2) Study programs in (Applied) Mathematics; and (3) Study programs in Economics, Business Administration and Management. Our experiences in teaching CSI&SE courses in study programs of all the three categories lead to the identification of typical students' and even teachers' behavioral patterns. In the paper we will discuss why such patterns lower the culture of interdisciplinarity, and how it can be raised by data science study programs. Also, we will transfer some our recent experiences from the execution of Information Engineering study programs, where we identify increasing awareness of students about the importance of Data Science in upcoming years, while still the polarization of students' population to one, with clear ideas about their future, vs. students with not clear recognition of their future opportunities is present.

Acknowledgements

The research presented in the paper was supported by the Ministry of Education, Science, and Technological Development of the Republic of Serbia under Grant III-44010, “Intelligent Systems for Software Product Development and Business Support based on Models”.

References

1. Luković, I.: Formal Education in Data Science – A perspective of Serbia, In Proceedings of Milićević I. (eds.) 7th International Scientific Conference Technics and Informatics in Education, Čačak, Serbia, University of Kragujevac, Faculty of Technical Sciences Čačak, ISBN: 978-86-7776-226-1, pp. 12–18, Čačak (2018).
2. Smith, J.: Data Analytics: What Every Business Must Know About Big Data and Data Science. Pinnacle Publishers, LCC, (2016). ISBN: 978-1-535-11415-8
3. Anderson, P., Bowring, J., McCauley, R., Pothering, and G., Starr C.: An undergraduate degree in data science: curriculum and a decade of implementation experience. In Proceedings of the 45th ACM technical symposium on Computer science education. ACM, USA, pp. 145–150 (2014). doi: 10.1145/2538862.2538936
4. Group of authors: Informatics education: Europe cannot afford to miss the boat. Report of the joint Informatics Europe & ACM Europe Working Group on Informatics Education. (2013).
5. Golshani, F., Panchanathan, S., and Friesen, O.: A Logical Foundation for an Information Engineering Curriculum. In Proceedings of 30th ASEE/IEEE Frontiers in Education Conference, USA, (2000). doi: 10.1109/FIE.2000.897639.

Big Data Processing and Analytics Inside DBMS

Mikhail Zymbler(✉)[0000-0001-7491-8656], Sachin Kumar,
Yana Kraeva, Alexander Grents, and Anastasiya Perkova

South Ural State University, Chelyabinsk, Russia
mzym@susu.ru, sachinagnihotri16@gmail.com, kraevaya@susu.ru,
grentsav@susu.ru, perkovaai@susu.ru

In the era of Big Data, there are two basic challenges for humans: how to effectively manipulate and analyze huge amounts of data. Currently, relational DBMSs remain the most popular tool for processing large tables in various data intensive domains, despite the widespread use of numerous NoSQL systems. At the same time, most of modern tools for mining the large data sets are non-DBMS and based on the MapReduce paradigm. If we consider DBMS only as a fast and reliable data repository, we get significant overhead for export large data volumes outside a DBMS, changing data format, and import results of analysis back into DBMS. That is why integration of data mining methods with relational DBMS is a topical issue.

There exist parallel DBMSs that can efficiently process transactions and SQL queries on very large databases. Such DBMSs could be a subject for integration of data mining methods but they are expensive and oriented to custom hardware that is difficult to expand. Open-source DBMSs are now being a reliable alternative to commercial DBMSs but there is a lack of open-source parallel DBMSs since the development of such software is rather expensive and takes a lot of time.

In the talk, we will consider an approach to deal with the problems described above. A parallel DBMS can be developed not from scratch but by small-scale modifications of the original codes of an open-source serial DBMS to encapsulate parallelism. Large- and small-scale data mining problems can be solved inside such a parallel DBMS.

Acknowledgments

This work was financially supported by the Russian Foundation for Basic Research (grant No. 17-07-00463) and by the Ministry of Science and Higher Education of the Russian Federation (government orders 2.7905.2017/8.9 and 14.578.21.0265).

ADVANCED DATA ANALYSIS METHODS

Framework for Automated Food Export Gain Forecasting

Dmitry Devyatkin [1] and Yulia Otmakhova [2]

[1] Federal Research Centre “Computer Science and Control” RAS, Moscow, Russia
[2] Novosibirsk State University, Novosibirsk, Russia
[1]devyatkin@isa.ru
[2]otmakhovajs@yandex.ru

Abstract. The food and agriculture could be a driver of the economy in Russia if intensive growth factors were mainly used. In particular, it is necessary to adjust the food export structure to fit reality better. This problem implies long-term forecasting of the commodity combinations and export directions which could provide a persistent export gain in the future. Unfortunately, the existing solutions for food market forecasting tackle mainly with short-term prediction, whereas structural changes in a whole branch of an economy can last during years. Long-term food market forecasting is a tricky one because food markets are quite unstable and export values depend on a variety of different features.

The paper provides a multi-step data-driven framework which uses multimodal data from various databases to detect these commodities and export directions. We propose the quantile nonlinear autoregressive exogenous model together with pre-filtering to tackle with such long-term prediction tasks. The framework also considers textual information from mass-media to assess political risks related to prospective export directions. The experiments show that the proposed framework provides more accurate predictions then widely used ARIMA model. The expert validation of the obtained result confirms that the framework could be useful for export diversification.

Keywords: data-driven market forecasting, international trade, quantile regression, multimodal data

1 Introduction

Sanctions and trade confrontations set difficulties for persistent economic growth. The essential way to overcome them is making the economy more independent and diversified [1]. Due to limited resources, the efforts should be focused on a limited set of development directions. Therefore, the developing of a particular economy field implies discovering a restricted set of the new prospective commodity items and export directions. In this paper, we consider this problem in the case of food and agriculture field. Thus our aim consists in finding the pairs <*Trade partner, Agriculture_OR_FoodCommodity*> with a high probability of the persistent export value growth from a particular country (in our case – from Russia) in several next years.

More precisely, we predict summary export value gain in the following two years based on information about current and two past years. We believe that modern data-driven approaches could be useful to tackle this problem.

This goal is not trivial because of the following issues:

1. Unstable character of many trade flows.

2. Too many features influence on trade flows. If we used them all, it would lead to over-complex prediction models, which aren't trainable with the dataset. Long-term forecasting requires the using of complex models that consider a large number of features and parameters, but the size of the training dataset is strictly limited. Therefore, complex models can be easily overfitted and in some cases give incorrect results on unseen data.

3. Political decisions, economic sanctions strongly affect trade flows, but they hardly ever can be predicted using only statistic databases.

4. Existing regression and classification metrics such as MSE or F1-score poorly reflect the accuracy of the solution to the highlighted problem, since even a small ranking error can lead to the omission of a very profitable direction.

In this paper, we propose a data-driven framework which can mitigate the highlighted problems.

At first, we apply a quantile regression loss since it allows estimating the distribution parameter for the predicted value so that we can process unstable trade flows more accurately.

Secondly, we believe it is possible to mitigate the overfitting problem and instability problems both if one pre-filters pairs with high probabilities of a decline in the future. This can be done with training a binary classifier, which is much simpler than regression and can be performed using simpler models which are not overfitted. Then the "large" errors of the regression model will have less impact on the final result. We also propose compositional features which can describe the market demand for a commodity item compactly to simplify the regression model.

Thirdly, we extract sentiment features from texts, more precisely, from news to assess political risks.

Finally, we suggest calculating ratios between the export value of the top predicted pairs and the export value of the actual top pairs with the highest export gain to assess the usefulness of the prediction.

The rest of the paper is organized as follows: in Section 2 we review related studies; in Sections 3 we present the proposed framework; in Section 4 we describe the results of the experimental evaluation; Section 5 contains conclusion and directions of the future work.

2 Related Work

The vastest branch of studies is devoted to short-term food market forecasting with basic regression and autoregression models. For example, Mor with colleagues propose linear regression and Holt–Winters' models to predict short-term demand for dairy products [2]. The more complex autoregressive integrated moving average model (ARIMA) allows dealing with non-stationarity time series. This model also widely used for food market forecasting, for example, in [3] to forecast harvest prices based on past monthly modal prices of maize in particular states.

Ahumada et al. [4] proposed an equilibrium correction model for corn price. Firstly they use an independent model for each corn. Then they also observe whether the forecasting precision of individual price models can be improved by considering their cross-dependence. The results show that prediction quality can be improved using models that include price interactions. The multi-step approach is proposed in [5]. The researchers consider the balance between production and market capacity to be the key factor for trade flows forecasting.

For forecasting in volatile markets, it is necessary to reveal detail information about the distribution of the predicted variables, not their mean values only. Quantile regression is a common solution in this case. [6], [7]. For example, researchers [8] apply linear quantile loss to train Support Vector Regressor and use it to assess confidence intervals for predicted values. The paper [9] combines hybrid ARIMA and Quantile Regression (ARIMA-QR) approaches to construct high and low quantile predictions for non-stationary data. The obtained results show that the model yield better forecasts at out-sample data compared to baseline forecasting models.

Let us briefly highlight some studies related to features for food market forecasting which can better explain trade flow dynamic than trade and production values themselves. Paper [10] provides a conclusion that finance matters for export performance, as commodities with higher export-related financial needs disproportionately benefit from better economic development. Jaud with colleagues uses level of outstanding short-term credit and trade credit insurance, reported in the Global Development Finance and Getting Credit Index (EGCc) from the World Bank Doing Business Survey as features related to the level of financial development.

Political factors also influence on the food market. Makombe with colleagues studies the relationship between export bans and food market [11]. The researchers conclude that the prohibitions cause market uncertainty which may have long-run implications for future food security and trade flows. The critical problem here lies in uncertainty in the way how to formalize and consider these factors in models. It is well-known fact, that political decisions often follow by outbursts in mass media, so one can easily predict possible political decisions if he or she analyses the new sentiment. This idea is widely used for short-term analysis in financial markets [12], thus we believe it could be helpful in the proposed framework.

Long-term export prediction assumes considering arbitrary dependencies between the model outcome and the lots of factors in the past. Duration of these dependencies can vary from single days for price movements to dozens of years for political decisions or climate changing. The mentioned approaches cannot model linear, non-linear dependencies and consider a broad set of sophisticated features at the same time though. A natural way to model such complex features and dependencies is to use neural network framework. Pannakkong with colleagues [13] uses a dense multilayer feed-forward network and ARIMA to forecast cassava starch export value. The results show that feed-forward neural network models overcome the ARIMA models in all datasets. Hence, the neural network models can predict the cassava starch exports with higher accuracy than the baseline statistical forecasting method such as the ARIMA. However, such a simple architecture cannot model long-term interaction.

There are particular network architectures for long-term prediction. In [14], researchers suggested the nonlinear autoregressive exogenous model (NARX) artificial network architecture for market forecasting. They proposed a feed-forward Time Delay Neural Network, i.e. the network without the feedback loop of delayed outputs, which could reduce its predictive performance. The main benefit of the model compared to model compositions is the ability of joint training of linear and non-linear parts of models. Similarly, in [15] authors proved that the generalized regression neural network with fruit fly optimization algorithm (FOA) is effective for forecasting of the non-linear processes.

Unfortunately, neural network approach often leads to inadequately complex models which are needed large datasets to be reliably fitted. We have relatively small dataset thus it is required to find the most straightforward network architecture and tightest feature set which however could achieve satisfactory forecast accuracy.

Because food market is volatile, it would be helpful if the forecasting model provided more information about predicted variables as quantile regression does. Although there are few works in which quantile regression-like loss function was used for training neural networks [16].

The methodological aspects of creating models for export forecasting require further study. Existing models consider some important indicators, but they can be based on erroneous assumptions that cast doubt on the obtained results. For example, the predictive model for assessing the country's diversification of exports provided in the Atlas of Economic Complexity (Feasibility charts) [17]. This model predicts a very curious output, namely that tropical palm oil could be one of the products for diversifying Russia's exports. This is due to the neglecting country's climatic and infrastructure capabilities. That is why the feature set is still not obvious for this problem.

The review shows that the most applicable solution for the food export gain forecasting is to combine long-term prediction models, such as NARX and quantile loss functions. In addition to basic features such as trade flows and production levels, these models should consider heterogeneous macroeconomic and climate indicators. Since the addition of political factors would complicate the regression model, it makes sense to consider them separately. That is, after the regression, we filter obtained export directions if they are related to high political risks.

3 Framework for Export Gain Forecasting

As a test dataset for the framework, we use annual information about trade flows (from UN Comtrade [18]), production values (UN FAOSTAT [19]) and macroeconomic indicators (International Monetary Foundation [20]). We consider the following macroeconomic features: state GDP, inflation level, population level etc. Due to heuristic reasons the dataset includes only the items which are produced in Russia and presented in its trade flows, so the final dataset contains 70 export directions and 50 commodities. We also do not consider records earlier than 2009, because the international financial crisis could lead to changes, which we cannot model adequately. Daily climate (temperature, wind speed, humidity, pressure) features were downloaded from RP-5 weather database [21]. The highest, the lowest and average values for each season were calculated, because the time step of the framework is one year. We also used open-available Russian news corpus from Kaggle [22].

It is no doubt to say that trade flows between particular country and its partners depends on trade flows between these partners and the other countries. Unfortunately, if we added all these features directly to the regression model, the model would become too complex and would tend to overfit in the dataset. We propose the *SPR* (Substantial PRoduct) composite features to resolve this problem. The *SPR* shows contribution of an arbitrary exported commodity item from Russia on the global demand satisfying (expression (1)):

$$SPR_i = \frac{X_i}{\sum_{i \in I, j \in D} C_{ij}}. \quad (1)$$

Here I is a set of leading export commodities, D is a set of export directions, X_i is total Russian export value for commodity i, C_{ij} is export value from Russia to country j for commodity i. We consider all trade flows between Russia and its' partners directly and encode the rest flows with the *SPR* features. The comprehensive feature list is presented in Table 1.

Table 1. Feature set for export gain forecasting

Group of features	**Frequency**	**For**	**Features**
Trade flows	Annual	Country, Commodity	Export value
			Import value
			Re-export value
			Re-import value
SPR	Annual	Commodity	SPR
Production	Annual	Country, Commodity	Production value
Macro-economic indicators	Annual	Country	Trade balance
			GDP
			Inflation (CPI)
			Inflation (PPI)
			Population
			Purchasing power parity (PPP)
			Unemployment rate
Climate indicators	Per season: max, min, median	City, town	Temperature
			Humidity
			Wind speed
			Precipitation
			Pressure
			Cloudiness

We propose the following framework for prediction of the promising pairs <*export item,direction*> (Fig. 1). The framework contains regression step and several filtering steps. Pre- and post-filtering steps are proposed to deal with the trade flows instability.

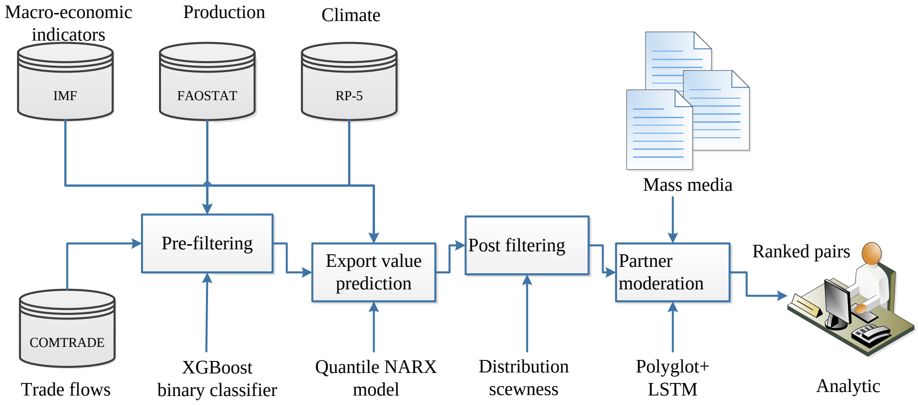


Fig. 1. Framework for prospective export pairs prediction

At first step of the framework we detect pairs which likely tend to decrease. On the one hand the filtering model should be much simpler, than the regression model, but on the other hand it should learn complex non-linear dependencies. The most appropriate approach in this case is decision tree ensembles. We tested several methods such as Random Forest [23], Gradient Boosting [24] and XGBoost [25] to fit these ensembles. The next two steps we realize with the modified NARX quantile regression model (Fig. 2). A single model is used for all directions and commodities since the use of individual models can lead to the loss of information about the interaction between the export value for commodities. We used the following loss function instead of mean squared error to obtain quantile NARX model:

$$L(\omega,\theta) = \sum_{t \vee y_{t \geq f(x_t,\omega)}} \theta(y_t - f(x_t,\omega)) + \sum_{t \vee y_{t<f(x_t,\omega)}} (1-\theta)(y_t - f(x_t,\omega)\,), \quad (2)$$

here θ is quantile level, x_t is features for time *t*, $f(x_t,\omega)$ is network output for time *t* and ω is parameters of the network.

This function was firstly introduced in [16]; it is a direct application of quantile regression [6] for networks training. Thanks to error-backpropagation framework the network architecture does not have any affection on the function (2). The modified NARX model allows predicting values for different quantile levels. We predict export flows with quantile levels 0.25, 05 and 0.75 and assessed skewness of the results. Than pairs with positive distribution skew are filtered. We also applied the Autoregressive Integrated Moving Average (ARIMA) model as a baseline.

In the last step, we filter unreliable trade partners with various models for sentiment analysis. We tested two neural network models, namely Attention-based Long-Short Term Memory (LSTM) [26] and Contextual LSTM [27]. Polyglot sentiment analyzer was also tested as a baseline. We used the Kaggle corpus with more than

10K news reports in Russian to train these models. Post-filtering itself consists of two parts. At first, we apply the Polyglot library [28] together with country name dictionary to extract news, mentioned Russia and some other country together. Then we apply a sentiment analyzer and filter trade partners with highly negative sentiment scores from the results.

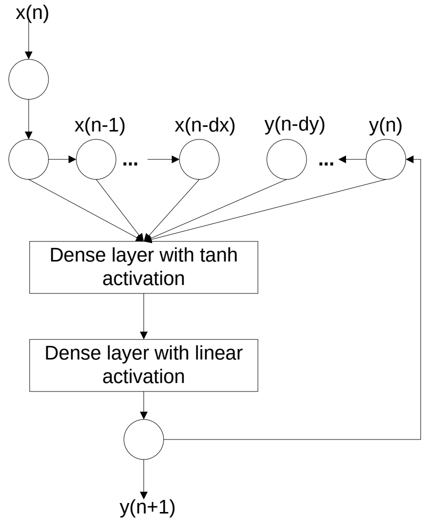

x(n)
x(n-1)
x(n-dx)
y(n-dy)
y(n)
...
...
Dense layer with tanh activation
Dense layer with linear activation
y(n+1)

Fig. 2. Nonlinear autoregressive exogenous model for export value forecasting

4 Results and Discussion

We evaluated the proposed framework as well as its crucial parts. At first, the classification performance for the filtering step was evaluated. We tested several decision tree ensembles (Random Forest, Gradient Boosting, XGBoost) and Linear Support Vector Machine (SVM) classifier as a baseline. The XGBoost method revealed the best accuracy in cross-validation, so we add it to the framework. It is worth to note that the filtering itself can be done with relatively high quality (about 73% F1 on 5-fold cross-validation, see Table 2, “filtering” column) using a simple feature-set because this task is much simpler than the whole regression task.

Table 2. Evaluation results for filtering methods

Method for filtering	**F_1-binary**
Random Forest	0.64±0.04
Gradient Boosting	0.59±0.02
XGBoost	0.73±0.05
LinearSVM	0.63±0.02

We also assessed the importance of different types of features for the filtering. The classifier was trained and tested on modified feature sets, in which distinct group of features had been omitted (see Table 3, “filtering” column). This column contains difference between binary F_1 score obtained on the full feature set and the score obtained on clipped feature set. The higher the F_1 score drop is, the more important related subset of features is. Results show that the most important features are SPR, climate and macro-economic indicators.

Table 3. Importance of the particular feature groups for the pre-filtering and for regression results

	Filtering	**Regression**
Group of features	**ΔF_1**	**ΔPredicted export value gain, in %**
Trade flows	0.05	7.1
SPR	0.17	6.8
Production	0.12	1.2
Macro-economic indicators	0.29	24.7
Climate indicators	0.05	17.5

Then we studied the importance of the different types of features for the regression. As for filtering step, we separated features into distinct subsets and trained the regressor with them (see Table 3, “regression” column). The “Predicted export value gain” here and in the next tables means ratios between the export value of the top-10 predicted pairs and export value of the top-10 actual pairs with the highest export gain. The “ΔPredicted export value gain” column contains difference between the gain obtained on the full feature set and the gain obtained on clipped feature set. The results showed that the most significant features are macro-economic indicators, climate and past export flows. This confirms the limitation of models which do not consider these features, for example [17].

Table 4. Importance of the particular filtering steps

Filtering	**Predicted export value gain, %**
Without	12.5
Pre-filtering (only)	38.9
Post-filtering (only)	22.7
All (combined)	61.6

We also evaluated the contribution of the filtering steps to the results. The filtering steps together help significantly improve the obtained results, as one can conclude from Table 4. These steps allow removing the pairs with the highest decline risk.

Table 5 contains results for considered sentiment analysis methods. We evaluated these methods on the test subset of the Kaggle sentiment dataset. Attention LSTM model shows slightly better result on this task, so we added it to the proposed framework.

Table 5. Results for the sentiment analysis

Method	**F_1-macro**
Polyglot sentiment	0.62
CLSTM	0.79
Attention LSTM	0.82

Fig. 3 shows average sentiment for mass-media news related to main trade partners of Russia, which are often mentioned in Russian news. The news was gathered from Lenta.ru dataset [29], because the timeline of the Kaggle dataset is not appropriate. We filtered news, contained both “Russia” and other country names and evaluated average sentiment for them with the Attention LSTM model. It’s easy to note that unreliable partners get lower marks.

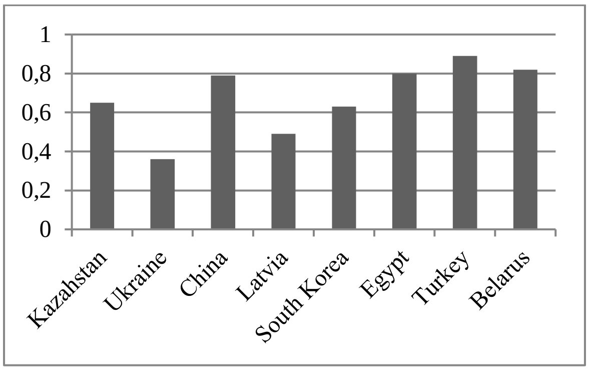


Fig. 3. Partner assessment with sentiment analysis

We tested the overall framework on retrospective data, more precisely records from 2009 to 2014 were used to train and other data (2015–2016) we left for the evaluation. The detailed results for the whole framework are presented in Table 6. The “Actual” column contains ranked pairs with the highest average export gain in 2015–2016 for Russian Federation. Summary average gain for the top 10 pairs amounted to 1.5 billion USD. The “Predicted” columns contain results of the forecasting.

The economic analysis of the detailed results showed that the list of the partners with the highest summary export gain did not match with the list of top importers for the study period.

The proposed NARX model allows predicting the most growing export commodities quite precisely. Linseed is the only mismatched position, but it reflects a new prospective market. Moreover, linseed production for export has been strongly supported by the Russian government since 2016. Thereby the model detected this potential market with past data and predicted that decision.

Table 6. Detailed results of the export value prediction

		Predicted			
Actual		**ARIMA**		**NARX + pre-filtering + quantile-filtering**	
Partner	Commodity	Partner	Commodity	Partner	Commodity
Bangladesh	Wheat	Egypt	Wheat	Egypt	Wheat
Egypt	Wheat	Bangladesh	Wheat	Saudi Arabia	Barley
China	Soybeans	China	Soybeans	Nigeria	Wheat
Nigeria	Wheat	Turkey	Wheat	Morocco	Wheat
China	Oil of Sunflower Seed	China	Oil of Sunflower Seed	Sudan	Maize
Rep. of Korea	Maize	Algeria	Oil of Soybeans	Turkey	Barley
Lebanon	Wheat	Azerbaijan	Wheat	Turkey	Linseed
Algeria	Oil of Soybeans	Saudi Arabia	Barley	Bangladesh	Wheat
Saudi Arabia	Barley	Lebanon	Wheat	Italy	Wheat
China	Oil of Soybeans	China	Oil of Soybeans	China	Oil of Soybeans
Export value gain, M USD	1474.4; 100%	674.8; 45.7%		908.65; 61.6%	

The most often cause of the NARX model errors is neglecting features, related to technological development. However, the appearance of new technologies leads to dramatic changes in the markets. New deep-processed commodities appear, and prices for existing raw products can decline, which leads to an export value drop for traditional providers. Existing counterparties (Turkey, for example) may switch to other commodities. Therefore, there is a need to add technological features to the model, which would make it possible to predict prospect commodities with an assessment of the related technologies for primary and deep processing.

To sum up, the results of the proposed framework could be useful for export diversification since NARX model provides new prospective commodity items. The variants of the NARX model and ARIMA are also helpful for counterparty countries exploration.

5 Conclusion

In this paper, we propose a data-driven framework for food export gain forecasting. The framework considers multimodal open data from many data sources and corpora. In this research, we tried to mitigate the set of problems, related to machine forecasting of food export gain: large feature set dimension, volatility of markets, factors which are difficult to formalize (political risks).

In the experiments, we used open data from FAOSTAT, UN Comtrade, information about global economic situation from International Monetary Foundation, climate information and reports from news corpora. According to the results, quantile loss function and NARX model is a promising combination for long-term prediction of trade flows for food commodities.

In the future research we plan to consider logistical and infrastructure conditions as well as technological features in the framework. The next steps of our research also include detailed analysis of the obtained commodity items and finding technologies which could help to push the export for these commodities up.

Acknowledgements

The project is supported by the Russian Foundation for Basic Research, project No. 16-29-12877 “ofi_m”.

References

1. Awokuse, T.: Does agriculture really matter for economic growth in developing countries? In: 2009 Annual Meeting. Agricultural and Applied Economics Association, vol. 49762. Milwaukee, Wisconsin (2009).
2. Mor, R. and Bhardwaj, A.: Demand forecasting of the short-lifecycle dairy products. In: Chahal, H., Jyoti, J., Wirtz, J. (eds.) Understanding the Role of Business Analytics, pp. 87–117. Springer, Singapore (2019).
3. Darekar, A. and Reddy, A.: Price forecasting of maize in major states. Maize Journal **6** (1&2), 1–5 (2017).
4. Ahumada, H. and Cornejo, M.: Forecasting food prices: The case of corn, soybeans and wheat. International Journal of Forecasting **32** (3), 838–848 (2016).
5. Burlankov, S., Ananiev, M., Gazhur, A., Sedova, N., and Ananieva, O.: Forecasting the development of agricultural production in the context of food security. Scientific Papers Series-Management, Economic Engineering in Agriculture and Rural Development **18** (3), 45–51 (2018).
6. Koenker, R. and Hallock, K.: Quantile regression. Journal of economic perspectives **15** (4), 143–156 (2001).
7. Maciejowska, K., Nowotarski, J., and Weron, R.: Probabilistic forecasting of electricity spot prices using Factor Quantile Regression Averaging. International Journal of Forecasting **32** (3), 957–965 (2016).
8. Li, G. Xu, S., Li, Z., Sun, Y., and Dong, X.: Using quantile regression approach to analyze price movements of agricultural products in China. Journal of Integrative Agriculture **11** (4), 674–683 (2012).

9. Arunraj N. and Ahrens D.: A hybrid seasonal autoregressive integrated moving average and quantile regression for daily food sales forecasting. International Journal of Production Economics **170**, 321–335 (2015).
10. Jaud, M., Kukenova, M., and Strieborny, M.: Financial Development and Sustainable Exports: Evidence from Firm-product Data. The World Economy **38** (7), 1090–1114 (2015).
11. Makombe, W. and Kropp, J.: The effects of Tanzanian maize export bans on producers' welfare and food security. In: Selected Paper prepared for presentation at the Agricultural & Applied Economics Association, vol. 333-2016-14428. Boston, MA (2016).
12. Nassirtoussi, A., Aghabozorgi, S., Yuing Wah, T., and Chek Ling Ngo, D.: Text mining of news-headlines for FOREX market prediction: A Multi-layer Dimension Reduction Algorithm with semantics and sentiment. Expert Systems with Applications **42** (1), 306–324 (2015).
13. Pannakkong, W., Huynh, V., and Sriboonchitta S.: ARIMA versus artificial neural network for Thailand's cassava starch export forecasting. Causal Inference in Econometrics, pp. 255–277. Springer, Cham (2016).
14. Menezes, Jr J. M. P. and Barreto, G.: A. Long-term time series prediction with the NARX network: An empirical evaluation. Neurocomputing **71** (16–18), 3335–3343 (2008).
15. Li, H., Guo, S., and Sun, J.: A hybrid annual power load forecasting model based on generalized regression neural network with fruit fly optimization algorithm. Knowledge-Based Systems **37**, 378–387 (2013).
16. Taylor, J.W.: A quantile regression neural network approach to estimating the conditional density of multiperiod returns. Journal of Forecasting **19** (4), 299–311 (2000).
17. The atlas of economic complexity, http://atlas.cid.harvard.edu, last accessed 2019/07/01
18. UN Comtrade: International Trade Statistics, https://comtrade.un.org/data/, last accessed 2019/04/28
19. Food and Agriculture Organization of the United Nations, http://www.fao.org/faostat/en/ last accessed 2019/04/28
20. International monetary foundation, http://www.imf.org/en/Data, last accessed 2019/04/28
21. RP5 weather archive, http://rp5.ru, last accessed 2019/04/28.
22. Kaggle Russain news dataset for sentiment analysis, https://www.kaggle.com/c/sentiment-analysis-in-russian/overview, last accessed 2019/04/28
23. Breiman, L.: Random forests. Machine learning **45** (1), 5–32 (2001).
24. Friedman, J.H.: Greedy function approximation: a gradient boosting machine. Annals of statistics, 1189–1232 (2001).
25. Chen, T. and Guestrin, C.: Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd ACM sigkdd international conference on knowledge discovery and data mining, pp. 785–794. ACM (2016).
26. Wang, Y., Huang M., Zhao L., and Zhu X.: Attention-based LSTM for aspect-level sentiment classification. In: Proceedings of the 2016 conference on empirical methods in natural language processing, pp. 606–615. Association for Computational Linguistics, Austin, Texas (2016).
27. Ghosh, S., Vinyals, O., and Strope B.: Contextual LSTM (CLSTM) models for large scale NLP tasks. In: arXiv preprint arXiv:1602.06291. ACM (2016).
28. Al-Rfou, R., Kulkarni V., and Perozzi, B.: Polyglot-NER: Massive multilingual named entity recognition. In: Proceedings of the 2015 SIAM International Conference on Data Mining, pp. 586–594. Society for Industrial and Applied Mathematics (2015).
29. Lenta.ru Russian news dataset, https://github.com/yutkin/Lenta.Ru-News-Dataset, last accessed 2019/04/28.

A Comparative Study on Feature Selection in Relation Extraction from Electronic Health Records

Ilseyar Alimova[0000−0003−4528−6631] and Elena Tutubalina[0000−0001−7936−0284]

Kazan (Volga Region) Federal University, Kazan, Russia
{alimovailseyar,evtutubalina}@gmail.com

Abstract. In this paper, we focus on clinical relation extraction; namely, given a medical record with mentions of drugs and their attributes, we identify relations between these entities. We propose a machine learning model with a novel set of knowledge and context embedding features. We systematically investigate the impact of these features with popular distance and word-based features. Experiments are conducted on a benchmark dataset of clinical texts from the MADE 2018 shared task. We compare the developed feature-based model with BERT and several state-of-the-art models. The obtained results show that distance and word features are significantly beneficial to the classifier. The knowledge-based features increase classification results on particular types of relations only. The context embedding feature gives the highest increase in results among the other explored features. The classifier obtains state-of-the-art performance in clinical relation extraction with 92.6% of F-measure improving F-measure by 3.5% on the MADE corpus.

Keywords: relation extraction, electronic health records, natural language processing, machine learning, clinical data, hand-crafted features

1 Introduction

Electronic health records (EHRs) contain rich information that can be applied to different research purposes in the field of medicine such as adverse drug reaction (ADR) detection, revealing unknown disease correlations, design and execution of clinical trials for new drugs, clinical decision supports and evidence-based medicine [16, 1, 14, 9, 12, 2]. Despite the enormous potential contained in the clinical notes, there are a lot of technical challenges devoted to the extraction of necessary information from EHRs [16]. EHRs describing the treatment of patients represents a massive volume of an underused text data source. Natural language processing (NLP) can be a solution to provide fast, accurate, and automated information extraction methods that can yield high cost and logistical advantages.

The relation extraction, which identifies important links between entities is one of the crucial steps of natural language processing (NLP). In this paper, we consider the relation extraction task as a binary classification. The classifier takes as an input pre-annotated pairs of entities and have to identify the relation between them. Let us consider the sentence: "The patient has received *4 cycles* of *Ruxience* plus *Cyclophosphamide* in the last day". In this sentence the entities *Ruxience* and *4 cycles* are related to each other, while *Cyclophosphamide* and *4 cycles* are not related.

Considerable efforts have been devoted to relation extraction research in biomedical domain, including MADE shared-task challenge [15], i2b2 competition [31] and BioCreative V chemical-disease relation extraction task [33]. The aim of the MADE competition was unlocking ADR related information, which can be further used by pharmacovigilance and drug safety surveillance. The organizers provided EHRs texts annotated with medications and their relations to corresponding attributes, indications, and adverse events. All participants of competition developed system based on the machine learning approaches [4, 7, 20, 34]. The winning system obtained 86.8%, while other participants achieved comparable results [15]. However, for the real-world application of extracting drug-related information, the results need to be further improved. Moreover, the contribution of different feature types has not been extensively investigated yet.

To fill this gap, we systematically evaluate four types of features on drug-related information extraction from EHRs: distance, word-based, knowledge, and embedding. In addition to popular features, we propose novel features: (i) number of sentences and punctuation characters between entities, (ii) the previous co-occurrence of entities in biomedical documents from different sources, (iii) semantic types from Medical Subject Headings (MeSH), and (iv) context embedding feature obtained with sent2vec model [6]. We apply a random forest model and perform experiments on the MADE corpus. For comparison, we evaluate a classifier based on Bidirectional Encoder Representations from Transformers (BERT) and approaches of teams participated in the MADE shared task.

The classifier with a combination of baseline and context embedding feature obtains the best results of 92.6% of F-measure and outperforms the previous state-of-the-art results [22] on 3.5%. BERT achieves 90.5% of F-measure. The obtained results show that distance and word features are significantly beneficial to the machine learning classifier. The knowledge features can increase results only on particular types of relations. We also found out that the context embedding feature gives the highest increase in results among the other explored features.

The rest of the paper is structured as follows. We discuss related work in Section 2. Section 3 devoted to corpus description. We describe our set of features in Section 4. Section 5 provides experimental evaluation and discussion. Section 6 concludes this paper.

2 Related Work

The first attempts to relation extraction from EHRs were made in 2010. One of the challenges of i2b2 competition was devoted to assigning relation types that hold between medical problems, tests, and treatments in clinical health records

[31]. This challenge aimed to classify relations of pairs of given reference standard concepts from a sentence. The system based on maximum entropy with a set of features from [25], semantic features from Medline abstracts and parsing trees feature performed the best results among challenge participants [27]. The described system obtained 73.7% of F-measure. The model developed by the team from NRC Canada achieved 73.1% of F-measure [3]. This model is also based on the maximum entropy classification algorithm with the following set of features: based on parsing trees, Pointwise Mutual Information between two entities calculated on Medline abstracts, word surface, concept mapping and context, section, sentence, document-level features. Besides, category balancing and semi-supervised training were applied. The third-place system is based on a hybrid approach that combines machine-learning techniques and constructed linguistic patterns matching [13]. The authors trained SVM with three types of features: surface, lexical, and syntactic. The system obtained 70.9% of F-measure. The rest of the participants applied supervised machine-learning approaches and achieved the results varying from 70.2% to 65.6% of F-measure [24, 17, 11, 28, 8]. One of the main problems faced by participants was varying number of examples for each relation types. The developed classifiers could capture the larger classes accurately by using basic textual features. However, to recognize less relevant relation types, hand-built rules have to be developed.

Natural Language Processing Challenge for Extracting Medication, Indication, and Adverse Drug Events from Electronic Health Record Notes was organized in 2018 [15]. The aim of the competition was extracting ADRs and detecting relations between drugs, their attributes, and diseases. In contrast to i2b2 competition, in this case, only entities are annotated in the corpus. Thus, it is necessary to make candidate pairs and then determine if there is a relation between them. The first place obtained system based on a random forest model with following a set of features, including, candidate entity types and forms, the number of entities between and their types, tokens and part of speech tags between and neighboring the candidate entities [4]. According to the competition resulting table, the described system obtained 86.8% of micro-averaged F_1. Dandala et al. applied the combination of Bidirectional LSTM and attention network and achieved the second place results with 84% of micro-averaged F_1 score [7]. The third place was taken by the system based on the support vector machine model [34]. The classifiers use four types of features: position, distance, a bag of words, and a bag of entities and obtained 83.1% of micro-averaged F_1 measure. Magge et al. employed random forest with entity types, number of the word in entities, number of words between entities, averaged word embeddings of each entity and indicator of presence in the same sentence as a feature [20]. This approach obtained 81.6% of micro-averaged F_1. As can be seen, the most participant teams applied machine learning models, and the only one utilized neural networks while the results were on par.

Munkhdalai et al. conducted additional experiments on MADE corpus and explored three supervised machine learning systems for relation identification: (1) a support vector machines (SVM) model, (2) an end-to-end deep neural network system, and (3) a supervised descriptive rule induction baseline system [22]. For the SVM system entity types, a number of clinical entities, tokens between entities, n-grams between two entities and of surrounding tokens, character n-grams of named entities were applied as features. The combination of BiLSTM and attention was utilized as a neural network model. The maximum averaged F-measure of 89.1% was obtained by the SVM based approach, while the neural network achieved only 65.72% of F-measure.

According to the reviewed studies, the machine learning approaches have a high potential for clinical relation extraction task. However, for real-world biomedical applications, the results need to be improved [15]. The error analysis of systems shows that the most common errors: (i) related entities more than two sentences away from each other, (ii), not related entities occur together in a small distance marks as related (iii) there is more than one entity related to the same entity and only the closest relation is detected. We suppose that these errors can be eliminated if the context is taken into account. Also, most of the previously proposed studies devoted to relation extraction from EHRs largely ignore valuable supportive information, such as the context and knowledge sources. Therefore, the machine learning approach proposed in this paper can be viewed as an extension of the previous work on extracting relations from clinical notes.

3 Corpus

We evaluated our model on the MADE competition corpus [15]. MADE corpus consists of de-identified electronic health records (EHRs) from 21 cancer patients. The EHRs include discharge summaries, consultation reports, and other clinic notes. The overall number of records is 1089, where 876 records were selected for the training split, and the remaining 213 notes formed the testing split. Several annotators participated in the annotation process, including physicians, biologists, linguists, and biomedical database curators. Each document was annotated with two annotators, one of which carried out the initial annotation, the second reviewed the annotations and modified them to produce the final version.

Each record annotated with the following types of entities: drug, adverse drug reaction (ADR), indication, dose, frequency, duration, route, severity, and SSLIF (other signs/symptoms/illnesses). There are 7 types of relations: drug−ade (adverse), sslif−severity (severity), drug−route (route), drug−dosage (do), drug−duration (du), drug−frequency (fr), drug−indication (reason). The detailed statistic of annotated relations is presented in Table 1. According to statistics, the most common relation types are drug-dose, drug-indication, and frequency. Two types of relationships (reason and adverse) have the maximum distance between entities more than 900 characters, which complicates the identification of relations between them.

Table 1. The overall statistic for MADE corpus

	Number			**Avg. distance**			**Max. dist**		
Relation type	**train**	**test**	**all**	**train**	**test**	**all**	**train**	**test**	**all**
do	5176	866	6042	8.4	7.7	8.3	215	143	215
reason	4523	870	5393	89.3	63.8	85.2	981	868	981
fr	4417	729	5146	17.7	18.6	17.8	201	178	201
severity_type	3475	557	4032	2.6	1.8	2.5	259	188	259
adverse	1989	481	2470	59.4	45.6	56.7	937	718	937
manner/route	2550	455	3005	13.5	12.9	13.4	191	137	191
du	906	147	1053	18.5	15.0	18.0	272	121	272
all	23036	4109	27145	30.6	26.0	29.9	981	868	981

4 Features

We have divided features into four categories: distance, word, embedding, and knowledge. Distance features are based on counting different metrics between entities. Word features were derived using various properties of context and entity words. Embedding features were received from word embedding models pre-trained on a large number of biomedical texts. Knowledge features were obtained from biomedical resources. The description of each type of feature set out below.

1. Distance features:
   - *word distance* (word_dist): the number of words between entities;
   - *char distance* (char_dist): the number of characters between entities;
   - *sentence distance* (sent_dist): the number of sentences between entities;
   - *punctuation* (punc_dist): the number of punctuation characters between entities;
   - *position* (position): the position of the entity candidate (drug or SSLIF type entity) with respect to the attribute among the entire entity candidates of the attribute, where the position of medical attribute is set to 0.
2. Word features:
   - *bag of words* (bow): all words within a 10-word window before and after the entities plus the entities text. We utilized as features only words that appeared in such context windows with frequencies $\geq$500 across the dataset. Thus, for each entity pair we generated 847 features;
   - *bag of entities* (boe): the counts of all annotation types between the entities;
   - *entity types* (type): binary vector with the number of entities length and units at positions of entity types.
3. Embedding features:
   - *entities embeddings* (ent_emb): the vectors obtained from pre-trained word embedding models for each entity. We explored two word embedding models, including trained on the concatenation of Wikipedia and

PubMed, PMC abstracts [21], and BioWordVec created using PubMed and the clinical notes from MIMIC-III Clinical Database [6]. For entities represented by several words the averaged vector value was applied;
   - *context embedding* (cont_emb): vector obtained from pre-trained BioSentVec model for words between two entities [6]. BioSentVec was obtained using sent2vec library and consists of 700-dimensional sentence embeddings;
   - *similarity* (sim): similarity measure between entities embedding vectors. Four types of similarity measures were employed: taxicab, Euclidean, cosine, coordinate. The vectors were obtained from BioWordVec model [6].

4. Knowledge features:
   - *UMLS concept types* (umls): UMLS[1] (Unified Medical Language System) semantic types of entities represented with binary vector;
   - *MeSH concept types* (mesh): MeSH[2] (Medical Subject Headings) categories of entities represented with a binary vector;
   - *fda clinical trials occurrence* (fda): the number of co-occurrence of both entities in approval document received from FDA[3] for each drug of dataset;
   - *biomedical texts co-occurrence* (bio_texts): the number of entities co-occurrence in biomedical texts. The detailed description of this feature is provided below.

Prior knowledge retrieved from available sources is essential for today's health specialists to keep up with and incorporate new health information into their practices [23]. This process of retrieving relevant information is usually carried out by querying and checking medical articles. We propose a set of features based on primary sources of information to analyze the influence of this process on clinical decision making. In particular, we utilize statistics from various resources using *Pharmacognitive*[4]. This system provides access to databases of grants, publications, patents, clinical trials, and others.

For our experiments, we focus on three sources: (i) scientific abstracts from MEDLINE, (ii) USPTO patents, and (iii) projects from the grant-making Agencies of USA, Canada, EU, and Australia. The *Pharmacognitive* system allows retrieving statistics such as the number of documents or overall funding per year matching a query. The queries are generated using terms from entities of three types: Medication, Indication, and ADR. We extend all queries with terms' synonyms provided by the Pharmacognitive tools. We consider the following features for a individual query *Medication*, *Condition*, *ADR*:

- the number of publications/patents/projects published in the particular year (3 features for each year from 1952 to 2018);

[1] https://www.nlm.nih.gov/research/umls/
[2] https://www.nlm.nih.gov/mesh/meshhome.html
[3] https://www.fda.gov/
[4] https://pharmacognitive.com

- the number of publications/patents/projects published before the particular year (3 features for each year from 1953 to 2018);
- the total number of publications/patents/projects published for all time (3 features);
- the average and sum of projects' funding published in the particular year (2 features for each year from 1974 to 2018);
- the average and sum of projects' funding published before the particular year (3 features for each year from 1975 to 2018);
- the average and sum of projects' funding published for all time (2 features).

We also generate features based on statistics of publications and projects for joint queries of two terms: *Drug* and a disease-related entity (*ADR* or *Indication*).

5 Experiments

In this section, we describe our classifier model, entity pair generation, experiments, and results.

5.1 Classifier

We build a system to resolve the task as a set of independent Random Forest classifiers, one for each relation type. The Random Forest model was implemented with the Scikit-learn library [26]. We tuned the parameters on 5-fold cross-validation and set the number of estimators equal to 100 and the weight balance: 0.7 for positive and 0.3 for negative classes to mitigate the imbalanced class issues.

5.2 Bidirectional Encoder Representations from Transformers (BERT)

BERT (Bidirectional Encoder Representations from Transformers) is a recent neural network model for NLP presented by Google [10]. The model obtained state-of-the-art results in various NLP tasks, including question answering, dialog systems, text classification and sentiment analysis [18, 35, 30, 5, 29]. BERT neural network based on bidirectional attention-based transformer architecture [32]. One of the main model advantages is the ability to give it a row text as the input. In our experiments, we utilized the entity texts combined with a context between them as an input.

5.3 Entities Pair Generation

For each entity we obtained a set of candidate entities following the rules from [34]: the number of characters between the entities is smaller than 1000, and the number of other entities that may participate in relations and locate between the candidate entities is not more than 3. These restrictions allow to reduce infrequent negative pairs and mitigate the imbalanced class issues, while more than 97% of the positive pairs remain in the dataset.

Table 2. The results of F-measure for each relation type and averaged micro F-measure of all relation types for MADE corpus. The distance and word features are applied as a baseline.

Features	severity	route	reason	do	du	fr	adverse	all
baseline: distance & word feat-s	.933	.918	.806	.906	.905	.896	.729	.866
Munkhdalai et al. [22]	.950	.960	.750	.880	.910	**.950**	.850	.891
Li et al. [19]	-	-	-	-	-	-	-	.872
baseline-word_dist	.923	.922	.812	.900	.860	.909	.716	.864
baseline-char_dist	.929	.916	.810	.908	.869	.890	.731	.864
baseline-sent_dist	.933	.919	.807	.910	.880	.906	.719	.866
baseline-punc_dist	.926	.912	.798	.907	.836	.906	.735	.863
baseline-position	.931	.917	.803	.897	.865	.883	.723	.858
distance	.918	.843	.683	.859	.713	.780	.525	.766
baseline-boe	.932	.897	.775	.888	.861	.868	.715	.845
baseline-bow	.918	.906	.726	.895	.810	.843	.712	.828
baseline-type	.934	.906	.779	.899	.891	.891	.562	.839
word	.542	.777	.645	.662	.718	.846	.511	.672
baseline+emb_pubmed_pmc_wiki	.927	.898	.730	.887	.684	.900	.605	.827
baseline+emb_bio	.920	.903	.772	.893	.602	.908	.613	.833
baseline+cont_emb	.936	.954	**.937**	.929	.854	.938	**.869**	**.926**
cont_emb	.932	.935	.909	.915	.854	.835	.782	.884
baseline+sim	.920	.908	.796	.905	.880	.902	.737	.862
baseline+umls	.936	.915	.815	.922	.883	.891	.734	.870
baseline+mesh	.938	.918	.812	.910	.856	.904	.730	.868
baseline+fda	.936	.912	.808	.906	.895	.909	.730	.868
baseline+bio_text	.934	.918	.805	.906	.905	.896	.749	.866
baseline+knowledge	.936	.914	.806	.916	.889	.896	.736	.848
BERT	**.951**	**.976**	.845	**.934**	**.946**	**.950**	.767	.905

5.4 Experiments and Results

We utilize the model with distance and word features as a baseline. In addition, we compare our results with two state-of-the-art approaches: proposed by Munkhdalai et al. [22] and by Li et al. [19]. Munkhdalai et al. applied SVM with following features: (i) token distance between the 2 entities, (ii) number of clinical entities between the 2 entities, (iii) n-grams between the 2 entities, (iv) n-grams of surrounding tokens of the 2 entities, (v) one-hot encoding of the left and right entities types, (vi) character n-grams of the named entities. Li et al. utilized modern capsule networks.

For distance and word features evaluation, we removed each of the features individually and in combination. To determine the most significant features from embedding and knowledge features sets, we add each of the features separately to the baseline model. The F_1-measure for each relation type and micro-averaged over all classes F_1 were used as evaluation metrics. The evaluation scripts provided by competition organizers were applied to compute these values. The results for each relation type and micro-averaged F-measure are shown in Table 2.

The combination of baseline selected features achieved 86.8% of micro F-measure. This result stays in pair with the best 86.84% F-measure achieved in the competition. The combination of baseline and context embedding features obtained the best results of 92.6% of micro-averaged F-measure. Thus our model outperformed the Munkhdalai et al. results on 3.5%, Li et al. approach on 5.4% and baseline approach on 6%. All reported improvements of the baseline model with context embedding feature over baseline and both state-of-the-art approaches are statistically significant with p-value < 0.01 based on the paired sample t-test. Further, we provided a more detailed analysis of the presented results.

According to Table 2, the classifier with distance features achieves 76.6% of micro-averaged F-measure. Different types of distance features seemed to be complementary to each other due to the absence of one of them leads to approximately the same loss of results. The baseline model without distance set of feature (see row ‘word’ in Table 2) decrease results on 19% of micro F-measure, which evidences the importance of these parameters for relation classification.

The word-based features also improved the performance of the relation extraction system. The most significant improvement of micro F-measure obtained with a bag of words feature (+3.8 %), which can be explained by a larger vector size compared to the rest of the word-based features. The entity type and a bag of entities feature increased the results of the baseline on 2.7% and 2.1% respectively (see rows ‘baseline-type’ and ‘baseline-boe’).

The results for embedding features show that entity embeddings and similarity feature decrease the results regardless of a word embedding model used. The context embedding feature achieved the most considerable improvement of baseline results and obtained 92.6% of micro F-measure. Moreover, the model trained only with the sent2vec feature, outperformed the baseline by 1.8%. This result leads to the conclusion that the context between candidate entities contains more useful information to make a conclusion about relations than candidate entities.

To evaluate knowledge features, it is better to consider the results for different relation types separately. The supplement of UMLS based feature to baseline model increased the results of baseline for severity, reason, and adverse relation types on 0.3%, 0.9% and 0.5% of F-measure respectively. The model with a combination of baseline and MeSH semantic types feature increased the results of baseline for severity and reason types on 0.5% and 0.6% of F-measure, respectively. The FDA co-occurrence feature increased the results for frequency type on 1.3%, while for the rest of the types results are in par. The number of co-occurrence in the biomedical texts feature improved the classifier performance for adverse relation type on 2%. Thus, the knowledge features improved model results for selected types of features.

BERT model achieved the best results for the severity, route, dose, duration, and frequency types of relation. However, for a reason and adverse types, this model obtained F-measure approximately lower on 10% than a random forest with baseline and context embedding features. Thus, BERT gained 90.5% of micro F-measure, and this is the second result among all evaluated models. We suppose that the results reducing for adverse and reason types can be caused for two reasons: (i) the same disease in different cases could be an adverse drug reaction and a reason, (ii) the average length of the context for these relation types is too long to catch the relation between entities.

A comparison of results for different types of relation shows that the best result was achieved for route (97.6%). This result roughly stays on par with the best results for severity, reason, dose, duration, and frequency types, while the best results for adverse type lower on 10.7%. This difference in results could be due to the greater lexicon variety of adverse drug reaction entity type.

To sum up this section, three important conclusions can be drawn. First, the distance and word-based features are beneficial for the relation classifier. Second, the context embedding has more impact on entities relations than entities embeddings. Finally, the prior knowledge improves the results on particular relation types and the most improvement achieved on adverse relation type with biomedical text co-occurrence feature.

6 Conclusion

In this study, we have investigated the different types of features for drug-related information extraction tasks from EHRs. Our evaluation on MADE competition corpus shows that context embedding, distance, and word features bring the most beneficial to relation extraction task. The classifier with a combination of these sets of features outperformed state-of-the-art results. These facts lead to the conclusion that the context between entities plays a crucial role in relation detection. The detailed analysis of results showed that prior knowledge about entities co-occurrence improved the results for adverse relation type. Our future research will focus on the investigation of modern neural networks for relation extraction from EHRs. We also plan to analyze various context representation methods and extend experiments on other biomedical corpora.

Acknowledgments

This research was supported by the Russian Foundation for Basic Research grant no. 19-07-01115.

References

1. Bates, D.W., Cullen, D.J., Laird, N., Petersen, L.A., Small, S.D., Servi, D., Laffel, G., Sweitzer, B.J., Shea, B.F., Hallisey, R., et al.: Incidence of adverse drug events and potential adverse drug events: implications for prevention. Jama **274** (1), 29–34 (1995).
2. Batin, M., Turchin, A., Sergey, M., Zhila, A., Denkenberger, D.: Artificial intelligence in life extension: From deep learning to superintelligence. Informatica 41(4) (2017)
3. de Bruijn, B., Cherry, C., Kiritchenko, S., Martin, J., and Zhu, X.: Nrc at i2b2: one challenge, three practical tasks, nine statistical systems, hundreds of clinical

records, millions of useful features. In: Proceedings of the 2010 i2b2/VA Workshop on Challenges in Natural Language Processing for Clinical Data. Boston, MA, USA: i2b2 (2010).

4. Chapman, A.B., Peterson, K.S., Alba, P.R., DuVall, S.L., and Patterson, O.V.: Detecting adverse drug events with rapidly trained classification models. Drug safety, 1–10 (2019).
5. Chen, Q., Zhuo, Z., and Wang, W.: Bert for joint intent classification and slot filling. arXiv preprint arXiv:1902.10909 (2019).
6. Chen, Q., Peng, Y., and Lu, Z.: Biosentvec: creating sentence embeddings for biomedical texts. arXiv preprint arXiv:1810.09302 (2018).
7. Dandala, B., Joopudi, V., and Devarakonda, M.: Adverse drug events detection in clinical notes by jointly modeling entities and relations using neural networks. Drug safety, 1–12 (2019).
8. Demner-Fushman, D., Apostolova, E., Islamaj Dogan, R., et al.: Nlms system description for the fourth i2b2/va challenge. In: Proceedings of the 2010 i2b2/VA Workshop on Challenges in Natural Language Processing for Clinical Data. Boston, MA, USA: i2b2 (2010).
9. Demner-Fushman, D., Chapman, W.W., and McDonald, C.J.: What can natural language processing do for clinical decision support? Journal of Biomedical Informatics **42** (5), 760–772 (2009).
10. Devlin, J., Chang, M.W., Lee, K., and Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018).
11. Divita, G., Treitler, O., Kim, Y., et al.: Salt lake city vas challenge submissions. In: Proceedings of the 2010 i2b2/VA Workshop on Challenges in Natural Language Processing for Clinical Data. Boston, MA, USA: i2b2 (2010).
12. Frankovich, J., Longhurst, C.A., Sutherland, S.M.: Evidence-based medicine in the emr era. N Engl J Med 365(19), 1758–1759 (2011)
13. Grouin, C., Abacha, A.B., Bernhard, D., Cartoni, B., Deleger, L., Grau, B., Ligozat, A.L., Minard, A.L., Rosset, S., and Zweigenbaum, P.: Caramba: concept, assertion, and relation annotation using machine-learning based approaches. In: i2b2 Medication Extraction Challenge Workshop (2010).
14. Gurwitz, J.H., Field, T.S., Harrold, L.R., Rothschild, J., Debellis, K., Seger, A.C., Cadoret, C., Fish, L.S., Garber, L., Kelleher, M., et al.: Incidence and preventability of adverse drug events among older persons in the ambulatory setting. Jama **289** (9), 1107–1116 (2003).
15. Jagannatha, A., Liu, F., Liu, W., and Yu, H.: Overview of the first natural language processing challenge for extracting medication, indication, and adverse drug events from electronic health record notes (made 1.0). Drug safety, 1–13 (2018).
16. Jensen, P.B., Jensen, L.J., and Brunak, S.: Mining electronic health records: towards better research applications and clinical care. Nature Reviews Genetics **13** (6), 395 (2012).
17. Jonnalagadda, S. and Gonzalez, G.: Can distributional statistics aid clinical concept extraction. In: Proceedings of the 2010 i2b2/VA workshop on challenges in natural language processing for clinical data. Boston, MA, USA: i2b2 (2010).
18. Le, H., Hoi, S., Sahoo, D., and Chen, N.: End-to-end multimodal dialog systems with hierarchical multimodal attention on video features. In: DSTC7 at AAAI2019 Workshop (2019).
19. Li, F. and Yu, H.: An investigation of single-domain and multidomain medication and adverse drug event relation extraction from electronic health record notes

using advanced deep learning models. Journal of the American Medical Informatics Association **26** (7), 646–654 (2019).

20. Magge, A., Scotch, M., and Gonzalez-Hernandez, G.: Clinical ner and relation extraction using bi-char-lstms and random forest classifiers. In: International Workshop on Medication and Adverse Drug Event Detection, 25–30 (2018).
21. Moen, S. and Ananiadou, T.S.S.: Distributional semantics resources for biomedical text processing. Proceedings of LBM, 39–44 (2013).
22. Munkhdalai, T., Liu, F., and Yu, H.: Clinical relation extraction toward drug safety surveillance using electronic health record narratives: classical learning versus deep learning. JMIR public health and surveillance **4** (2), (2018).
23. Pao, M.L., Grefsheim, S.F., Barclay, M.L., Woolliscroft, J.O., McQuillan, M., and Shipman, B.L.: Factors affecting students' use of medline. Computers and Biomedical Research **26** (6), 541–555 (1993).
24. Patrick, J., Nguyen, D., Wang, Y., and Li, M.: I2b2 challenges in clinical natural language processing 2010. In: Proceedings of the 2010 i2b2/VA Workshop on Challenges in Natural Language Processing for Clinical Data. Boston, MA, USA: i2b2 (2010).
25. Patrick, J. and Li, M.: A cascade approach to extracting medication events. In: Proceedings of the Australasian Language Technology Association Workshop 2009, 99–103 (2009).
26. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011).
27. Roberts, K., Rink, B., and Harabagiu, S.: Extraction of medical concepts, assertions, and relations from discharge summaries for the fourth i2b2/va shared task. In: Proceedings of the 2010 i2b2/VA Workshop on Challenges in Natural Language Processing for Clinical Data. Boston, MA, USA: i2b2 (2010).
28. Solt, I., Szidarovszky, F.P., and Tikk, D.: Concept, assertion and relation extraction at the 2010 i2b2 relation extraction challenge using parsing information and dictionaries. Proc. of i2b2/VA Shared-Task. Washington, DC (2010).
29. Sun, C., Huang, L., and Qiu, X.: Utilizing bert for aspect-based sentiment analysis via constructing auxiliary sentence. arXiv preprint arXiv:1903.09588 (2019).
30. Uglow, H., Zlocha, M., and Zmyślony, S.: Semeval 2019 task 6: An exploration of state-of-the-art methods for offensive language detection. arXiv preprint arXiv:1903.07445 (2019).
31. Uzuner, Ö., South, B.R., Shen, S., and DuVall, S.L.: 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. Journal of the American Medical Informatics Association **18** (5), 552–556 (2011).
32. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., and Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems, 5998–6008 (2017).
33. Wei, C.H., Peng, Y., Leaman, R., Davis, A.P., Mattingly, C.J., Li, J., Wiegers, T.C., and Lu, Z.: Assessing the state of the art in biomedical relation extraction: overview of the biocreative v chemical-disease relation (cdr) task. Database **2016** (2016).
34. Xu, D., Yadav, V., and Bethard, S.: Uarizona at the made1. 0 nlp challenge. Proceedings of machine learning research **90**, 57 (2018).
35. Zhu, C., Zeng, M., and Huang, X.: Sdnet: Contextualized attention-based deep network for conversational question answering. arXiv preprint arXiv:1812.03593 (2018).

A New Approach for Approximately Mining Frequent Itemsets

Timur Valiullin, Joshua Zhexue Huang, Jianfei Yin, and Dingming Wu

Big Data Institute, College of Computer Science and Software Engineering
Shenzhen University Shenzhen, China

Abstract. Mining frequent itemsets in transaction databases is an important task in many applications. This task becomes challenging when dealing with a very large transaction database because traditional algorithms are not scalable due to the memory limit. In this paper, we propose a new approach for approximately mining of frequent itemsets in a transaction database. First, we partition the set of transactions in the database into disjoint subsets and make the distribution of frequent itemsets in each subset similar to that of the entire database. Then, we randomly select a set of subsets and independently mine the frequent itemsets in each of them. After that, each frequent itemset discovered from these subsets is voted and the one appearing in the majority subsets is determined as a frequent itemset, called a popular frequent itemset. All popular frequent itemsets are compared with the frequent itemsets discovered directly from the entire database using the same frequency threshold. The recalls and precisions of the frequent itemsets from selected subsets are analyzed against the entire database. The experiment results demonstrate that the use of less than 10 percent of the transaction data in the database can achieve more than 87 percent accuracy. The new approach is very suitable for parallel implementation for large transaction database mining.

Keywords: Approximate Frequent Itemsets Mining, Random Sample, Partition

1. Introduction

Frequent itemsets mining is the first and most critical stage of finding association rules from a transaction database. Association rule mining is one of the main data mining tasks in many applications, such as basket analysis, product recommendation, cross-selling, inventory control, etc. Huge research efforts are devoted to solving frequent itemsets mining problem. Many of these works had considerable impact and led to a plenty of sophisticated and efficient algorithms for association rules mining, such as Apriori [1, 2], FP-Growth (Frequent Pattern Growth) [3–6], Eclat [7–9] and some others. However, the decade fast development of e-commerce, online and off-line shopping has resulted in fast growth of transaction data, which present a tremendous challenge to these existing algorithms, because these algorithms require a large memory to run efficiently on large transaction databases.

Parallel and distributed association rule mining algorithms were developed to handle large transaction databases. Parallel association rule mining algorithms use in-memory computing to efficiently mine association rules from a large transaction database. However, their scalability is limited by the size of memory of the parallel system. Distributed association rule mining algorithms [10, 11] were developed using MapReduce [12] and run on a Hadoop cluster platform. The algorithms have better scalability, but they are not efficient in mining large transaction datasets because of frequent I/O operations and communication overhead between nodes.

In this paper, we propose a new approach for mining frequent itemsets from a big transaction dataset. Similar to the distributed algorithms in MapReduce, we partition the dataset into disjoint subsets of the same size. However, we make the distribution of frequent itemsets in each subset similar to the distribution of frequent itemsets in the entire dataset. Then, we randomly select a set of subsets and run a frequent itemset mining algorithm independently to find the local frequent itemsets from each subset. After all frequent itemsets are discovered from the set of subsets, each frequent itemset is voted by all subsets and the one appearing in the majority subsets is determined as a frequent itemset, called a popular frequent itemset. All popular frequent itemsets are compared with the frequent itemsets discovered directly from the entire database using the same frequency threshold. The recalls and precisions of the popular frequent itemsets from the selected subsets are analyzed against the entire database to show how many true frequent itemsets in the entire transaction dataset can be discovered from the selected subsets.

We have conducted experiments to evaluate the new approach on two datasets. Empirically we have shown that the proposed method is not only capable of producing highly accurate frequent itemsets but also approximating the global frequency of frequent itemsets with very small error.

The remaining of this paper is organized as follows. Related works are discussed in Section 2. Section 3 describes the proposed approach. In Section 4, the details of the algorithm are described. Experiments evaluation is presented in Section 5. Finally, conclusions and future work are drawn in Section 6.

2 Related Work

Frequent itemsets mining is a well-studied problem in computer science. However, the enormous data growth made traditional methods inadequate. Therefore, parallel and distributed algorithms came in use.

Researchers in [13] introduced the parallel implementation of the FP-growth algorithm on GPU. In [10] and [11], the authors introduced two different approaches for mining frequent itemsets in a large database based on MapReduce. In [10], researchers presented two methods for frequent itemsets mining based on Eclat algorithm. The first one is a distributed version of Eclat that partitions the search space more evenly among different processing units, and the second one is a hybrid approach, where k-length frequent itemsets are mined by an Apriori variant, and then the found frequent itemsets are distributed to the mappers where frequent itemsets are mined using Eclat. Authors of [11] presented a novel zone-wise approach for frequent itemsets mining based on sending computations to a multi-node cluster. All mentioned approaches have obtained a speed increase over the traditional algorithms and allowed to increase the size of the dataset used for mining. However, all introduced approaches require using the entire dataset to get the result. In [14], M. Riondato first introduced PARMA (Parallel Randomized Algorithm for Approximate association rule mining). Algorithm sends random subsets of the database to various machines in the cluster as an input. Then, each machine mines the received subset, and reducers combine the result. Research in [15] is the basis for the current work. Random sample partition (RSP) data model was presented, which showed that the block-level samples from an RSP data model can be efficiently used for data analysis.

3 A New Approach

In our approach, we split a big dataset into smaller disjoint subsets such that the distribution of frequent itemsets in each subset is similar to the distribution of frequent itemsets in the entire dataset. Mining smaller subsets allows using traditional frequent itemset mining algorithms without experiencing memory limit problems. By combining the results of random subsets, we are able to produce highly accurate approximate frequent itemsets.

3.1 Definitions

A transactional dataset $D=\{t_1,t_2,...,t_n\}$ is represented by a collection of n transactions, where each transaction t is a subset of the set of items $I=\{I_1,I_2,...,I_m\}$. An itemset A with k distinct items is referred as *k-itemset*. In this paper, we do not distinguish itemsets with different numbers of unique items. Given an itemset A, define $T_D(A)$ as the set of transactions in D which contain A. The number of transactions in $T_D(A)$ is defined as the support of A by D and denoted as $support(A)=|T_D(A)|$. The frequency of A, i.e., proportion of transactions containing A in D, is denoted as

$$freq_D(A) = \frac{|T_D(A)|}{n}.$$

Under the above definitions, the task for finding frequent itemsets from D with respect to a minimal frequency threshold θ is defined as follows.

Definition 1. *Given a minimum frequency threshold θ for $0<\theta\leq 1$, the frequent itemsets mining with respect to θ is finding all itemsets $\{A_i\}$ for $1\leq i\leq M$ with $freq(A_i)\geq\theta$, where M is the total number of frequent itemsets found in D. Formally, we define the whole set of frequent itemsets in D as*

$$FI(D,I,\theta)=\{(A_i,freq_D(A_i)): A_i \subset I, freq_D(A_i)\geq\theta\}.$$

Definition 2. *Let FI(D,I,θ) be the set of frequent itemsets in D with respect to θ and M=|FI(D,I,θ)| the number of frequent itemsets in FI. The accumulative distribution of frequent itemsets in FI is defined as*

$$P(f) = \frac{1}{M} \sum_{\forall A_i \in FI} I(freq_D(A_i) \leq f$$

where I() is an indicator function and f is a frequency value for 0≤f≤1. The example of P(f) is shown in Fig. 1.

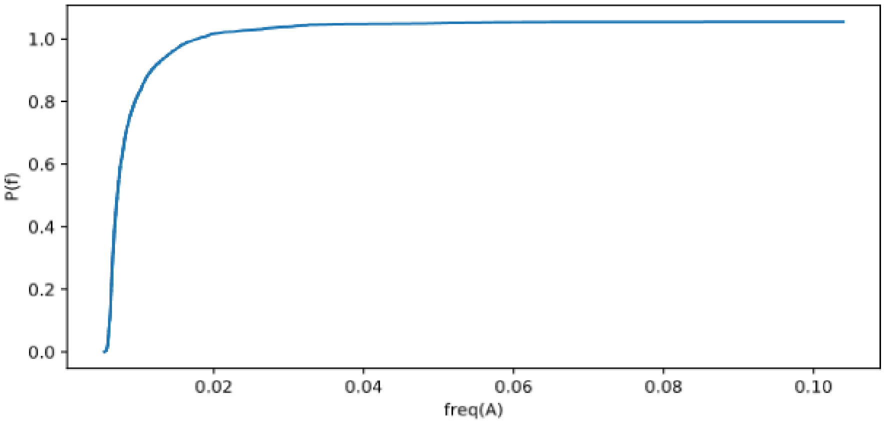


Fig. 1. Example of the accumulative frequent itemsets distribution

Let *D* be a big transactional dataset and *P*={*D*$_1$,*D*$_2$,...,*D*$_k$} a partition of *D*, where $\bigcup_{i=1}^{k} D_i = D$ and D_i $^{\mathrm{T}}$$D_j$ = ∅ for *i* 6=*j*. D_i for 1≤*i*≤*k* is named as a block of transactions of dataset *D*.

Definition 3. *Let P*$_D$*(f) be the accumulative distribution of frequent itemsets FI(D,I,θ) and P*$_{D_i}$*(f) the accumulative distribution of frequent itemsets FI(D*$_i$*,I,θ) for 1≤i≤k. P is a random sample partition of D if*

$$P_{D_i}(f) \rightarrow P_D(f) \text{ as } |D_i| \rightarrow |D|. \quad (1)$$

Definition 3 is a redefined definition of random sample partition in [15] with respect to frequent itemsets by replacing the condition of $E[\tilde{F}_k(t)]=F(t)$ with condition (1), where $\tilde{F}_k(t)$ denotes the sample distribution function of D_k and $E[\tilde{F}_k(t)]$ denotes its expectation.

Definition 4. *FI*$_D$*(D,I,θ) is called the set of global frequent itemsets and FI*$_{D_i}$*(D*$_i$*,I,θ) the set of local frequent itemsets. Accordingly, P*$_D$*(f) is the accumulative distribution of global frequent itemsets and P*$_{D_i}$*(f) is the accumulative distribution of the local frequent itemsets in D*$_i$*.*

3.2 Approximate Computing

When the transactional dataset D is big and cannot be held in memory, we cannot run a frequent itemset mining algorithm on D to find all frequent itemsets $FI_D(D,I,\theta)$. In this situation, we randomly select a set of l transaction blocks $\{D_1,D_2,...,D_l\}$ from the partition P and use the set of local frequent itemsets $FI_{Di}(D_i,I,\theta)$ for $1\leq i\leq l$ to estimate the set of global frequent itemsets $FI_D(D,I,\theta)$. This approach is called approximate frequent itemset mining.

Definition 5. *Let itemset A be a frequent itemset in $FI_{Di}(D_i,I,\theta)$ for $1\leq i\leq l$. A is called a popular frequent itemset if*

$$\sum_{i=1}^{l} I(A \in FI_{D_i}(D_i, I, \theta)) \geq a \tag{2}$$

where I() is an indicator function and a is a given integer greater than l/2.

Definition 6. *The frequency of a popular frequent itemset A is defined as*

$$freq(A) = \frac{\sum_{i=1}^{l} freq_{D_i}(A) \times I(A \in FI_{D_i}(D_i, I, \theta))}{\sum_{i=1}^{l} I(A \in FI_{D_i}(D_i, I, \theta))}$$

The set of all popular frequent itemsets PFI from $FI_{Di}(D_i,I,\theta)$ for $1\leq i\leq l$ is the estimation of the set of global frequent itemsets $FI_D(D,I,\theta)$. Given PFI and assuming $FI_D(D,I,\theta)$ is known, an itemset A has the following status:

- true positive if $A\in PFI$ and $A\in FI_D(D,I,\theta)$ – false positive if $A\in PFI$ but $A/\in FI_D(D,I,\theta)$;
- true negative if $A/\in PFI$ and $A/\in FI_D(D,I,\theta)$ – false negative if A /∈ PFI but $A\in FI_D(D,I,\theta)$.

4 An Approximate Frequent Itemsets Finding Algorithm

In this section, we propose an approximate algorithm for finding the set of popular frequent itemsets from a set of l transaction blocks $\{D_1,D_2,...,D_l\}$ randomly selected from the partition of a big transactional dataset D, and using the popular frequent itemsets to estimate the set of frequent itemsets in D with respect to a given frequency threshold θ. The algorithm consists of three steps: converting the dataset D into a partition of k transaction blocks and randomly selecting l blocks from the partition; finding the local frequent itemsets for each of l selected transaction blocks; finding the popular frequent itemsets from the local frequent itemsets.

4.1 Generate Partition of Transaction Blocks

Given a transaction dataset *D*, the first step is to convert it to a partition of transaction blocks. *D* is preprocessed such that each record represents one purchase transaction and the transactions with one purchased item are removed. The pseudo code for creating the random partition is given in Algorithm 1.

Algorithm 1 RSP Blocks generation and selection

```
Input:
-D: preprocessed data;
-l: number of subsets;
-m: subset size;
procedure RSPBlocks(D,l,m)
    $k = \frac{|D|}{m}$
    for each $D_i$, 1 <= i <= k do
        randomly assign m transactions from D to the i-th block without replacement
    end for
    randomly select l transaction blocks from the set of created k blocks, l <= k
Output: set of l transaction blocks of D
end procedure
```

4.2 Finding Local Frequent Itemsets

In this step, Apriori algorithm is called to find the local frequent itemsets from each of *l* transaction blocks $\{D_1,D_2,...,D_l\}$ with respect to a given minimum frequency threshold θ. Finally, *l* sets of local frequent itemsets are obtained. The pseudo code of obtaining local frequent itemsets is presented in Algorithm 2.

Algorithm 2 Local frequent itemsets mining

```
Input:
-{$D_l$}: set of l transaction blocks of D;
-$\theta$: minimum frequency threshold
procedure LocalFIs({$D_l$},$\theta$)
    for each $D_i$, 1 <= i <= l do
        $FI_i = Apriori(D_i,\theta)$
    end for
Output: {$FI_l$} - set of l sets of the local frequent itemsets
end procedure
```

4.3 Finding Popular Local Frequent Itemsets

The *l* sets of local frequent itemsets are united into one set of unique local frequent itemsets. For each frequent itemset in the united set, the number of its appearances in the *l* sets is checked with Eq. (2). If the condition is satisfied, the frequent itemset is considered as a popular frequent itemset. Otherwise, it is dropped. All local frequent itemsets in the united set are checked and the set of popular frequent itemsets is obtained. These popular frequent itemsets are used as the approximate set of the frequent itemsets in *D* with respect to the same minimum frequency threshold *θ*. The pseudo code is given in Algorithm 3.

Algorithm 3 Popular Frequent Itemset mining

```
Input:
-{$FI_l$}: set of $l$ local frequent itemsets;
procedure PopularFIs({$FI_l$})
$FI = dictionary(\cup_{i=1}^{l} FI_i)$ // for all frequent itemsets found, creating <key, value>pair, where itemset is a key and number of its repeats in all blocks is a value
    for each frequent itemset $\in FI$ do
        if value$>\frac{k}{2}$ then
            include frequent itemset to the set of popular frequent itemsets
        end if
    end for
Output: set of popular frequent itemsets
end procedure
```

5 Experiments

To demonstrate the performance of the approximate frequent itemsets algorithm, we conducted a series of experiments on two datasets. We run our algorithm several times with different numbers of transaction blocks and different block size, and compared the set of popular frequent itemsets with the exact set of frequent itemsets obtained from the entire dataset.

5.1 Datasets

We evaluated the proposed approach on 2 datasets downloaded from Kaggle.com and Open-Source Data Mining Library. Properties of datasets used in the experiments are described in Table 1.

Table 1. Properties of the datasets used in experiments

	Kaggle dataset	Online Retail dataset
Number of transactions	729148	541908
Number of items	791	2603
Average transaction length	8	4

5.2 Experiment Settings

In order to test the proposed algorithm, the set of popular frequent itemsets was compared against the set of frequent itemsets in the entire dataset to compute the accuracy, recall, and precision. We conducted 50 experiments for each set of parameters specified in Table 2 and averaged obtained results afterward. For both datasets, we used the same minimum frequency threshold for both local and global frequent itemsets. We chose threshold to be small enough to produce a big collection of the output frequent itemsets and set θ to be 0.005 for all experiments.

Testing was started with a comparison of the accumulative distribution of frequent itemsets for local and global frequent itemsets and proceeded with the evaluation of the different metrics of the popular frequent itemsets.

Table 2. Parameters used for experiments

Number of blocks	Block size
50	10000, 5000, 3500, 2000, 1000, 500
30	10000, 5000, 3500, 2000, 1000, 500
15	10000, 5000, 3500, 2000, 1000, 500
10	10000, 5000, 3500, 2000, 1000, 500
5	10000, 5000, 3500, 2000, 1000, 500

5.3 Evaluation Methods

For evaluation of the accuracy and sufficiency of obtained approximate frequent itemsets, we used the confusion matrix in our research. Using the confusion matrix allows analyzing the efficiency of the proposed approach more detailed by introducing three measures, namely recall, precision, and accuracy.

$$Recall = \frac{TP}{TP + FN}$$

shows the fraction of the global frequent itemsets that are contained in the popular FIs.

$$Precision = \frac{TP}{TP + FP}$$

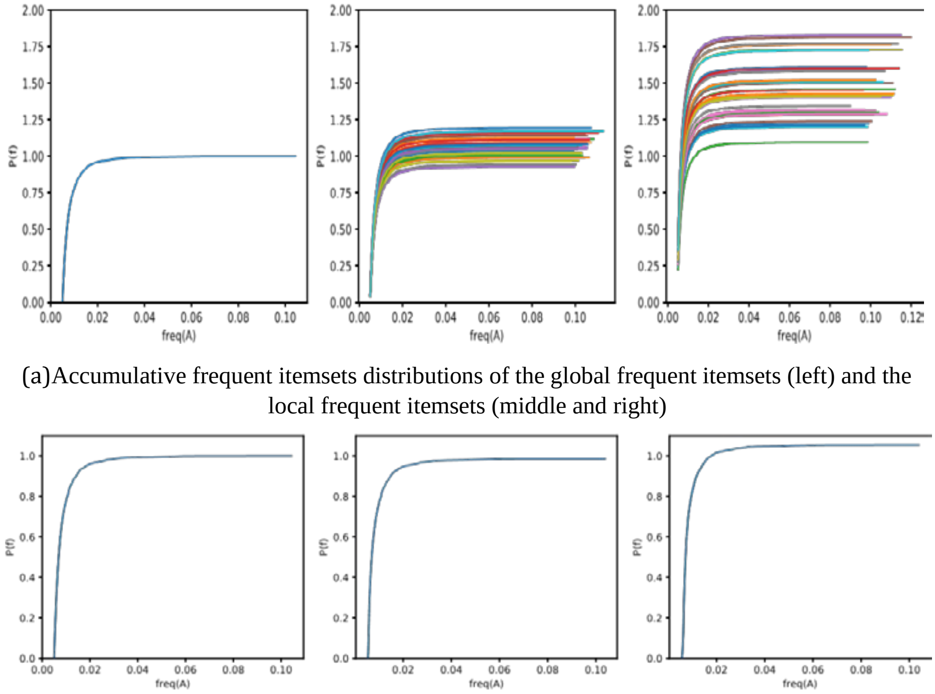
shows the fraction of the popular frequent itemsets that are contained in the set of the global frequent itemsets.

$$Accuracy = \frac{TP + TN}{TP + FN + TN + FP}$$

shows the proportion of accurate results among the total number of cases examined.

5.4 Experiment Results

We started with a comparison of accumulative distributions of the local and global FIs. Fig. 2(a) clearly shows that the accumulative distribution of global frequent itemsets is similar to the accumulative distributions of local frequent itemsets with a sufficiently large subset size (Fig. 2(a) left and middle graphs). Different colors show the differences of the accumulative distributions between the local frequent itemsets. However, decreasing the size of the subset leads to the growth of the number of local frequent itemsets (Fig. 2(a) right graph) and results in a significant difference between the accumulative distributions of the global and local frequent itemsets. Nevertheless, the accumulative distribution of the popular frequent itemsets shows almost identical accumulative distribution to the global frequent itemsets. The accumulative distributions of the global and popular frequent itemsets are represented in Fig. 2(b). It shows that the number of the popular FIs almost matches to the number of global FIs, and the overall frequency of the popular frequent itemsets is consistent with the frequency of the global frequent itemsets.



(a)Accumulative frequent itemsets distributions of the global frequent itemsets (left) and the local frequent itemsets (middle and right)

(b) Accumulative frequent itemsets distribution of the global frequent itemsets(left), accumulative frequent itemsets distributions of the popular frequent itemsets

Fig. 2. Accumulative frequent itemsets distributions. Number of subsets = 30, subset size (middle) = 10000, subset size (right) = 2000 (Kaggle dataset)

The proposed algorithm approximates the exact set of frequent itemsets in the entire dataset. The difference between approximate and global collections is false-positive FIs. The number of false-positive frequent itemsets affects one of the accuracy measures, namely precision. Fig. 3 shows, how the precision value is affected by different experimental parameters. It is observed that using more transaction blocks decreases the number of false-positive itemsets, therefore increasing the precision. From the graph, we can see that the number of falsepositive frequent itemsets decreases as the growth of the block size.

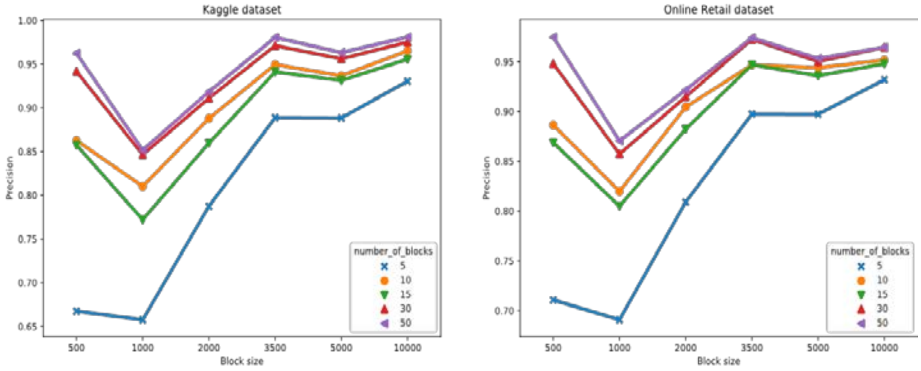


Fig. 3. Precision changes with different parameters

The overall change of the accuracy defined in terms of a confusion matrix is represented in Fig. 4. From the graphs, we can see that accuracy increase can be obtained by increasing both the number of subsets used and the subset size.

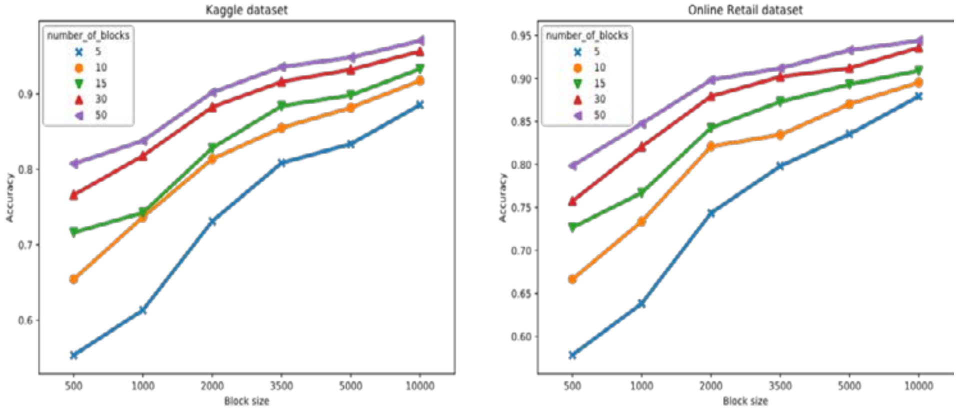


Fig. 4. Accuracy changes with different parameters

5.5 Result Analysis

To evaluate the efficiency of our approach, we conducted 50 independent experiments on both datasets with specified parameters in Table 2 to estimate the performance for each set of parameters. For each test run, the set of approximate frequent itemsets was compared to the exact set of the frequent itemsets obtained by mining the entire dataset. As a result, we received 50 different observations of elapsed time, recall, precision and accuracy for each set of parameters, and then averaged all values. Fig. 5 illustrates how the average accuracy changes with the change in the amount of data being mined. The graphs show that the proposed algorithm is capable of producing approximate frequent itemsets with above 87% of the accuracy, using only a little less than 10% of data.

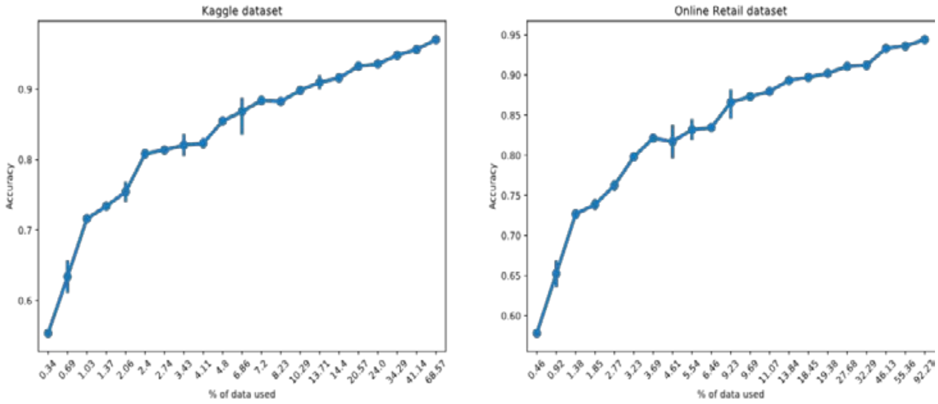


Fig. 5. Accuracy increases with the rise of data used to mine

We also conducted an evaluation of the estimated frequency error in the approximate frequent itemsets for all experimental parameters. In Fig. 6, we depict the distribution of the average absolute error in the frequency estimation, defined as:

$$\frac{\sum_{i=1}^{l}|freq_{D_i}(A) - freq(A)|}{|PFI|-|FI(D,I,\theta)|}$$

for all itemsets *A* that are contained in both the approximate and the global frequent itemsets. We can see that the flustration of the error decreases with the increase of the subset size. The error is reduced as the increase of the number of transaction blocks and the block size.

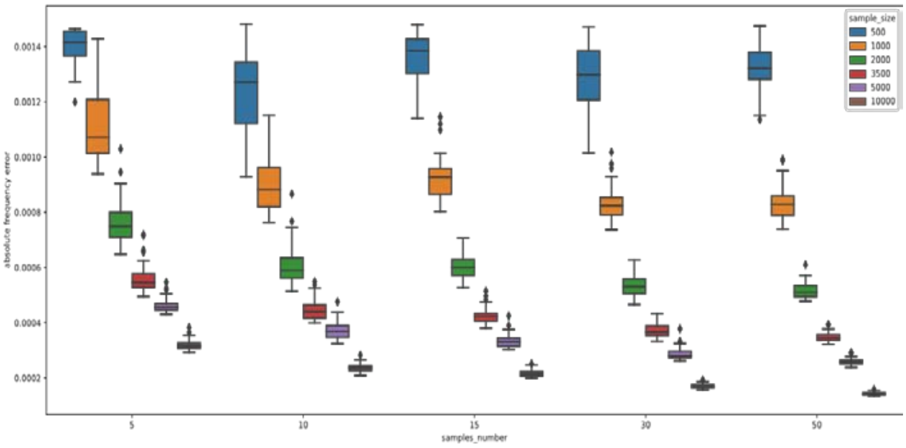


Fig. 6. Error in frequency estimations for different parameters (Kaggle dataset)

6 Conclusions and Future Work

In this paper, we have presented a new approach for mining approximate collections of frequent itemsets based on a random sample partition of the data. We have shown that using the RSP data model in big data can be very beneficial, especially in the frequent itemset mining task, since the size of transaction database grows much faster than the contained patterns change.

For the further work, we are going to implement the parallel version of the algorithm on a cluster and to conduct experiments on big datasets in terabyte scale. We will also conduct a theoretical analysis of the approach.

References

1. Agrawal, R., Imielinski, T., and Swami, A.: Mining association rules between sets of items in large databases. In Proceedings of SIGMOD, 1993.
2. Agrawal, R. and Srikant, R.: Fast algorithms for mining association rules in large data bases. In Proceedings of VLDB, 1994.
3. Han, J., Pei, J., and Yin, Y.: Mining frequent patterns without candidate generation. In Proceedings of the 19th ACM International Conference on Management of Data (SIGMOD), 2000.
4. Grahne, G. and Zhu, J.: Efficiently using prefix-trees in mining frequent itemsets. In Proceedings of the CEUR Workshop Proceedings, 2003.
5. Racz, B. An fp-growth variation without rebuilding the fp-tree. In Proceedings of the CEUR Workshop Proceedings, 2003.
6. Grahne, G. and Zhu, J.: Reducing the main memory consumptions of fpmax* and fpclose. In Proceedings of the CEUR Workshop Proceedings, 2004.
7. Zaki, M.J., Parthasarathy, S., Ogihara, M., and Li, W.: New algorithms for fast discovery of association rules. In Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining, 1997.

8. Zaki, M.J. and Gouda, K.: Fast vertical mining using diffsets. In Proceedings of the 9th ACM International Conference on Knowledge Discovery and Data Mining, 326–335, Washington, DC, USA, 2003.
9. Schmidt-Thieme, L.: Algorithmic features of eclat. In Proceedings of the Workshop Frequent Item Set Mining Implementations, 2004.
10. Moens, S., Aksehirli, E., and Goethals, B.: Frequent itemset mining for big data. In 2013 IEEE International Conference on Big Data, 2013.
11. Prajapati, D., Garg, S., and Chauhan, N.C.: Interesting association rule mining with consistent and inconsistent rule detection from big sales data in distributed environment. Future Computing and Informatics Journal **2** (1), 19–30 (2017).
12. Dean, J. and Ghemawat, S.: Mapreduce: Simplified data processing on large clusters. In Proceedings of the CACM, 107–113 (2004).
13. Jiang, H. and Meng, H.: A parallel fp-growth algorithm based on gpu. In 2017 IEEE 14th Int. Conf. E-bus. Eng, 97–102 (2017).
14. Riondato, M., DeBrabant, J.A., Fonseca, R., and Upfal, E.: Parma: a parallel randomized algorithm for approximate association rules mining in mapreduce. In Proceedings of the ACM International Conference on Information and Knowledge Management, 2012.
15. Salloum, S., He, Y., Huang, J.Z., Zhang, X., and Emara, T.Z.: A random sample partition data model for big data. In [Online]. Available: https://arxiv.org/abs/1712.04146, 2017.

DIGITAL LIBRARIES AND DATA INFRASTRUCTURES.

INVITED PAPER

Big Math Methods in Lobachevskii-DML Digital Library

A.M. Elizarov[0000−0003−2546−6897] and E.K. Lipachev[0000−0001−7789−2332]

Higher School of Information Technologies and Intelligent Systems; N.I. Lobachevskii Institute of Mathematics and Mechanics of Kazan (Volga region) Federal University,
35 Kremlevskaya ul., Kazan, Russia, 420008
amelizarov@gmail.com, elipachev@gmail.com

Abstract. We offer methods for the formation of digital collections from a set of documents (scientific articles, monographs, collections of reports), which are presented in various storage formats. Based on the analysis of the structure of documents and the stylistic features of their design, we have developed an algorithm for extracting the metadata of these documents. We present a software tool for dividing collections of articles into separate documents and the formation of their semantic presentation. On the example of a collection "Proceedings of N.I. Lobachevskii Mathematical Center", which have a different format and structure, we describe the algorithm for creating a digital collection and its inclusion in the Lobachevskii-DML.
Algorithms for replenishing the electronic collections of the Lobachevskii-DML digital library and generating metadata of documents of these collections in selected formats are presented. Services for the normalization of the Lobachevskii-DML digital library collection metadata were developed in accordance with the DTD rules and the NISO JATS and DBLP XML schemas. Algorithms for creating a mandatory and fundamental collection of metadata collections are presented in accordance with the rules of the European digital mathematical library EuDML.

Keywords: Electronic mathematical collections · Digital libraries · Formation and extraction of metadata · Semantic links of information objects · Metadata normalization services · Mandatory and fundamental metadata sets · Lobachevskii-DML digital library.

1 Introduction

The term "Big Data", which is currently widely used in various subject areas [1], in relation to mathematics requires certain clarifications: big data in mathematics is fundamentally different from big data in the current traditional understanding. In mathematics, all data is essential; moreover, in mathematical documents, many of their parts, especially formulas, are a peculiar code that requires decoding and special interpretation.

Information and communication technologies (ICT) have completely changed the life cycle of scientific documents both at the stages of their preparation and dissemination, and at the stage of obtaining results. The above fully applies to scientific activities in the field of mathematics. But when solving mathematical problems, the expectations from the use of ICT are significantly higher. Here we can draw an analogy with the way computers have completely eliminated manual calculations. The complexity of manual calculations, moreover, their routine, can be demonstrated by the example of numerous tables of function values. Such, for example, are the four-digit tables of V.M. Bradis [2], familiar to all mathematicians: they were created in 1921 and reprinted more than 60 times.

Computations have always required the use of specific methods and non-standard organizational solutions to cope with the volume (Volume is one of the characteristics of big data) and to overcome the barrier of the computing power of the individual. Riche de Prony (Gaspard Clair François Marie Riche de Prony) in 1791–1802 to compile "cadastral tables" containing logarithms (8 characters), created a "factory for calculation" ("usine à calculer"), dividing the calculators into sections: a section of theorists from five prominent mathematicians, including Legendre, a section of "calculators", the number which was 7–8 people familiar with numerical and analytical calculations, and a section of 60-80 people who were engaged only in addition and subtraction [3]–[6]. Speaking of Velocity as one of the characteristics of big data, the duration of manual calculations illustrates an example of calculating π: V. Shanks (William Shanks, 1873) spent 15 years calculating 707 characters of this number (but only 555 of them turned out to be true).

"Manual" calculations are a typical calculation practice used almost until the middle of the 20th century. After that, the created computers saved the scientists from tedious arithmetic operations. Today, on the simplest laptop, the calculation of Pi using the same algorithm that was used for manual counting will take less than a second. Humankind is expecting the same progress now not only in calculations. In the same way, intelligent computer tools should leave in the past time-consuming routine (and not only!) operations in Mathematics. In addition to computing and document preparation, intelligent search tools are needed, including recommendation systems for finding scientific articles that are close in content; terminological annotation services; personal information assistants and information platforms for publishing automation.

This article describes approaches to managing large collections of digital mathematical documents based on semantic methods and consistent with the principles of the World Digital Mathematical Library (WDML), as well as related to the areas constituting the Big Math concept. These approaches are being developed and already partially implemented in the framework of the project for creating the Lobachevskii–DML digital math library.

2 Big Data in Mathematics and Big Math

Mathematicians, as well as scientists of other specialties, in recent decades have faced such volumes of scientific documents that require the involvement of new methods of working with information. These methods should be primarily based on the use of intelligent software tools. Estimates of the growth in the volume of scientific production made today are fairly approximate and take into account only articles in scientific journals. As an example, we present the results of a calculation carried out by the Center for Science and Technology Research at the University of Leiden (SBF 2007). According to this Center (see, for example, [7]), the number of scientific publications in professional journals worldwide increased from about 686 thousand in 1990 to about 1,260 thousand in 2006, which corresponds to an increase of 84%. The annual growth rate calculated on this basis was more than 5%. At the same time, the number of scientific publications is growing faster than the world economy. In addition to journal articles, scientific knowledge is being disseminated today through such new forms of publications as academic blogs, social networks, and dynamic publications. These forms have already become widespread on the Web (see, for example, [8]).

Specialized software services are being developed for working with scientific content. Currently, computer support is used at all stages of the life cycle of a scientific document. Mathematical content has features that do not always allow using general-purpose software tools to work with it. The specificity of mathematical documents is determined, first of all, by the logical structure of texts presented in the form of a strict sequence of objects – definitions, statements and proofs. They clearly indicate or are implicitly hidden links with objects from other documents that are understandable only to a specialist in mathematics.

The presence of specialized formulas is another feature of mathematical documents, which requires the use of specialized software tools at all stages of the document life cycle, in particular, for their input and display. Such tools, as a rule, are developed by mathematicians themselves (for example, [9]–[13]).

Documents that contain similar texts may differ significantly in terms of the content laid down in the formulas contained in these documents. Moreover, absolutely identical formulations of theorems can have qualitative differences on the results declared in them. Examples are theorems on the improvement of approximation estimates or reference books on special sections of mathematics (see, for example, [14]). Therefore, without methods that use the semantics of not only texts, but also formulas, effective work with mathematical documents is impossible [15]–[18].

Big data in mathematics also manifests itself in studies that require consideration and analysis of numerous cases. For example, the classification of finite simple groups required the long-term efforts of a large group of mathematicians and is presented on more than 10,000 journal pages. An overview of this grand study is given in [19, 20]. The well-known problem of four colors was reduced to 1936 configurations and to create an algorithm for checking them on a computer [21, 22]. The validity of the computer proof was confirmed by G. Gonthier by the formalization in the Coq language in 2005 [23].

J. Carette, W.M. Farmer, M. Kohlhase and F. Rabe [24] proposed to use, by analogy with the term Big Data, the term Big Math to denote the field of creating methods and developing software systems to support mathematical research. They highlighted 5 main aspects of Big Math:

- Inference (output of statements by deduction);
- Computation (algorithmic transformation of representations of mathematical objects into forms that are easier to understand);
- Tabulation (creating static, specific data related to mathematical objects and structures that can be easily stored, queried and shared);
- Narration (bringing the results into a form that people can assimilate);
- Organization (modular organization of mathematical knowledge).

The main task of mathematical software systems today is to integrate the aspects that make up Big Math.

3 Integrating Mathematical Knowledge with Digital Mathematical Libraries

The system of digital mathematical libraries currently being created is intended to consolidate and make accessible both modern mathematical knowledge and the knowledge contained in articles and books published in the pre-digital period. To achieve this goal, in the framework of digital libraries, methods for managing digital information are developed that take into account the characteristics of the presentation of mathematical content (see, for example, [25, 26]).

The most important tasks in the management of mathematical knowledge are highlighted in [17, 26, 27]. The defining part of these problems can be solved with the help of digital mathematical libraries built using semantic technologies [26].

An overview of digital mathematical libraries from the point of view of the DELOS Digital Library Reference Model is given in [25]. These libraries are mainly national and carry out the task of consolidating the mathematical documents of their countries, primarily books and journal articles. Examples of such libraries are The Numdam French digital mathematics library [28] and the All-Russian Mathematical Portal Math-Net.Ru [29].

In the field of integration of mathematical knowledge, the most significant is the Global Digital Mathematics Library (GDML) initiative [30, 31]. The World Digital Mathematics Library (WDML) project put forward the idea of combining the entire corpus of digital mathematical documents in the distributed system of electronic collections as the main task [26]. The European Digital Mathematics Library (EuDML, https://initiative.eudml.org/) [32] project is aimed at integrating European mathematical resources. This project is considered as one of the stages of building the World Digital Mathematical Library.

4 Lobachevskii Digital Mathematical Library

In accordance with the basic principles of WDML, a digital library Lobachevskii Digital Mathematics Library (Lobachevskii-DML, https://lobachevskii-dml.ru/)

is being created at the Kazan University [33]. The construction of this library involves the development of management tools for mathematical content that take into account not only the specifics of mathematical texts, but also the peculiarities of processing Russian-language texts. Another objective of this digital library is the integration of the mathematical resources of Kazan University and their inclusion in the global scientific infrastructure, in particular, Math-Net.Ru and EuDML. To solve this problem, methods for the normalization of metadata are being developed in accordance with the schemes of international scientometric databases.

4.1 Use in the Organization of Digital Collections of Semantic Analysis Methods

In the project WDML [26] in the organization of digital collections proposed to use an object approach. It involves the analysis and processing of not only the documents themselves included in the collections, but also the objects contained in these documents (in particular, definitions, mathematical statements and their proofs). This section presents a number of methods that have been developed within the framework of this approach and are implemented in the formation of Lobachevskii-DML's digital scientific collections. These collections were formed as a result of processing an array of unstructured digitized mathematical documents, presented in various formats (.pdf, .tex, .doc, .docx), using the developed special methods. Approbation of the methods is performed on the journal archive “Proceedings of N.I. Lobachevskii Mathematical Center” for 1998–2018, containing more than 60 volumes.

Note that the main purpose of the “Proceedings ...” is the publication of materials of mathematical conferences. As a result, the majority of the volumes of the “Proceedings ...” contain several dozen articles with a limited (from a modern point of view) composition of metadata. Since 1998 (since the release of the first volume), several style rules have been used to prepare materials, which influenced the choice of formats and the design of articles in the collected collections. The prerequisites for creating a digital collection from the array of files “Proceedings ...” were the division of volumes into separate articles, the selection of metadata describing each article, the generation of additional metadata containing, in particular, the bibliographic description of the article, a relation to the article file in the digital collection, as well as relations to the profiles of the authors of the article on academic portals and scientometric databases (kpfu.ru, MathNet.ru, Scopus, etc.). The main steps in creating this digital collection are as follows.

At the first stage, the processed archive was clustered: the volumes of “Proceedings ...” were divided into classes in accordance with the similarity of their structure and design. For each class, a set of regular expression patterns was developed that define the rules for searching information blocks. The basis of this algorithm is the approach proposed in [34, 35]. The algorithm is implemented in the form of programs in the C# language, allowing to process files in $\TeX$, OpenXML (.docx) and .pdf formats. $\TeX$ files were processed using standard functions that implement text string operations. PDFLib (https://www.pdflib.com) and iTextSharp libraries (https://www.nuget.org/packages/iTextSharp/) were used to process PDF files. For documents presented in the form of .docx files, the word/document.xml file was extracted from the .docx archive in accordance with the Office OpenXML format (see, for example, [36]).

At the next stage, the metadata that describe both the volume as a whole and the articles included in it were selected from the array of files of the"Proceedings..." volumes. In particular, for all the articles of each volume were allocated their names, as well as the page numbers of their beginning and end. For this, an algorithm was developed that uses the structural homogeneity of each volume and the style uniqueness in the design of articles in it. In addition to the listed metadata, this algorithm allowed us to also highlight lists of authors, bibliography blocks and other metadata (for example, e-mail addresses and keywords), if they are present in the text.

Further, an XML-language was proposed for describing digital mathematical collections, which consists of a set of tags and XML-schemas based on the Journal Archiving and Interchange Tag Suite (https://jats.nlm.nih.gov/1.2d2/). In the notation of this language, on the basis of the data obtained at the stage of processing the initial array of files, a description of the collection "Proceedings..." was carried out.

Using the methods of text analysis [1, 37] from the documents of the digital collection, we have isolated the terms that make up the sets of keywords for inclusion in the metadata. The term extraction algorithm is a development of the approach proposed in [34, 35, 38].

The next step in creating a digital collection included the procedures for dividing each volume of "Proceedings..." into separate articles. To do this, from XML-files containing meta-descriptions of volumes, we read tags, whose attributes point to the starting and ending pages of articles. Next, we divide the files into separate documents, which are assigned names in accordance with the rules adopted in the digital collection. The process of selecting articles was organized using a program developed in Python using the functions of the PyPDF2 library (http://pybrary.net/pyPdf/).

Such metadata as authors' email addresses and their affiliation, we imported from authors' profiles that are presented on academic sites and in various scientific databases, and in parallel they were refined. In this procedure, the semantic links established in the process of forming a digital collection were applied. The corresponding algorithm is based on the method of [33, 35, 38].

The implementation of the algorithms described above allowed us to form a digital collection of the "Proceedings of N.I. Lobachevskii Mathematical Center" and together with the specified set of metadata to include it in the digital library Lobachevskii-DML.

4.2 Formats and Normalization of Metadata Documents of Digital Math Libraries

Metadata Formats. At present, publications on mathematics are indexed in many scientometric databases. These databases impose different requirements on the composition of the metadata of the documents included in them and the schemes for their presentation. On the other hand, digital math libraries also use various metadata formats when building their collections. This is partly due to the fact that the articles included in such collections, being published in journals in accordance with the rules established in them, differ in the requirements for the metadata used. These differences can be quite significant, primarily related to the composition of metadata and their format, and are most noticeable in the archival collections of scientific journals. For example, in many articles published before 2000, there are no keywords and annotations, and the affiliation of authors appeared only in articles of recent years. At the same time, the constantly expanding set of metadata used today testifies to their increasing role in the improvement of modern scientific communications. Thus, there is a need to develop both methods for extracting missing metadata from documents and methods for converting already created metadata into the formats of relevant scientometric databases. Note also that participation in such projects of integration of mathematical resources as EuDML (The European Digital Mathematics Library, https://initiative.eudml.org/) [32, 39], involves the provision of sets of metadata generated according to schemes of aggregators of mathematical resources.

Note that the metadata scheme of the digital mathematical library EuDML is described in [40]: the metadata is divided into basic, fundamental, and additional [41]. To describe journal articles in the EuDML project, XML schemas (NISO JATS V1.0) [42] are used. The mandatory set of EuDML metadata is minimal in composition and contains the title of the article in the original language, the names and surnames of the authors, the list of bibliographies, the unique identifier of the article (for example, doi) and the URL of its full text. The fundamental set of metadata, in addition to the required metadata, includes annotation of the article and keywords.

A number of electronic collections of the digital library Lobachevskii-DML are physically located in other digital libraries. Our tasks are to replenish such collections with additional metadata, as well as automatically selecting objects and establishing semantic links between them.

When forming the fundamental set of metadata of electronic collections stored on external resources, the metadata presented on these resources is initially imported. For this purpose, a program for extracting metadata from web pages and writing them in the XML-format of the digital library Lobachevskii-DML, as well as replenishing and subsequent conversion by EuDML schemas.

As an example, we will point out the archive of articles of the journal “Russian Mathematics (Izvestiya VUZ. Matematika)”. This journal collection is digitized, supplied with meta descriptions, presented on the portal MathNet.Ru (http://www.mathnet.ru/php/journal.phtml?jrnid=ivm) (see also [29]), and is also one of the collections digital library Lobachevskii-DML. The following steps are implemented for this collection.

Part of the metadata was imported from the “Citation in AMSBIB format” block of the MathNet.Ru portal. Then, keywords and a hyperlink to the

Springer Link portal page (https://link.springer.com/journal/11982) with the English version of the article were read from the web page. This information is included in the metadata, and a hyperlink is made.

The next step involves analyzing the web page of the English version of the article, extracting and recording metadata. Next, a personal identifier of this article was generated, which was proposed to be created as a string concatenation – journal identifier (attribute value "jrnid =") and article identifier (attribute value "paperid =") on the portal MathNet.Ru.

Normalization of Metadata. By normalization, we mean the use of methods for generating or transforming document metadata in accordance with the rules and XML-schemas of digital libraries and scientometric databases.

One of the most popular and respected computer science libraries is "Dblp Computer Science Bibliography" (DBLP, https://dblp.uni-trier.de/). A prerequisite for the inclusion of electronic collections in this library is the reorganization and normalization of the metadata of the relevant documents. Among the collections of the digital library Lobachevskii-DML, such is the collection of the "Russian Digital Libraries Journal" (https://elbib.ru/). An archive of articles published in this journal, starting in 2015, was chosen to prepare for indexing in DBLP. The necessary metadata are: publication identifier, the names and surnames of the authors, title of work, year of publication, volume, number, starting and ending pages of the article in the journal number and URL of the full text of the article.

Normalization to the DBLP format occurs in three stages: the extraction of the required metadata, the addition of metadata and their normalization into the desired format.

Using the program developed in C# and the System: XML extension tools, the collection files are processed sequentially and, as a result, a set of metadata is generated for each document. At the next stage, the metadata is updated with information about the article and its authors in English. This information is imported from the English version of the journal's site using the HTMLAgilityPack extension tool. Since the English-language information about the authors is incomplete – only the names and initials are indicated – the names are translated from the Russian-language page. The result of this work was the inclusion of the Russian Digital Libraries Journal and articles published in it in 2015–2018 in the DBLP database (https://dblp.uni-trier.de/db/journals/rdlj/).

Lobachevskii-DML Metadata Factory. As a rule, the term "metadata factory" refers to a set of software tools for managing metadata in digital libraries (see, for example, [28]). These tools are aimed at performing operations such as extracting metadata from digital documents, improving metadata, refining metadata, updating metadata and normalizing metadata into digital library formats and formats of scientometric databases. The structure of the metadata factory of the digital library Lobachevskii-DML also includes semantic transliteration services and a recommendatory system for refining scientific classifiers.

4.3 Digital Mathematical Ecosystem

On the Lobachevskii-DML digital library portal, the OntoMath digital ecosystem is presented, which is an essential part of this digital library [43]. The main components of this ecosystem are: mathematical ontologies Mocassin, OntoMath$^{\mathrm{Pro}}$ and OntoMath$^{\mathrm{Edu}}$, the semantic publishing platform, the semantic search service OntoMathSearch, recommender systems for the selection of scientific classifiers, search for related articles and terminological annotation.

Mathematical Ontologies. The concept of the Semantic Web assumes the semantic structuring of the Inter-net data space for its use by software agents, and the main tasks are the unification (compatibility) and binding of data from different sources. Most relevant to applying Linked Data principles is the LOD project. Its main advantage is in a standardized approach to the structuring and storage of integrated data that is loaded and presented in the form of RDF, that is, triplets of the "subject – predicate – object" type.

An important direction in the development of the Semantic Web domain was the development of ontologies of subject domains, including ontologies of the presentation of mathematical knowledge [44].

The representation and exchange of knowledge in any subject area is based on its conceptualization (see, for example, [17]). The communication process (both between people and between machines) uses a language with a dictionary containing a set of terms to denote elements of conceptualization. Successful communication requires that all its participants, first, share a common conceptualization and, second, use a common vocabulary. A means of solving this problem, as is known, is ontology. Ontology defines the basic concepts of a certain subject area and the relationship between them. The main components of ontology are classes, relations and axioms.

Mocassin Ontology [45] is an ontology of the logical structure of mathematical documents, designed for automatic analysis of mathematical publications in the LaTeX format. This ontology formally (in the OWL language) describes the semantics of the structural elements of mathematical documents (for example, theorems, lemmas, proofs, definitions, etc.) expressed in the form of classes and properties. In addition, the ontology contains the axioms of cardinality and transitivity.

The ontology of professional mathematics OntoMath$^{\mathrm{Pro}}$ [46, 47] is the ontology of mathematical knowledge, which is organized in the form of two hierarchies:

- hierarchies of areas of mathematics: mathematical logic, set theory, algebra, geometry, topology, and so on;
- hierarchies of mathematical objects: set, function, integral, elementary event, Lagrange polynomial, etc.

The OntoMath$^{\text{Pro}}$ ontology is developed in OWL-DL/RDFS and contains 3450 classes, 6 types of object properties, 3630 instances of the IS-A property, and 1140 instances of the remaining properties. It contains five types of relationships: Class → Subclass, Defined with the help, Associative relationship, Task → Solution method and Area of Mathematics → Mathematical object. Ontology concepts contain their name in Russian and English, definition, links to external resources from the Linked Open Data cloud, and links to other concepts. Objects of semantic annotation are also formulas associated with formulas, fragments of text that specify the descriptions of variable formulas.

Ontology of educational mathematics OntoMath$^{\text{Edu}}$. In the current version, this ontology is developed for the system description of the educational aspect of mathematical knowledge. The initial ontology design of OntoMath$^{\text{Edu}}$ is based on the OntoMath$^{\text{Pro}}$ ontology developed by us earlier and described above. A new conceptualization has been created, reflecting the conceptual system of mathematics that corresponds to school education. Professional terminology has been adapted to educational activities, in particular, the language of school mathematics. Relationships reflecting the didactic dependence between the concepts have been added to OntoMath$^{\text{Edu}}$. Ontology concepts contain their names in English, Russian, and Tatar languages, as well as basic definitions, relationships with other ontology concepts (associative relationships), and links to concepts from external data sets. The OntoMath$^{\text{Edu}}$ ontology is built on a set of OntoMath$^{\text{Pro}}$ basic ontology relationships such as taxonomic relation (ISA); the relationship between the mathematical object and the field of mathematics; the relationship between mathematical objects is "determined by"; the relationship between the task and the method of solving it; a new set of didactic relations was also introduced [49].

When creating the top level of ontology OntoMath$^{\text{Edu}}$, the planimetry section of the school mathematics course was selected as a pilot: the current version of the ontology contains 585 concepts related to the planimetry course of 5–9 classes of secondary school. The ontology structure contains type hierarchies; hierarchies of materialized relationships; hierarchy of roles and network of points of view. The specificity of school geometric knowledge was taken into account, therefore, when designing ontology, a number of relations between the concepts were singled out: "whole–part", "determined", relation of ontological dependence, "theorem–property", "theorem–characteristic", "found by formula" (see also [50, 51]).

5 Conclusion

This paper describes approaches to managing large collections of digital mathematical documents that are based on semantic methods and are consistent with the principles of the World Digital Mathematical Library (WDML). These approaches and methods fully relate to the areas that make up the new concept of Big Math. They are being developed and practically implemented as part of the creation of the Lobachevskii-DML digital math library. The main results mentioned are as follows.

Methods for the formation of digital collections from a set of documents – scientific articles, monographs, reports presented in various storage formats are proposed. Based on the analysis of the structure of documents and the stylistic features of their design, an algorithm for extracting their metadata has been developed.

In connection with the increasing role of metadata in the improvement of modern scientific communications, both methods for extracting missing metadata from documents and methods for converting already created metadata into the formats of relevant scientometric databases have been developed and described.

A software tool has been developed for dividing collections of articles into separate documents and forming their semantic presentation. For example, the set of “Proceedings of N.I. Lobachevskii Mathematical Center”, which have a different format and structure, describes an algorithm for creating a digital collection and its inclusion in the Lobachevskii-DML digital mathematical library.

Algorithms for enriching the electronic collections of the Lobachevskii-DML digital library and generating metadata of documents of these collections in selected formats are presented.

Services for the normalization of the collection metadata of the Lobachevskii-DML digital library have been developed in accordance with the DTD rules and NISO JATS and DBLP XML schemas. By normalization, we mean the use of methods for generating or transforming document metadata in accordance with the rules and XML schemas of digital libraries and scientometric databases.

Algorithms for creating a mandatory and fundamental collection of metadata collections are presented in accordance with the rules of the European digital mathematical library EuDML.

The digital ecosystem OntoMath, which is the most important part of the Lobachevskii-DML digital library, is described. The main components of this ecosystem are: mathematical ontologies Mocassin, OntoMath$^{\text{Pro}}$ and OntoMath$^{\text{Edu}}$, the semantic publishing platform, the semantic search service OntoMathSearch, recommender systems for the selection of scientific classifiers, search for related articles and terminological annotation.

Acknowledgments

This work was funded by the subsidy allocated to Kazan Federal University for the state assignment in the sphere of scientific activities, grant agreement No. 1.2368.2017) and with partial financial support of the Russian Foundation for Basic Research and the Government of the Republic of Tatarstan, within the framework of scientific project No. 18-47-160012.

References

1. Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data. EMC. Education Services (ed.), Wiley (2015).

2. Bradis, V.M.: Four-digit mathematical tables. Moscow: Drofa, 2019.
3. Riche de Prony: Tables des logarithmes, sinus et tangentes pour la division décimale du quart de cercle calcul?es avec 8 ou 9 décimales pour être imprimées avec 7 décimales exactes au bureau du Cadastre, https://patrimoine.enpc.fr/exhibits/show/dataincognita/item/1817. Last accessed 16 May 2019
4. Bulletin de bibliographie, d'histoire et de biographie mathématiques. Notice sur la découverte des logarithmes. Nouvelles annales de mathématiques. Journal des candidats aux écoles polytechnique et normale, Serie 1, vol. 14, pp. 1–204 (1855) (Additional pages), http://www.numdam.org/item/NAM_1855_1_14__S1_0/. Last accessed 16 May 2019
5. Peaucelle, J.L.: Le détail du calendrier de calcul des tables de Prony de 1791 à 1802. http://rybn.org/human_computers/articles/calcul_des_tables_de_prony.pdf. Last accessed 16 May 2019
6. Roegel, D.: A reconstruction of the "Tables des logarithmes à huit decimals" from the French "Service géographique de l'armée" (1891). [Research Report] 2010. inria-00543952. https://hal.inria.fr/inria-00543952. Last accessed 16 May 2019
7. Binswanger, M.: Excellence by Nonsense: The Competition for Publications in Modern Science. In: Bartling, S., Friesike, S. (eds). Opening Science. The Evolving Guide on How the Internet is Changing Research, Collaboration and Scholarly Publishing, pp. 49–72. Springer International Publishing (2014). https://doi.org/10.1007/978-3-319-00026-8_3
8. Heller, L., The, R., and Bartling, S.: Dynamic Publication Formats and Collaborative Authoring. In: Bartling, S., Friesike, S. (eds). Opening Science. The Evolving Guide on How the Internet is Changing Research, Collaboration and Scholarly Publishing, pp. 191–211. Springer International Publishing (2014). https://doi.org/10.1007/978-3-319-00026-8_13
9. Knuth, D.E.: The TeX book. Addison-Wesley Publishing Company (1984, 1986, 1991).
10. Cervone, D.: Math Jax: A Platform for Mathematics on the Web. Notices of the AMS **59**, 312–316 (2012).
11. Tantau, T.: The TikZ and PGF Packages. Manual for version 3.1.4a (2019). https://pgf-tikz.github.io/pgf/pgfmanual.pdf. Last accessed 16 May 2019
12. Tools & Technical Specifications. EuDML Enhancer toolset demos. https://initiative.eudml.org/tools-technical-specifications. Last accessed 16 May 2019
13. OpenDreamKit. https://kwarc.info/projects/odk/. Last accessed 16 May 2019
14. Polyanin, A.D. and Zaitsev, V.F.: Handbook of Ordinary Differential Equations. Exact Solutions, Methods, and Problems. CRC Press. Taylor & Francis Group (2018).
15. Kohlhase, M.: Semantic Markup in TeX/LaTeX (2019). http://ctan.altspu.ru/macros/latex/ contrib/stex/sty/stex/stex.pdf. Last accessed 16 May 2019
16. Kohlhase, M.: OMDoc – an open markup format for mathematical documents [Version 1.2]. Springer, Berlin (2006).
17. Elizarov, A.M., Kirillovich, A.V., Lipachev, E.K., Nevzorova, O.A., Solovyev, V.D., and Zhiltsov, N.G.: Mathematical knowledge representation: semantic models and for-malisms. Lobachevskii Journal of Mathematics **35** (4), 348–354 (2014). https://doi.org/10.1134/S1995080214040143
18. Elizarov, A., Kirillovich, A., Lipachev, E., and Nevzorova, O.: Semantic formula search in digital mathematical libraries. Proc. of the 2nd Russia and Pacific Conf.

on Comp. Technology and Applications (RPC 2017). IEEE, pp. 39-43 (2017). https://doi.org/10.1109/RPC.2017.8168063

19. Gorenstein, D.: The Enormous Theorem. Scientific American **253** (6), 104–115 (1985).
20. Solomon, R.: A brief history of the classification of the finite simple groups. Bulletin of the AMS. New Series **38** (3), 315–352 (2001).
21. Appel, K. and Haken, W.: Every map is four Colourable. Bulletin of the AMS **82**, 711–712 (1986).
22. Appel, K. and Haken, W.: Every map is four Colourable. Contemporary Mathematics **98** (1989).
23. Gonthier, G.: Formal Proof – The Four-Color Theorem. Notices of the AMS **55** (11), 1382–1393 (2008).
24. Carette, J., Farmer, W.M., Kohlhase, M., and Rabe, F.: Big Math and the One-Brain Barrier. A Position Paper and Architecture Proposal. arXiv:1904.10405v1 [cs.MS] 23 Apr 2019.
25. Elizarov, A.M., Lipachev, E.K., and Zuev, D.S.: Digital mathematical libraries: Overview of implementations and content management services. CEUR Workshop Proceedings **2022**, 317–325 (2017).
26. Developing a 21st Century Global Library for Mathematics Research. The National Academies Press,Washington (2014). https://doi.org/10.17226/18619
27. Elizarov, A.M., Kirillovich, A.V., Lipachev, E.K., and Nevzorova, O.A.: Mathematical Knowledge Management: Ontological Models and Digital Technology. CEUR Workshop Proceedings **1752**, 44–50 (2016).
28. Bouche, T. and Labbe, O.: The New Numdam platform. CICM 2017: Intelligent Computer Mathematics, 70–82 (2017).
29. Chebukov, D.E., Izaak, A.D., Misyurina, O.G., Pupyrev, Yu.A., and Zhizhchenko, A.B.: Math-Net.Ru as a Digital Archive of the Russian Mathematical Knowledge from the XIX Century to Today. Intelligent Computer Mathematics. LNCS **7961**, 344–348 (2013). https://doi.org/10.1007/978-3-642-39320-4_26
30. Ion, P.: The Effort to Realize a Global Digital Mathematics Library. In: Greuel, G.-M. et al. (eds.). ICMS 2016, LNCS, vol. 9725, pp. 458–466. Springer (2016). https://doi.org/10.1007/978-3-319-42432-3 59
31. Ion, P.D.F. and Watt, S.M.: The Global Digital Mathematics Library and the International Mathematical Knowledge Trust. ICM 2017: Intelligent Computer Mathematics, 2017. LNAI, vol. 10383, pp. 56–69. Springer, 2017. https://doi.org/10.1007/978-3-319-62075-6_5
32. Ion, P.D.F. and Watt, S.M.: The Global Digital Mathematics Library and the International Mathematical Knowledge Trust. ICM 2017: Intelligent Computer Mathematics, 2017. LNAI, vol. 10383, pp. 56–69. Springer, 2017. https://doi.org/10.1007/978-3-319-62075-6_5
33. Bouche, T.: Reviving the free public scientific library in the digital age? the EuDML project. In: Kaiser, K., Krantz, S.G., Wegner, B. (eds.) Topics and Issues in Electronic Publishing JMM/AMS Special Session. FIZ Karlsruhe, pp. 57–80 (2013). https://www.emis.de/proceedings/TIEP2013/05bouche.pdf. Last accessed 16 May 2019
34. Elizarov, A., Khaydarov, S., and Lipachev, E.: Scientific documents ontologies for semantic representation of digital libraries. Second Russia and Pacific Conf. on Computer Technology and Applications (RPC). Vladivostok, Russky Island, Russia 25–29 September, pp. 1–5 (2017). https://doi.org/10.1109/RPC.2017.8168064

35. Batyrshina, R.R.: Method for extracting terms in digital mathematical collections. Proc. of the N.I. Lobachevskii Math. Center. Kazan: Kazan Math. Soc. Publ. **55**, 24–26 (2017).
36. Standard ECMA-376 Office Open XML File Formats. http://www.ecmainternational.org/publications/standards/Ecma-376.htm. Last accessed 16 May 2019
37. Ingersoll, G.S., Morton, T.S., and Farris, A.L.: Taming Text. How to Find, Organize, and Manipulate It. Manning Publications Co. (2013).
38. Sabitova, E.M.: Algorithm for extracting connections in scientific digital collections. Proc. of the N.I. Lobachevskii Math. Center. Kazan: Kazan Math. Soc. Publ. **55**, 123–126 (2017).
39. Bouche, T. and Rákosník, J.: Report on the EuDML External Cooperation Model. In: Kaiser K., Krantz S.G., Wegner B. (eds.) Topics and Issues in Electronic Publishing, JMM, Special Session, San Diego, 99–108 (2013).
40. Jost, M., Bouche, T., Goutorbe, C., and Jorda, J.P.: D3.2: The EuDML metadata schema. http://www.mathdoc.fr/publis/d3.2-v1.6.pdf. Last accessed 16 May 2019
41. EuDML metadata schema specification (v2.0-final). https://initiative.eudml.org/eudml-metadata-schema-specification-v20-final. Last accessed 16 May 2019
42. Journal Article Tag Suite. NISO JATS V1.0. https://jats.nlm.nih.gov/1.0/. Last accessed 16 May 2019
43. Khaydarov, S. and Yamalutdinova, G.: Recommender System of Physical and Mathematical Documents Classification. CEUR Workshop Proceedings **2260**, 480–486 (2018).
44. Elizarov, A., Kirillovich, A., Lipachev, E., and Nevzorova, O.: Digital Ecosystem OntoMath: Mathematical Knowledge Analytics and Management. CCIS **706**, 33–46. Springer (2017). https://doi.org/10.1007/978-3-319-57135-5_3
45. Lange, C.: Ontologies and languages for representing mathematical knowledge on the Semantic Web. Semantic Web **4** (2), 119–158 (2013). https://doi.org/10.3233/SW-2012-0059
46. Solovyev, V. and Zhiltsov, N.: Logical Structure Analysis of Scientific Publications in Mathematics. Proc. of the Int. Conf. on Web Intelligence, Mining and Semantics (WIMS'11). ACM **21**, 1–9 (2011)
47. Elizarov, A.M., Zhizhchenko, A.B., Zhil'tsov, N.G., Kirillovich, A.V., and Lipachev, E.K.: Mathematical Knowledge Ontologies and Recommender Systems for Collections of Documents in Physics and Mathematics. Doklady Mathematics **93** (2), 231–233 (2016). https://doi.org/10.1134/S1064562416020174
48. Nevzorova, O., Zhiltsov, N., Kirillovich, A., and Lipachev, E.: OntoMathPRO Ontology: A Linked Data Hub for Mathematics. CCIS **468**, 105–119. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-11716-4_9
49. Elizarov, A., Kirillovich, A., Lipachev, E., Nevzorova, O., and Shakirova, L.: Open Linked Data and Ontologies in Mathematics Education. CEUR Workshop Proceedings **2260**, 186–196 (2018).
50. Kirillovich, A., Shakirova, L., Falileeva, M., and Lipachev, E.: Towards an Educational Mathematical Ontology. L. Gómez Chova, et al. (eds). 13th International Technology, Education and Development Conference (INTED2019), Valencia, Spain, March 11-13, 2019. IATED, 6823–6829 (2019).
51. Elizarov, A.M., Lipachev, E.K., and Khaydarov, S.M.: Method of automated selection of reviewers of scientific articles, implemented in the scientific journal information system. Proceedings of the 21th Conference Scientific Services & Internet (SSI 2019), Novorossiysk-Abrau, Russia, September 23-28, 2019.

DIRAC System as a Mediator Between Hybrid Resources and Data Intensive Domains

Vladimir Korenkov[1,3], Igor Pelevanyuk[1,3], and Andrei Tsaregorodtsev[2,3]

[1] Joint Institute for Nuclear Research, Dubna, Russia
korenkov@jinr.ru, pelevanyuk@jinr.ru
[2] CPPM, Aix-Marseille University, CNRS/IN2P3, Marseille, France
atsareg@in2p3.fr
[3] Plekhanov Russian Economics University, Moscow

Abstract. Data and computing-intensive applications in scientific research are becoming more and more common. And, since different computing solutions have different protocols and architectures, they should be chosen wisely during the design stage. In a modern world of diverse computing resources such as grids, clouds, and supercomputers the choice can be difficult. Software developed for integration of various computing and storage resources into a single infrastructure, the so-called interware, is intended to facilitate this choice. The DIRAC interware is one of these products. It proved to be an effective solution for many experiments in High Energy Physics and some other areas of science. The DIRAC interware was deployed in the Joint Institute for Nuclear Research to serve the needs of different scientific groups by providing a single interface to a variety of computing resources: grid cluster, computing cloud, supercomputer Govorun, disk, and tape storage systems. The DIRAC based solution was proposed for the Baryonic Matter at Nuclotron experiment which is in operation now as well as for the future experiment Multi-Purpose Detector on the Nuclotron-based Ion Collider fAcility. Both experiments have requirements making the use of heterogeneous computing resources necessary.

Keywords: Grid computing, Hybrid distributed computing systems, Supercomputers, DIRAC

1 Introduction

Data intensive applications became now an essential mean for getting insights of new scientific phenomena while analyzing huge data volumes collected by modern experimental setups. For example, the data recording to the tape system at CERN exceeded in total 10 Petabytes per month in 2018 for all the 4 LHC experiments. In 2021, with the start of the Run 3 phase of the LHC program, the experiments will resume data taking with considerably increased rates. The needs for the computing and storage capacity of the LHCb experiment, for instance, will increase by an order of magnitude [1].

In a more distant future, with the start of the LHC Run 4 phase, the projected data storage needs of the experiments are estimated to exceed 10 Exabyte’s. These are unprecedented volumes of data to be processed by distributed computing systems, which are being adapted now to cope with the new requirements.

Other scientific domains are quickly approaching the same collected data volumes: astronomy, brain research, genomics and proteomics, material science [3]. For example, the SKA large radio astronomy experiment [2] is planned to produce about 3 Petabytes of data daily when it will come into full operation in 2023.

The needs of the LHC experiments in data processing were satisfied by the infrastructure of the World LHCb Computing Grid (WLCG). The infrastructure still delivers the majority of computing and storage resources for these experiments. It is well suited for processing the LHC data ensuring massively parallel data treatment in a High Throughput Computing (HTC) paradigm. WLCG succeeded in putting together hundreds of computing centers of different sizes but with similar properties, typically providing clusters of commodity processors under control of one of the batch systems, e.g. LSF, Torque or HTCondor. However, new data analysis algorithms necessary for the upcoming data challenges require new level of parallelism and new types of computing resources. These resources are provided, in particular, by supercomputers or High Performance Computing (HPC) centers. The number of HPC centers is increasing and there is a clear need in setting up infrastructures allowing scientific communities to access multiple HPC centers in a uniform way as it is done in the grid systems.

Another trend in massive computing consists in provisioning resources via cloud interfaces. Both private and commercial clouds are available now to scientific communities. However, the diversity of interfaces and usage policies makes it difficult to use multiple clouds for applications of a particular community. Therefore, providing uniform access to resources of various cloud providers would increase flexibility and the total amount of available computing capacity for a given scientific collaboration.

Large scientific collaborations typically include multiple participating institutes and laboratories. Some of the participants have considerable computing and storage capacity that they can share with the rest of the collaboration. With the grid systems this can be achieved by installing complex software, the so-called grid middleware, and running standard services like Computing and Storage Elements. For managers of local computing resources who are usually not experts in the grid middleware, this represents a huge complication and often results in underused resources that would otherwise be beneficial for the large collaborations. Tools for easy incorporation of such resources can considerably increase the efficiency of their usage.

The DIRAC Interware project is providing a framework for building distributed computing systems using resources of all different types mentioned above and putting minimal requirements on the software and services that should be operated by the resources providers. Developed originally for the LHCb experiment at LHC, CERN, the DIRAC interware was generalized to be applicable for a wide range of applications. It can be used to build independent distributed computing infrastructures as well as to provide services for existing projects. DIRAC is used by a number of High Energy Physics and Astrophysics experiments but it is also providing services for a number of general-purpose grid infrastructures, for example, national grids in France

[4] and Great Britain [5]. The EGI Workload Manager is the DIRAC service provided as part of the European Grid Infrastructure service catalog. It is one of the services of the European Open Science Cloud (EOSC) project inaugurated in the end of 2018 [6]. The EGI Workload Manager provides access to grid and cloud resources of the EGI infrastructure for over 500 registered users.

In this paper we describe the DIRAC based infrastructure deployed at the Joint Institute for Nuclear Research, Dubna, putting together a number of local computing clusters as well as connecting cloud resources from JINR member institutions.

2 DIRAC Interware

The DIRAC Interware project provides a development framework and a large number of ready-to-use components to build distributed computing systems of arbitrary complexity. DIRAC services ensure integration of computing and storage resources of different types and provide all the necessary tools for managing user tasks and data in distributed environments [7]. Managing both workloads and data within the same framework increases the efficiency of data processing systems of large user communities while minimizing the effort for maintenance and operation of the complete infrastructure.

The DIRAC software is constantly evolving to follow changes in the technology and interfaces of available computing and storage resources. As a result, most of existing HTC, HPC and cloud resources can be interconnected with the DIRAC Interware. In order to meet the needs of large scientific communities, the computing systems should fulfill several requirements. In particular, it should be easy to describe, execute and monitor complex workflows in a secure way respecting predefined policies of usage of common resources.

2.1 Massive Operations

Usual workflows of large scientific collaborations consist and creation and execution of large numbers of similar computational and data management tasks. DIRAC is providing support for massive operations with its Transformation System. The system allows definition of Transformations – recipes to create certain operations triggered by the availability of data with required properties. Operations can be of any type: submission of jobs to computing resources, data replication or removal, etc. Each Transformation consumes some data and derives (“transforms”) new data, which, in turn, can form input for another Transformation. Therefore, Transformations can be chained creating data driven workflows of any complexity. Data production pipelines of large scientific communities based on DIRAC are using heavily the Transformation System defining many hundreds of different Transformations. Therefore, each large project developed its own system to manage large workflows each consisting of many Transformations. There was a clear need to simplify the task of managing complex workflows for the new communities. In order to do that a new system was introduced in DIRAC – Production System. The new system is based on the experience of several community specific workflow management systems and provides a uniform way to create a set of Transformations interconnected via their input/output data filters. It helps production managers to monitor the execution of so created workflows, evaluate the overall progress of the workflow advancement and validate the results with an automated verification of all the elementary tasks.

2.2 Multi-community Services

In most of the currently existing multi-community grid infrastructures the security of all operations is based on the X509 PKI infrastructure. In this solution, each user has to, first, obtain a security certificate from one of Certification Authorities (CA) recognized by the infrastructure. The certificate should be then registered in a service holding a registry of all the users of a given Virtual Organization (VO). The user registry keeps the identity information together with associated rights of a given user. In order to access grid resources, users are generating proxy certificates which can be delegated to grid remote services in order to perform operation on the user's behalf.

The X509 standard based security is well supported in academia institutions but is not well suited for other researchers, for example, working in universities. On the other hand, there are well-established industry standards developed mostly for the web applications that allow identification of users as well as delegation of user rights to remote application servers. Therefore, grid projects started migration to the new security infrastructure based on the OAuth2/OIDC technology. With this technology, user's registration is done by local identity providers, for example, a university LDAP index. On the grid level a Single-Sign-On (SSO) solution is provided by federation of multiple identity providers to ensure mutual recognition of user security tokens. In particular, the EGI infrastructure has come up with the Check-In SSO service as a federated user identity provider.

The DIRAC user management subsystem was recently updated in order to support this technology. Users can be identified and registered in DIRAC based on their Check-In SSO tokens which contain also additional user metadata, e.g. membership in VOs, user roles and rights. This metadata are used to define user membership in the DIRAC groups, which define user rights within the DIRAC framework. This allows managing community policies, such as resources access rights and usage priorities that will be applied by DIRAC to the user payloads. The DIRAC implementation of the new security framework is generic and can be easily configured to work with other SSO systems.

2.3 DIRAC Software Evolution

The intensity of usage of the DIRAC services is increasing and the software must evolve to cope with the new requirements. This process is mostly driven by the needs of the LHCb experiment, which remains the main consumer and developer of the DIRAC software. As was mentioned above, the order of magnitude increase in the data acquisition rate of LHCb in 2021 dictates revision of the technologies used in its data processing solutions.

Several new technologies were introduced recently into the DIRAC software stack. The use of Message Queue (MQ) services allows passing messages between distributed DIRAC components in an asynchronous way with the possibility of message buffering in case of system congestions. The STOMP message passing protocol is used and all the MQ service supporting this protocol can be used, e.g. ActiveMQ, RabbitMQ and others. The MQ mechanism for the DIRAC component communications is considered to be complementary to the base Service Oriented Architecture (SOA) employed by DIRAC. This solution increases the overall system scalability and resilience.

The DIRAC services states are kept in relational databases using MySQL servers. The MySQL databases have shown very stable operation over the years of usage. However, the increased amount of data to be stored in databases limits the efficiency of queries and new solutions are necessary. The so-called NoSQL databases have excellent scalability properties and can help in increasing the efficiency of the DIRAC components. The ElasticSearch NoSQL (ES) database solution was applied in several DIRAC subsystems. In particular, the Monitoring System, which is used to monitor the current consumption of the computing resources, was migrated to the use of the ES based solution. This information is essential in implementation of the priority policies based on the history of the resources consumption to ensure fair sharing of the common community resources.

This and other additions and improvements in the DIRAC software aim at the overall increase of the system efficiency and scalability to meet requirements of multiple scientific communities relying on DIRAC services for their computing projects.

3 JINR DIRAC Installation

The Joint Institute for Nuclear Research is an international intergovernmental organization, a world-famous scientific center that is a unique example of the integration of fundamental theoretical and experimental research. It consists of seven laboratories: Laboratory of High Energy Physics, Laboratory of Nuclear Problems, Laboratory of Theoretical Physics, Laboratory of Neutron Physics, Laboratory of Nuclear Reactions, Laboratory of Information Technologies, Laboratory of Radiation Biology. Each laboratory being comparable with a large institute in the scale and scope of investigations performed.

JINR has powerful high-productive computing environment that is integrated into the world computer network through high-speed communication channels. The basis of the computer infrastructure of the Institute is the Multifunctional Information Computer Complex (MICC). It consists of several large components: grid cluster, computing cloud, supercomputer Govorun. Each component has its features, advantages, and disadvantages. Different access procedures, different configuration and connection with different storage systems do not allow simple usage of all of them together for one set of tasks.

3.1 Computing Resources

Grid cluster. The JINR grid infrastructure is represented by the Tier1 center for the CMS experiment at the LHC and the Tier2 center.

After the recent upgrade, the data processing system at the JINR CMS Tier1 consists of 415 64-bit nodes: 2 x CPU, 6–16 cores/CPU that form altogether 9200 cores for batch processing [8]. The Torque 4.2.10/Maui 3.3.2 software (custom build) is used as a resource manager and a task scheduler. The computing resources of the Tier2 center consist of 4,128 cores. The Tier2 center at JINR provides data processing for all four experiments at the LHC (Alice, ATLAS, CMS, LHCb) and apart from that supports many virtual organizations (VO) that are not members of the LHC (BES, BIOMED, COMPASS, MPD, NOvA, STAR, ILC).

Grid cluster is an example of a High-Throughput Computing paradigm. It means that the primary task of this cluster is to run thousands of independent processes at the same time. Independent means that once a process has started and until it finishes, the process does not rely on any input that is being produced at the same moment by other processes.

Jobs may be sent to the grid using CREAM Computing Element – service installed in JINR specifically for grid jobs. Computing element works as an interface to the local batch farm. Its primary task is to authenticate the owner of the job and redirect it to the right queue. For the users, it is required to have X509 certificate and be a member of Virtual Organization supported by the Computing Element.

Cloud infrastructure. The JINR Cloud [9] is based on an open-source platform for managing heterogeneous distributed data center infrastructures – OpenNebula 5.4. The JINR cloud resources were increased up to 1564 CPU cores and 8.1 TB of RAM in total. Cloud infrastructure is used primarily for two purposes: to create personal virtual machines and to create virtual machines to serve as worker nodes for jobs. We are going to focus on the second purpose.

The biggest advantage of cloud resources as computing capacity is their flexibility. In case of grid or batch resources, several jobs working on one worker node share between them: operating system, CPU cores, RAM, HDD/SSD storage, disk Input/Output capabilities, and network bandwidth. If a job needs more disk space or RAM it is not straightforward to submit the job to the grid without the help of administrators, who in most cases have to create a dedicated queue for this particular kind of jobs. In the case of clouds, it is much easier to provide a specific resource that the job requires. It may be a virtual machine with a large disk, specific operating system, required number of CPU cores, RAM capacity and network.

When a job destinated to the cloud enters the system the corresponding virtual machine is created by DIRAC using the OpenNebula API. During the contextualization process, the DIRAC Pilot is installed in the VM and configured to receive jobs for this cloud resource. Once the job is finished, the pilot attempts to get the next job. If there are no more jobs for the cloud, the pilot will request the VM shutdown. The pilot in the cloud environment is not limited by the time and may work for weeks. These features make cloud resources perfect for specific tasks with unusual requirements.

Govorun supercomputer. The Supercomputer Govorun was put into production in March 2018[10]. It is a heterogeneous platform built on several processors' technologies: GPU part and two CPU parts. GPU part unites 5 servers DGX-1. Each server consists of 8 NVIDIA Tesla V100 processors. The CPU part is a high dense liquid-cooled system. Two types of processors are used inside: Intel Xeon Phi 7290(21 servers) and Intel Xeon Gold 6154(40 servers). The total performance of all the three parts is 1 PFlops for operations with single precision and 0.5 PFlops for double precision. SLURM 14.11.6 is used as the local workload manager. Three partitions were created to subdivide tasks in the supercomputer: gpu, cpu, phi.

The supercomputer is used for tasks, which require massive parallel computations. For example: to solve problems of lattice quantum chromodynamics for studying the properties of hadronic matter with high energy density and baryon charge and in presence of strong electromagnetic fields, mathematical modeling of the antiproton-proton and antiproton-nucleus collisions with the use of different generators. It is also used for simulation of collision dynamics of relativistic heavy ions for the future MPD experiment on the NICA collider.

Right now, the supercomputer utilizes its own authentication and authorization system. Every user of the supercomputer should be registered and allowed to send jobs. Sometimes a part of the supercomputer is free from parallel tasks and may be used as a standard batch system. Special user was created for DIRAC. All jobs sent to the Govorun are executed with this user identity. This frees actual users from additional registration procedures.

3.2 Storage Resources

EOS storage on disks. EOS [11] is a multi-protocol disk-only storage system developed at CERN since 2010 to store physics analysis data physics experiments (including the LHC experiments). Having a highly-scalable hierarchical namespace, and with the data access possible by the XROOT protocol, it was initially used for physics data storage. Today, EOS provides storage for both physics and user use cases. For the user authentication, EOS supports Kerberos (for local access) and X.509 certificates for grid access. To ease experiment workflow integration, SRM as well as GridFTP access is provided. EOS supports the XROOT third-party copy mechanism from/to other XROOT enabled storage services.

The EOS was successfully integrated into the MICC structure. The NICA experiments already use EOS for data storage. At the moment there are ~200TB of “raw” BM@N data and ~84GB of simulated MPD data stored in the EOS instance. EOS is visible as a local file system on the MICC worker nodes. It allows users authorized by the Kerberos5 protocol to read and write data. A dedicated service was installed to allow usage of X509 certificates with VOMS extensions.

dCache disk and tape storage. The core part of the dCache has been proven to efficiently combine heterogeneous disk storage systems of the order of several hundreds TBs and present its data repository as a single filesystem tree. It takes care of data, failing hardware and makes sure, if configured, that at least a minimum number of copies of each dataset resides within the system to ensure high data availability in case of disk server maintenance or failure. Furthermore, dCache supports a large set of standard access protocols to the data repository and its namespace. It supports DCAP, SRM, GridFTP, and xRootD [12].

dCache at JINR consists of two parts: disk storage and tape storage. The disk part operations are similar to EOS. The tape works through the dedicated disk buffer servers. When data are uploaded to the dCache tape part, they are first uploaded to the disk buffer. If the disk buffer is occupied above a certain threshold (which is 80% in our case), all the data is moved from disk to tape and removed from the disk buffer. While data stay in the buffer, access to them is similar to access to the dCache disk data. But once the data are moved to tape and removed from the disk, access to them may require time. The time required to select the right tape and transfer data from tape to the disk depends on the tape library task queue. Generally, the time varies from 20 seconds up to several minutes.

Tape library should be used only for archive storage and preferably for big files. Otherwise, it may bring unnecessary load on the tape library. It is much easier to write many small files to the tape than to read it back.

Ceph storage. Software-defined storage (SDS) based on the Ceph technology is one of the key components of the JINR cloud infrastructure. It runs in production mode since the end of 2017. It delivers object, block and file storage in one unified system. Currently, the total amount of raw disk space in that SDS is about 1 PB. Due to triple replication, effective disk space available for users is about 330 TB. Users of Ceph can attach part of the storage to a computer using the FUSE disk mounting mechanism. After that, it is possible to read and write data to the remote storage as if it is connected directly to the computer.

The Ceph storage was integrated into DIRAC installation for tests. Since Ceph does not allow authentication by X509 certificates with VOMS extensions, a dedicated virtual machine was configured to host DIRAC Storage Element – a standard service which works as a proxy to a file system. It checks certificates with VOMS extensions before allowing writing and reading to a dedicated directory. Right now, Ceph storage does not allow massive transfers since it relies on one server with Ceph attached by FUSE. The test demonstrated that the maximum speed of transfer is not exceeding 100 MB/s which is a consequence of 1Gb network connection. The way to increase the performance of this storage is an improvement of the network speed up to 10Gb/s and a possible creation of additional DIRAC Storage Elements which can share the load between themselves.

Performance test of EOS and disk dCache. In the case of massive data processing, it is crucial to know the limitation of different components. The limitations may depend on the usage of resources. In many use-cases it is crucial to transfer some amount of data first, so we decided to test storage elements. The synthetic test was proposed: run many jobs on one computing resource, make them start download of all the same data at the same moment, measure how much time it takes to get the file.

Every test job had to go through the following steps:

1. Start execution on the worker node.
2. Check the transfer start time.
3. Wait until the transfer time moment.

4. Start the transfer.

5. When the transfer is done, report the information about the duration of the transfer.

6. Remove the downloaded file.

Two storage systems were chosen for the tests: EOS and dCache since only they are accessible for read and write on all the computing resources right now. We chose the test file size to be 3 GB. The amount of test jobs in one test campaign depends on the number of free CPU cores in our infrastructure. We initiate 200 jobs during one test campaign. Not all of them could start at the same time, which means that during the test less than 200 jobs may download data. This is taken into consideration when we calculate total transfer speed.

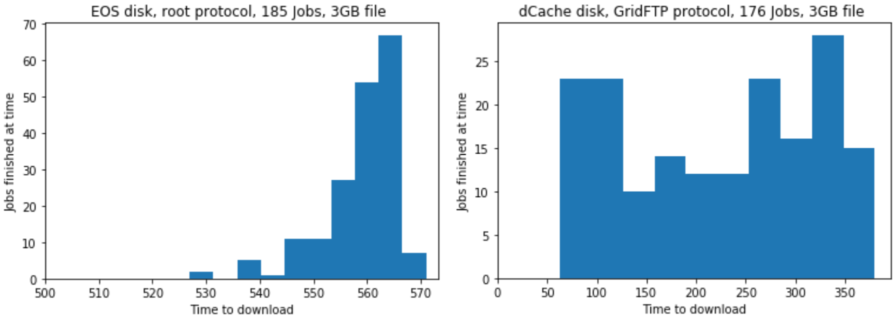


Fig. 4. Number of transfers finished at the time

Several test campaigns were performed to evaluate variance between the tests, but all of them showed similar results after all. Two representative examples were chosen to demonstrate the rates (see Fig. 1). To calculate transfer speed the following formula was used:

$$Transfer\ speed = \frac{Total\ data\ transferred}{Longest\ transfer\ duration}.$$

This formula allows calculation of the worst transfer speed of individual file during the test campaign. For EOS calculated transfer speed was 990 GB/s on 200 jobs and for dCache it was 1390 GB/s in 176 jobs. It should be mentioned that all the tests were performed on a working infrastructure, so some minor interference may be caused by other activities. On the other hand, demonstrated plots represent real transfers performed under normal conditions.

The numbers described above demonstrate that the computing and storage infrastructure at JINR is quite extensive and diverse. Nowadays, different components are used directly for different tasks. So, the workflows are bound to dedicated resources and switching between them would not be an easy task at least. Sometimes different components could be separated by a slower network, different authentication systems and different protocols. This problem becomes visible when one of the resources is overloaded while others are underloaded. In the case of good interoperability between the components, it would be possible to easily switch between them.

Of course, the resources are not fully available for all the tasks. They have to provide pledges for different tasks and experiments, but still, they could be underloaded. And since there are tasks that should not be necessarily bound to particular resources, it would be beneficial to have a mechanism to use them in some uniform way by scientific groups.

So, to improve the usage efficiency of all the resources, to provide a uniform way to store, access and process data, the DIRAC system was installed and evaluated.

4 JINR DIRAC Installation

The DIRAC installation in JINR consists of 4 virtual machines. Three of them placed on a dedicated server to avoid network and disk I/O interference with other virtual machines. The operating system on these virtual machines is CentOS 7. It appeared that some of the LCG software related to grid job submission is not compatible with CentOS. To cope with that, we created a new virtual machine with Scientific Linux 6 installed there. Flexibility of the DIRAC modular architecture allowed us to do so. The characteristics of the virtual machines hosting DIRAC services are presented in Table 1.

Table 7. Virtual machines hosting DIRAC services

	dirac-services	dirac-conf	dirac-web	dirac-sl6
OS	CentOS	CentOS	CentOS	Scientific Linux
Version	7.5	7.5	7.5	6.10
Cores	8	4	4	2
RAM	16 GB	8 GB	8 GB	2 GB

4.1 Use Cases in JINR

Up to now, we foresee two big possible use cases: Monte-Carlo generation for Multi Purpose Detector (MPD) at NICA and data reconstruction for Baryonic Matter at Nuclotron(BM@N).

Raw data was received by the BM@N detector and uploaded to the EOS storage. There are two data taking runs available now: run 6 and run 7. The data sizes are respectively: 16 TB and 196 TB. All data consists of files, for run 6 it is roughly 800 files and for run 7 it is around 2200 files. The main difficulty with these files is the fact that their sizes are very different: from several MBs up to 800 GBs per one file. This makes data processing a tough task especially on resources without small amount of local storage or bad network connection. The data could be processed using the Govorun Supercomputer, but the EOS is currently not connected to the storage. And the data may require full reprocessing one day, if the reconstruction algorithms will be changed.

So far, the best would be to process big files in the cloud, other files in the grid infrastructure and sometimes, when the supercomputer has free job slots, do some processing there. But without some central Workload Management system and Data Management system this is a difficult task. The data could be placed not only in EOS but also in dCache. This would allow data delivery to the worker nodes using grid protocols like SRM or xRootD. Once the X509 certificates start working for the EOS storage, it will also be included in the infrastructure and be accessible from everywhere.

The second use case is Monte-Carlo generation for the MPD experiment. Monte-Carlo generation could be performed almost on all the components of MICC at JINR. It is a CPU intensive task less demanding in terms of disk size and input/output rates. The file size could be tuned to be in a particular range for the convenience of the future use. The use of a central distributed computing system may not be critical right now, but it will definitely be useful later, when the real data arrive. It would allow for design and testing of the production workflows, and allow different organizations to participate in the experiment.

5 Conclusion

Joint Institute for Nuclear Research is a large organization with several big computing and storage subsystems. Most of the time they are used by particular scientific groups and there is no simple way to reorganize the load throughout the whole computing center. But with the new big tasks and with the improvement in technologies it became easier to integrate computing resources and use them as a single meta-computer. This leads to improvements in terms of efficiency of usage of the computing infrastructures.

The DIRAC Interware is a good example of a product for building distributed computing systems. It covers most of the needs in workload and storage management. Putting DIRAC services into operation at JINR allowed organization of data processing not in terms of tasks, but in terms of workflows. It also provides tools for removing barriers between the heterogeneous computing and storage resources.

DIRAC services were installed at JINR in order to integrate resources used by big experiments like MPD and BM@N. The initial tests and measurements demonstrated the possibility to use it for data reconstruction and Monte-Carlo generation on all the resources: JINR grid cluster, Computing Cloud and Govorun supercomputer.

References

1. Bozzi, C. and Roiser, S.: The LHCb software and computing upgrade for Run 3: opportunities and challenges, 2017 J. Phys.: Conf. Ser. **898** 112002; doi: doi:10.1088/1742-6596/898/10/112002
2. SKA telescope. https://www.skatelescope.org/software-and-computing/, last accessed 2019/08/19
3. Kalinichenko, L. et al.: Data access challenges for data intensive research in Russia, Informatics and Applications **10** (1), 2–22 (2016); doi: 10.14357/19922264160101
4. France Grilles. http://www.france-grilles.fr, last accessed 2019/08/19
5. Britton, D. et al.: GridPP: the UK grid for particle physics, Phil. Trans. R. Soc. A **367**, 2447–2457 (2009).
6. European Open Science Cloud. https://www.eosc-portal.eu, last accessed 2019/08/19
7. Gergel, V., Korenkov, V., Pelevanyuk, I., Sapunov, M., Tsaregorodtsev, A., and Zrelov, P.: Hybrid Distributed Computing Service Based on the DIRAC Interware, Communications in Computer and Information Science **706**, 105–118 (2017). doi: https://doi.org/10.1007/978-3-319-57135-5_8
8. Baginyan, A. et al.: The CMS Tier1 at JINR: five years of operations, Proceedings of VIII International Conference “Distributed Computing and Grid-technologies in Science and Education” **2267**, 1–10 (2018).
9. Baranov, A. et al: New features of the JINR cloud, Proceedings of VIII International Conference “Distributed Computing and Grid-technologies in Science and Education” **2267**, 257–261 (2018).
10. Adam, Gh. et al.: IT-ecosystem of the HybriLIT heterogeneous platform for high-performance computing and training of IT-specialists, Proceedings of VIII International Conference “Distributed Computing and Grid-technologies in Science and Education” **2267**, 638–644 (2018).
11. Peters, A.J. et al.: EOS as the present and future solution for data storage at CERN 2015, J. Phys.: Conf. Ser. **664** 042042, doi: doi:10.1088/1742-6596/664/4/042042
12. dCache, the Overview, https://www.dcache.org/manuals/dcache-whitepaper-light.pdf, last accessed 2019/08/19

Stages of Design of Digital Twin for Local Enterprise

Regina Starodubtseva[0000-0003-4414-368X], Vlada Kugurakova[0000-0002-1552-4910], and Darina Vorobyeva[0000-0001-9997-2907]

Kazan (Volga region) Federal University, Kremlevskaya 18, Kazan, Russia
vlada.kugurakova@gmail.com

Abstract. The work is devoted to the consideration of the stages of development of a digital twin for the local enterprise. With the aid of the digital twin it is possible to manage the enterprise, control problems in it, ensure the normal functioning of the enterprise objects due to the monitoring of a status of facilities and equipment based on big data obtained from Industrial IoT-sensors, as well as the monitoring of performance and current status of employees on the basis of the interpretation of bio-signals obtained in real time. The plant for preparation of super-viscous oil with disposal facilities was chosen as a local enterprise. This data is collected for the unit of predictive analysis and prediction in real time of pre-emergency situations for the rapid taking action to prevent these threats. In order to develop a plausible prototype of the digital twin of the enterprise, Unity cross-platform development environment for computer games was used.

Keywords: Digital Twin, Virtual Reality, Augmented Reality, Unity Real-Time Development Platform

1 Introduction

Digital systems for personal use have long moved from dry data to the more natural form of the information delivery to a user. Production systems tend to the same and, it seems, due to technological progress in the field of automation finally step over the abyss, after which the control systems become inherently more complex, but at the same time easier for direct human use. The improvement of such systems will require joint efforts of specialists of mathematical, computer, and, of course, psychological sciences, but will provide a huge increase in the efficiency of enterprises, as well as will contribute to the mental comfort of people working directly with the system.

Digital economy and the related terms such as digital twin, digital copy, digital double, digital transformation, digitalization are already firmly established in the terminology of both IT-specialists and economists, but there are different interpretations of them. “A Digital Twin is an integrated multiphysics, multiscale, probabilistic simulation of an as-built vehicle or system that uses the best available physical models, sensor updates, fleet history, etc., to mirror the life of its corresponding flying twin” [1]. According to this definition, in our understanding digital twin is an interactive 3D representation using user-friendly visual interfaces (clear and needless for additional interpretation) of the mathematical model of the technological object, updated in real time on the basis of information from the array of sensors. According to the vice president of GE Software Research Colin J. Parris, “Digital twin is a living model that drives a business outcome” [2].

Besides the remote comfortable work of top management with the current state of the enterprise, for the work of middle managers and their subordinates, workforce in the production itself, it is necessary to develop applications that allow to get the same interactive data in augmented reality format. Thus, the aim of the research is the development of the most effective and realistic simulation of the work of the local enterprise in real time, simulation of the sensors that monitor the condition of facilities and equipment, as well as recording of the status of employees; the creation of functionality of simple and intuitive user interface (UI/UX), data generation on the basis of which visualization of simulation of work of the enterprise will be implemented.

The remainder of this paper is organized as follows. In Section 2, there is a summary of the literature on digital twin research. Section 3 introduces the digital twin framework for the concrete enterprise, including the architecture and functional development. In Section 4, a case study shows the emergence of new problems while developing the digital twin framework. Conclusions are discussed in Section 5.

2 Related Works

The term “Twin” appeared as part of NASA's Apollo program for which two space crafts were built, one of which remained on earth to display the condition of the first one and was called “the twin” [3]. Usage of a digital twin creates new opportunities for improving business processes of companies due to the accurate and timely data. Nowadays the concept of a digital twin is considered actively as it is a strategic pathway. According to [4] a digital twin is a powerful tool which has the following advantages: Visibility, Predictive, What if Analysis Understand and explain behaviours, Connect disparate systems such as backend business applications. Described advantages provide great prospects for applying of a digital twin. However, the literature review reveals, that the development of the digital twin is still at its infancy as literature mainly include concept papers, but some concrete applied case-studies already exist.

The authors of [5] consider the concept of a digital twin in a whole, its application in product design, production scheduling, production itself and in prediction of normal and critical events. The article [6] specifically describes the application of the digital twin for modeling of the construction work of the railway station buildings for King’s Cross station. The results of this study can provide construction participants with reasonable guidance on the use of the digital twin in railway station projects for the planning, design and operation of an economical, efficient and environmentally friendly construction project. The work [7] describes some possible cases of usage of a digital twin for the wiring harness for Mercedes-Benz Cars. Digital twin allows to find vehicle faults caused by damage of the wiring harness, that consequently makes it possible to solve the problem quickly and cost-effectively. As for the petrochemical industry, the research [8] represents approaches to time series data processing as well as frequency unification, time lag issues, and the demand for immediacy for modelling of a digital twin.

The paper [9] concerns the use of sensors for Industrial Internet of Things and represents a SmartJacket. This solution is a jacket that carries data collecting sensors and safety elements such as RGB LED sleeve strips. Upon the occurrence of a specific event RGB LED sleeve stripe of the corresponding color illuminates, alerting the employee about the emergency situation.

An essential part of a digital twin is an ability to forecast some emergency situations in an enterprise. According to [10], a digital twin propose a solution of diagnostics (static and dynamic), optimization and prediction tasks due to the following components: the initial digital model, describing the processes and relations between some of control parameters and being optimized parameters; the knowledge base, filled with initial information about desired KPI; the database, able to store current data from control object; the ability to connect to DCS / PSS / ICS to send them a control signals; the execution environment, which can run digital models; the chosen algorithms of system identification and optimization. The paper [11] provides the model of emergency situations and accidents prediction for reduce the frequency of their occurrence at building sites. The authors analyzed data obtained from environmental sensors and generated association rules that represent the relationship between the accident types and causes. For this reason, preprocessing, association rule generation, and visualization are executed step by step. An experimental implementation using open-source R is conducted for demonstration of the accident prediction model. The investigation [12] provides an architecture framework to implement the cyber-physical production system, that monitors quality in metal-casting processes and predicts the occurrence of factory situations in real time based on the technologies such as the IoT, big data, and simulations. Moreover, such system acts as a coordinator to form optimal decisions related to the re-creation of production schedules.

In the current article we represent our own vision of a digital twin on the example of the development of the plant for preparation of super-viscous oil with disposal facilities.

3 Our Approaches

One of the small local enterprises in the Republic of Tatarstan was chosen as a preliminary site for the implementation of digital transformation of the enterprise production process management. The plant for preparation of super-viscous oil with disposal facilities “Kamenka” [13] was chosen as it provides not only the relative simplicity of a technical device, but also the ability to simulate all technological processes according to data obtained from IoT-sensors. A comfortable realistic three-dimensional interface for monitoring the situation at the site was developed for this enterprise. The architecture of the developed system is described below.

3.1 Architecture of the System

Database. For the specific solution database will have four entities (tables) such as User, Requirement, Notification, Task. More detailed database architecture is represented in the Figure (see Fig. 1).

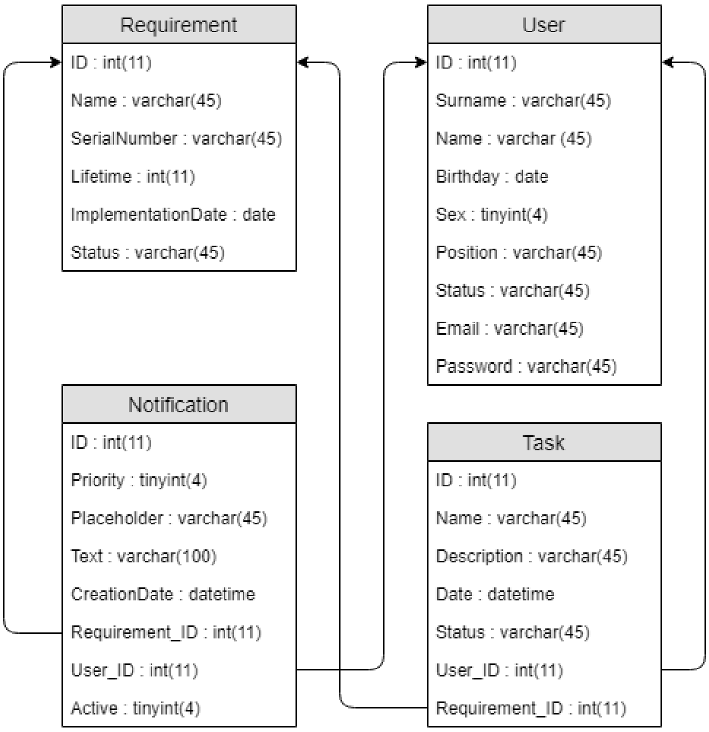


Fig. 1. Database architecture

The table *User* contains information about each member of staff, personal information, including username and password for authentication in the system. It is assumed that each employee will have access to the system through the application on a mobile device, which the employee will carry with him throughout the time spent in the enterprise, in other words, the entire working day. He will receive notifications (*Notification*) via this application. Notifications are simple messages from the administrator or other employees, or tasks (*Task*) that need to be executed (prevent or fix machinery breakdown, or take the place of another co-worker). The table *User* also includes such information as surname, name, date of birth, sex (male/female), job role, work status (busy/free), email and password.

The table *Requirement* contains all information about equipment of the enterprise including tank vessels, tubes, collectors, etc. It is assumed that each equipment introduced into the enterprise must be registered in the system, which will store such information as the equipment identification plate (category), serial number, run life, installation date and status (working properly/malfunction/exited run life). If the status of the equipment is “malfunctioning” or “exited run life”, an administrator receives an alert and creates a task (*Task*) to repair or remove the equipment from production. The task must be executed by someone of the employees.

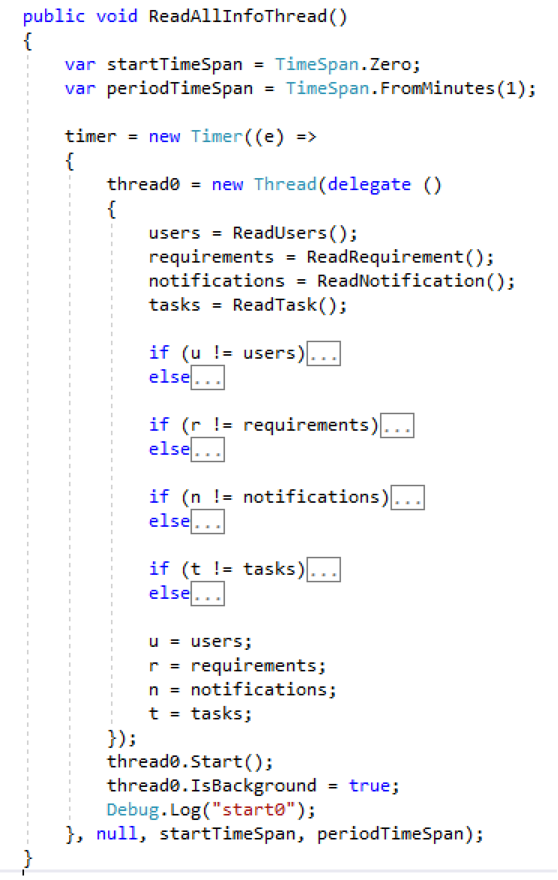
The table *Notification* contains information about equipment or employee status notifications, and it is associated with the *User* (User_ID) and *Requirement* (Requirement_ID) tables. Besides ID and Foreign Key the entity of a notification have the following attributes: priority (deviation from the norm is critical/not critical), title (by which the deviation is caused), notification text, the date of the deviation appearance and activity (the problem is fixed/not fixed).

Finally, the table *Task* contains information about tasks that are sent to employees for execution, and it is associated with the *User* (User_ID) and *Requirement* (Requirement_ID) tables. When an administrator receives automatic notifications about the status of enterprise’s facilities, he creates a task object. It is assumed that employees will choose a task from the list of tasks: “accept”, or skip the incoming task, i. e. “reject”. Thus, the entire workflow will consist of the execution of tasks that income on the mobile devices. It will allow to optimize production activities and save time. The *Task* entity will consist of attributes such as a task title, description, creation date, and status (completed/not completed).

The database architecture should contain status of all features such as Prediction entity including workload, integrator, cost part of the object for predicting the possibility of equipment replacement. The database architecture should be also extended with UserState entity, which has information about health and environment around the employee. This information includes data obtained from sensors.

Database query. All database queries are implemented in the *DbMySQLUtils* class. They act as methods, so inheriting from this class it is possible to access the database. Since some queries take a good deal of time and slow down the main stream, requests are sent to the database in separately created streams. For this reason, *System.Threading* utility is used.

All data from the database is read after a specified period of time to remove excessive load from the database. This algorithm is described in the *ReadAllInfoThread()* method (see Fig. 2). When starting the application, the system will immediately read the data, then update the data again after a specified pause, and so on. If other queries to the database (INSERT, UPDATE, DELETE) are executed, the stream, that reads all data, will stop, then the stream with a certain query will be started and then the *ReadAllInfoThread*() method with the newly created stream will be started again.



```
public void ReadAllInfoThread()
{
    var startTimeSpan = TimeSpan.Zero;
    var periodTimeSpan = TimeSpan.FromMinutes(1);

    timer = new Timer((e) =>
    {
        thread0 = new Thread(delegate ()
        {
            users = ReadUsers();
            requirements = ReadRequirement();
            notifications = ReadNotification();
            tasks = ReadTask();

            if (u != users)...
            else...

            if (r != requirements)...
            else...

            if (n != notifications)...
            else...

            if (t != tasks)...
            else...

            u = users;
            r = requirements;
            n = notifications;
            t = tasks;
        });
        thread0.Start();
        thread0.IsBackground = true;
        Debug.Log("start0");
    }, null, startTimeSpan, periodTimeSpan);
}
```

Fig. 2. Method that reads all data from the database

Movement of employees. The movement of employees will be carried out according to their tasks: the employee, having received an alert (a call from a head, automatic notification of the problem, scheduled tasks, etc.), must go to the location of the specified division.

For debugging, the procedure of the current location of the employee was simulated, which was implemented through the serialization of data into a JSON-file of several checkpoints. In the future, the movement will be implemented not via coordinates from the JSON-file, but relative to deviations from the norm of equipment and staff conditions, emergency situations, and other factors.

Indicators of equipment sensors. Serialization and deserialization of indicators of sensors of the equipment were implemented by the same principle. For tubes, tanks, collectors two indicators were selected – pressure and temperature. The condition of the equipment will be determined on the basis of these indicators. Each indicator has three zones: green, yellow and red. Each zone is an interval in which one or another indicator can be located. The green zone means that the indicator is normal, the yellow zone – a slightly deviation from the norm, the red one – a critical deviation from the norm.

Indicators of employee sensors. Similarly to serialization and deserialization of equipment sensors data of employees are recorded and read. First of all, it is pressure, pulse and temperature. Employee pressure is measured in two values: upper indicator (arterial tension) and lower indicator (venous blood pressure). Both indicators are measured in millimeter of mercury (mm Hg). The temperature is measured in standard degrees Celsius (°C). Pulse is measured in beats per minute.

However, for comprehensive workplace safety in the oil industry with the aid of sensors also it is possible to measure the electrocardiogram, record the respiratory rate and heart rate, body position and activity of the staff, as well as to supplement the system with sensors of environmental quality, for example, sensors that monitor the presence of poisonous gas in the working environment, that provides monitoring of the health of the employee. Moreover, with the aid of the alarm installed in the wearable device, the employee can be timely warned about dangerous situations or be able to warn about them by pressing the button.

3.2 Functional Development

Zoom and view from different angles. An important part of the implementation of digital twins are natural interfaces that give mental comfort to the user working directly with the system [14].

The zoom implementation for the desktop version was programmed as standard via a computer mouse Scroll Wheel. Viewing from different angles is implemented via User Interface (UI) buttons. When being clicked by a computer mouse wheel on a particular employee or equipment the point of click becomes closer.

View employee data and equipment information. It was necessary to add the ability to view employee data and equipment information with the aid of using entity-relationship model from the database (*Requirement* and *User*). The fields in which this information will be displayed are panels *Requirement InfoPanel* and *Employee InfoPanel* (see Fig. 3), located on the right side of the screen. These panels will appear as often as a user clicks the left mouse button on the corresponding digital object.

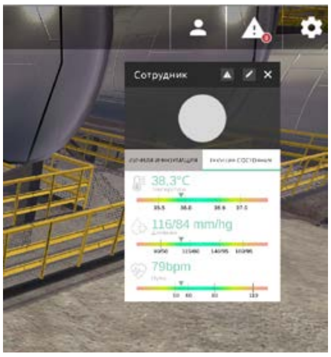


Fig. 3. Current employee data

Notifications about problems. When a problem occurs in the system (sensor indicators are in the red or yellow zone), a notification object (*Notification*) appears. Depending on the degree of criticality of each indicator, a certain algorithm determines in which zone the overall condition of the equipment or employee is located. Moreover, depending on the indicators of previous notifications, the decision is made whether to deactivate (disable) the latest notifications. Deactivated notifications are notifications included in the archive. They can be viewed in the employee or equipment card by clicking on the button with the warning icon (triangle with an exclamation mark). Notifications related to an equipment or employee can be sorted by the following categories: All messages, Accepted, Rejected, and Recent. When is being hovered by the mouse over the warning button on the main scene, a drop-down list with the same categories appears (see Fig. 4).

Label is an instant visualization of the status of an employee or equipment. In order to instantly determine the status of an employee or equipment, the label tool is developed. The label appears on the top of the object, and can be yellow if the deviation is not critical, red – in case of critical deviation (see Fig. 5).

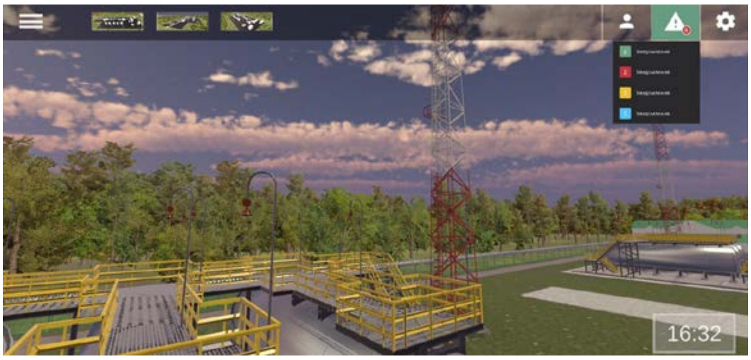


Fig. 4. Drop-down list of notifications

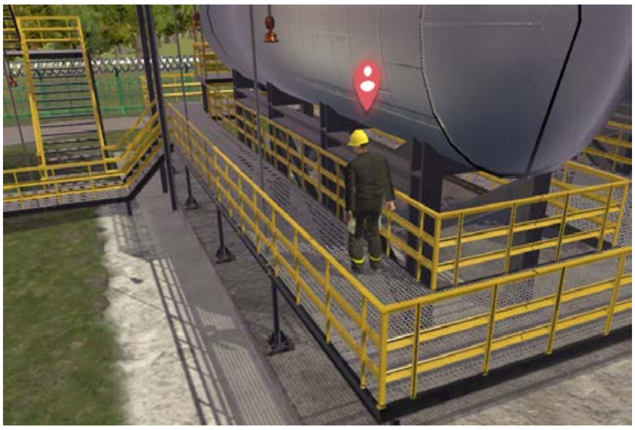


Fig. 5. Employee with a warning label

4 Emergence of New Problems

The problem of overload of connection of the application with the database occurs when there are too many requests sent to the server. In order to avoid this problem it was necessary to develop an algorithm by which all queries will be distributed in accordance with the priority given to them. Thus, a queue of queries is created, where the higher priority ones are executed at the beginning. If the amount of data in the database is too large, it is necessary to distribute it to several different servers. The next step in the implementation of the pilot application will be solving this load problem to determine the load limits for the physical architecture organization.

5 Conclusion

To sum up, digital twin is in the process of development and requires the introduction of many functions for the correct operation of the system. At this stage, first of all, it is necessary to solve the problem of overload of the communication channel with the database. In addition, it is necessary to add a display function and a corresponding sort of a notification on the stage-screen with notifications and on *Requirement InfoPanel* and *Employee InfoPanel* panels.

However, besides the current problems, many important tasks were solved: designing the application architecture, functionality for editing and displaying information about equipment and employees, creation of notifications and functional content of the scene such as scaling, camera movement, displaying UI/UX elements when pressing or hovering over certain buttons, as well as automatic display of warning labels in places of problems.

Acknowledgements

This work was funded by the subsidy of the Russian Federation Government to support the Program of competitive growth of Kazan Federal University among world-class academic centers and universities.

References

1. Glaessgen, Edward, and David Stargel: The digital twin paradigm for future NASA and US Air Force vehicles, 53rd AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics and Materials Conference 20th AIAA/ASME/AHS Adaptive Structures Conference 14th AIAA (2012).
2. Parris, C.J., Laflen, J.B., Grabb, M. L., and Kalitan, D.M.: The future for industrial services: the digital twin. Infosys Insights, 42–49 (2016).
3. Rosen, R., von Wichert G., Lo, G., and Bettenhausen, K.D.: About the importance of autonomy and digital twins for the future of manufacturing, IFAC-Papers OnLine **48** (3), 567–572 (2015).
4. Kuehn, W.: Digital twins for decision making in complex production and logistic enterprises. Int. J. of Design & Nature and Ecodynamics **13** (3), 260–271 (2018).
5. Qi, Q. and Tao, F.: Digital twin and Big Data towards smart manufacturing and industry 4.0: 360 degree comparison, in IEEE Access **6**, 3585–3593 (2018).
6. Kaewunruen, S. and Xu, N.: Digital twin for sustainability evaluation of railway station buildings, Frontiers in Built Environment **4**, art. no. 77 (2018).
7. Tharma, R., Winter, R., and Eigner, M.: An approach for the implementation of the digital twin in the automotive wiring harness field, Proceedings of International Design Conference, DESIGN **6**, 3023–3032 (2018).
8. Min, Q., Lu, Y., Liu, Z., Su, C., and Wang, Bo: machine learning based digital twin framework for production optimization in petrochemical industry, International Journal of Information Management (2019).
9. Marcon, P., Arm, J., Benesl, T., Zezulka, F., Diedrich, C., Schröder, T., Belyaev, A., Dohnal, P., Kriz, T., and Bradac, Z.: New approaches to implementing the SmartJacket into industry 4.0, Sensors (Switzerland) **19** (7), No. 1592 (2019).
10. Kostenko, D., Kudryashov, N., Maystrishin, M., Onufriev, V., Potekhin, V., and Vasiliev, A.: Digital twin applications: Diagnostics, optimisation and prediction, Annals of DAAAM and Proceedings of the International DAAAM Symposium **29** (1), 0574–0581 (2018).
11. Kwon, J.-H. and Kim, E.-J.: Accident prediction model using environmental sensors for industrial internet of things, Sens. Mater. **31** (2), 579–586 (2019).
12. Lee, J.H., Do Noh, S., Kim, H.-J., and Kang, Y.-S.: Implementation of cyber-physical production systems for quality prediction and operation control in metal casting, Sensors (Switzerland) **18** (5), No 1428 (2018).
13. Khayrullin, L.O., Kugurakova, V.V., and Starodubtseva, R.A.: Digital real time twin of enterprise entity. CEUR Workshop, in print (2019).
14. Kugurakova, V.V., Elizarov, A.M., Khafizov, M.R., Lushnikov, A.Yu., and Nizamutdinov, A.R.: Towards the immersive VR: measuring and assessing realism of user experience. In 23nd International Conference on Artificial ALife and Robotics, 146–152 (2018).

Formalization Impacts of Disasters on Enterprises and Population, Recommendations for Decision-making

E.D. Viazilov [1], A.S. Mikheev [1]

[1] RIHMI-WDC, 6, Koroleva St., 249035 Obninsk, Russia
vjaz@meteo.ru

Abstract. Information model proposed to describe situations of impacts disasters on industrial facilities and population, to make recommendations for decision-making. Methodical questions of knowledge formalization on the impacts of disasters and recommendations for preventive activity presented. The structure of the database of threshold values of disasters indicators developed. The database on impacts and recommendations has been created for more than 100 dangerous situations, for various objects and activities, in the period before disasters (based on climate and prognostic information), at the time of disasters (real time data) and after the disasters.

Key words: Disasters, Impacts, Recommendations, Formalization

1 Introduction

Disasters – strong wind, rain, extreme heat, floods, fog, waves and others – cause enormous material damage and even result in death [8]. Many losses could have been avoiding if business leaders and the public would not only receive timely information about disasters, but also to know what can happen because of their exposure to the disaster and what should be done to reduce or prevent adverse impacts. To this required automatically bring information about disasters to the decision makers of it at the initiative of the system, not the person; visualize information about dangerous situations in the form of a text description, interactive maps, the results of monitoring of hydrometeorological situations with an indication of the level of danger (yellow, orange, red), separately for each object of economy and type of activity. Impacts and recommendations for making decisions should input with assessing potential damage and calculating the cost of preventive measures. Such tools call decision support systems (DSS). The main approaches to the development of such hydrometeorological support are presenting in the articles [1, 5, 9–11].

Decision makers gain experience in dealing with natural disasters in the course of their activities. Moreover, they are confronting with certain disasters (for example, tsunami, and earthquakes) sometimes only once during the entire period of their activity, in result accumulated experience is lost. Existing experience is not always reflecting in instructions, is poorly formalized and often is presenting in a much-generalized form. This experience is stored in the memory of a person in the form of non-formalized information, skills, and abilities. Traditional forms of knowledge representation in the form of instructions, Internet pages have the following limitations: a long search time, knowledge is presenting in different sources, sometimes even a contradiction arises in them.

Despite the fact that the first studies on the creation the DSS in the field of preventing the disasters impacts began in 1990 [9], this field of research is still in the “embryonic” stage. In other branches of DSS, they are already working in permanent mode, for example, in the field of energy [12]. The most advanced solution support system currently available is Watson, developed by IBM. This system is using as a medical assistant. Using the approaches implemented in the system for our tasks may be possible, but this solution is too expensive, system operation algorithms are not clear.

If at the beginning 90-ies the main problem when creating DSS was the lack of methods and tools of creating them, then in the late 90s there was a shortage of materials with impacts and recommendations for decision-making. Currently, the Internet and other publication contains a lot of different manuals, instructions and regulations that need to be formalized and presented in the form of knowledge base.

For the development of hydrometeorological services at the modern level of IT technologies development, it is necessary to create a knowledge base in the form of formalized information on potential impacts and recommendations for decision-making, a database of threshold values for disasters indicators, software tools for identifying disasters, and searching of knowledge.

2 Information Model of Knowledge Description

The basic idea of creating a DSS is as follows. Knowing of the environment conditions is possible to determine in advance the list of impacts of natural disasters for the population and enterprises. Knowing these impacts, you can make a list of recommendations on the behavior of the population in these situations, as well as a list of recommendations to support decision-making. For the same hydrometeorological conditions at different enterprises and depending on the time of year, there may be different solutions.

Before applying the knowledge accumulated in traditional sources, a person must find and interpret them to solve a specific problem. This complicates and slows down the process of preparing a decision, at the same time in practice the use of knowledge and decision-making should be carrying out immediately after receiving the initial information in real time. In addition, in traditional forms of storing knowledge, the process of changing and supplementing new knowledge seriously hampered.

To create a DSS, formalized knowledge of scientists and specialists, registered in the literature, should be introducing in the knowledge base. An important point in the formalization of knowledge is the understanding that, depending on the time of use knowledge (before, at the time and after disaster) you must to use a particular type of information (observation, analysis, short-term or long-term forecasts, and climatic data).

To create a knowledge base for impacts and recommendations, it is necessary to define a situation description unit, select attributes, and define attribute properties. If the description unit in the selected situation determines solutions for one value of the observation, forecast or climatic indicator, is basing on three dangers levels, for 50 disasters, hundreds of typical objects, four type of activity, then the total number of cases may exceed ten thousand.

The basis of knowledge are list of rules (if, then, else). In a semi-formalized form, these rules look like this:

DISASTER: <name>; BACKGROUND: <text>; GEOGRAPHIC AREA: <list>; PERIOD (DATE): <from … <to …>>; OBJECTS OF IMPACTS: <list>; IMPACTS: <list>; RECOMMENDATIONS < list >.

Knowledge written in the form of rules has a disadvantage – with a large number of them, it becomes difficult to check their consistency. Therefore, the DSS is proposing to use the no classical knowledge bases in the form of rules, and the database of thresholds values of disasters indicators depending on the object type, type of activity on object, the location of the object, dangers level, season, climatic region. Danger levels create for the population, technical systems, building structures, ships, ports, and so on. The analysis of various situations connected with disasters and the formalization of information about these situations allows suggesting the following form for describing them:

1) Name of disaster;

2) The determination of the disaster from the meteorological dictionary;

3) The causes of the disaster (text description);

4) Photos with examples of the manifestation of disaster;

5) Impact objects (port, housing and communal economy, population);

5.1) Name of the object, which may be impact by disasters;

5.2) Information type (climate, forecast, in the moment disaster and after the disaster);

5.3) Impact indicators and their meanings;

5.4) Level of danger;

6) Impacts (name, type of activity affected by disaster, priority, author, possible potential damage);

7) Typical impacts (using for several disasters);

8) Recommendations (name, level of management to which the recommendation, priority, author is intended, cost of preventive measures of the activities, reference);

9) Typical recommendations (used for several disasters);

10) Reference to situations related to others disasters;

11) Sources of information (bibliography).

The identification of disasters is a procedure for determining the list of possible impact on object for various levels of danger. Here one can use such indicators of disasters, as threshold values, probability of disasters, risk, etc.

Disasters reasons are prerequisites for their occurrence. There, an important role is playing by environmental conditions (mountains, deserts) in which prerequisites lead to disasters (for example, heavy rains in the mountains lead to mudflows).

Each case of a dangerous situation is characterizing by the geographical area of manifestation, the duration of the risk, time of year, the climatic zone. If for a particular impact are several values of geographical and time conditions, then are several situations of dangerous impacts. The geographic area, as an element of the situation, is very important, since it often predetermines the fact of occurrence of a certain type of disasters (for example, flooding at the mouth of a river) and possible consequences of a phenomenon (for example, silt deposition on the banks of rivers from flooding). This property extends from the region “North hemisphere”, “subtropics”, “mountain area” to the level of names of settlements, transport objects (airport, seaport). There are the next types of geographic objects:

- global – “Southern Hemisphere”, “World Ocean”, “Arctic”;
- geophysical – “tropic region”, “tundra”, “lowland”;
- continent – “Europe”, “Australia”;
- ocean – “Pacific”, “Atlantic”;
- name of transport object.

The geographical region may have a specific value of is, at any level of the classification system (“pool Barents sea”, “Caspian Sea Coast”). Water areas can be open sea, bays, estuaries, port water area. To display different types of geographic areas, it is necessary, in addition to the value of the geographic region itself, to store the attribute “Type of geographic object” (countries, sea regions, mountains, lowlands).

It may note the following features associated with the identification of disasters and their manifestation. The situation may include several disasters (for example, wind, heavy precipitation), some of them are complex (for example, “storm” is “strong wind” + “waves”).

The description of the situations with disasters should contain information on the disasters indicators (air and water temperature, wind speed, height waves, speed for “strong wind”). Moreover, the consequences of the disasters impacts are determined not only by the threshold values of the indicators, but also by the conditions in which the “disasters” occurs. That is, depending on the conditions of exposure to disasters, disaster may be dangerous to one degree or another. For example, the same water level rise for settlements nearly river may be disasters, and for building on the elevation – no danger. That is, according to the sensitivity of objects to disasters, it is necessary to specify threshold values for individual objects and regions.

The situation with the disasters (for example, “flooding in the seaport”) may be associated with various variants of disasters impacts:

- one type of impact on several facilities or activities, for example, wind speed affects port cranes and loading or unloading, vessels traffic in the port;
- one type of impact – one object of impact – several consequences of disasters exposure, for example, high sea level affects the seaport (complication of mooring of vessels to the pier, moistening of cargo stored on the pier, impossibility of stay at the roadstead);
- several types of impact – one or several objects of impact and a wide variety of combinations leading to various consequences of in this situation, for example, a storm leads to flood in the river (heavy rain); tearing off roofs, trees falling down (strong wind); fires, of death to people (lightning).

The procedure for assessing impacts and issuing recommendations includes the following work:

- description of the existing situation (nature, causes of the disaster, objects of impact, activities);
- defining a list of possible objects that may be affected by natural disasters;
- assessment of natural disasters that affect the object;

- description of the sources of data used for impact assessment;
- list of all expected impacts (indicator, dangerous level);
- impact criteria (at what values of the indicator comes a negative impact);
- determining the significance of the negative impact for each disaster and the object for which the assessment is being conducted;
- loss assessment;
- identification of preventive actions;
- assessment of the cost of preventive actions;
- choice of alternative solutions;
- development of methods and tools for informing and consulting of the public about possible impacts.

3 Methodical Problems

3.1 Selection of Indicators

To create a knowledge base, it is necessary to use observable, predicted and climatic values of indicators of disasters – the probability of their occurrence; threshold levels of disasters indicators; impact lists and recommendations. The regulatory document of Roshydromet [13] determines the composition of disasters, their indicators and general threshold values. For some regions of Russia local threshold values of disaster indicators have already been introduced. This refers to the water level, wind speed, air temperature.

Spatial-temporal properties of disasters are important for assessing their level of danger. They determine the disaster scale (local, regional or global), the place of manifestation of disaster (the name of the settlement, the river), the response time to disaster, the period of possible impact (instant or gradual increase of the impact).

Important properties of disasters are their intensity, power, amplitude, magnitude, etc. Intensity meteorological processes (wind speed and height of the wave) translates them into the category disaster. For example, wind speed becomes dangerous if it is more than 15 m/s in 1/3 of a federal subject with phenomenon duration of 6 hours. Threshold values of wind speed for oceans, Arctic and Far East seas are not less than 30 m/s, and for the mountain regions – not less than 35 m/s [13]. Threshold values are establishing by regulation and are depend on their impacts on economic activities in specific geographic areas, taking into account their repeatability.

For some indicators dangerous both low and high values, e.g., pressure, humidity, temperature of air, etc. (Table 1). It should be noted that the temperature of the air, water and soil has yet another threshold value “transition through zero degrees”, which is considered as a separate disasters – frost (transition from a plus to a minus) and thaw (transition from minus to plus). Here it is important to use the forecasts and warnings of disasters [6], as well as climate risk assessments [7].

When specifying values of indicators of disasters, not only threshold values can be using, but also other types of indicators, for example:

- average (background) value – climatic value (rate) for the considered temporal and spatial resolution, for example, the medium amounts of precipitation;

- repeatability of climatic values – the number of cases (years) the manifestation of measured values of parameters in the specified intervals (wind speed is zero, 1–4, 5–9, 10–14, 15–25, >25 m/s; water level <150, 151–200, >200 cm) for a certain period, as a percentage;
- anomaly – the deviation of the parameter value from the climatic norm, which is triggering if the difference between the current and climatic indicator values is greater than the set value that is significant for air temperature, pressure, humidity.

Table 1. Thresholds of indicators

Indicators	Situations						
	Catastrophic	Dangerous	Moderately indignant	Normal	Moderately indignant	Dangerous	Catastrophic
Air temperature, ℃	<–35	–35 ÷ –25	–24 ÷ –20	–19 ÷ 20	21 ÷ 24	25 ÷ 35	>35
Air pressure, mb	<985	985–994	995–1004	1005–1015	1016–1020	1021–1030	>1030
Humidity, %	<15	15–30	31–40	41–80	81–90	91–95	>95

3.2 Determining of Thresholds Values

The most laborious and decisive step in the development of knowledge is the formation of threshold values for disaster indicators. Attributes of indicators description should include the activities type; time of year (the same indicator of disasters has different risks depending on the season of the year); geographical area (in different parts of the country, enterprises and people are differently prepared for the same disasters, for example, in areas of constant exposure to strong winds and frosts, the population has already adapted itself to surviving in such conditions). Therefore, the threshold values for every of activities type, season of the year, geographical area should be clarified.

For disasters, impacting on enterprises, refined threshold values of indicators are needed, within which it is possible to compensate for their negative impact with the help of preventive measures. The values of threshold values for specific objects and activities is basing on the existing experience of the manager. On this basis, the level of danger is the subject of an assessment of the safety of the vital activity of the population and industrial enterprises from disasters. For example, of the thresholds values are level of water, influencing the activities of the seaport or shops on the embankments in St. Petersburg; air temperature and precipitation during loading and unloading of perishable goods.

For shipping in the shallow strait requires a constant knowledge of specific values of the water level every hour and even more often. At a certain value of the water level can pass vessels with some draft. At the same time, for the construction of port facilities, on the contrary, it is necessary to know the extreme values of water level. Examples of indicators for disasters are presented in Table 2.

Table 2. Examples of disasters, which influences on different enterprises and populations

Disasters type	Name	Indicators of disasters	Objects (managing level)
Marine disasters	Waves	**Current or forecast information**: wave height>5 m. **Climate**: wave height recurrence with different wind directions.	Vessel (port authority, captain, passengers); tourism (business leaders, local government)
	Early ice cover	**Current or prognostic information**: date of occurrence of ice cover well before the average long-term date. **Climate**: max and min dates of ice cover.	Ship (port administration, ship captain)

3.3 Development of Impacts List

After determining of dangerous situations, it is necessary to evaluate possible impacts. The impacts assessment procedure includes the next steps: identification and analysis of disasters; determination of the impacts of disasters on enterprises and the population; identification of individual objects subjected to disasters with complex social or technical conditions; the identification of economic vulnerability to disasters; identification of secondary impacts from accidents and disasters at enterprises; identification of areas with a high probability manifestations impacts.

Determining the degree of impact of natural disasters on such objects as schools, hospitals, transportation plays a key role in determining the list of resources and preventive actions; identify the danger level and objects of impacts; vulnerability assessments of these objects; choice of decision criteria (loss minimization, safety of people). When assessing impacts, it is necessary to use accumulated experience in environmental impact assessment and disaster risk management [2–4, 16].

Identification of individual objects prone to natural disasters with complex social or technical conditions, also includes, for example, determining the percentage of people with low living wages, the number of elderly, children, uninsured dwellings, people without transport; potentially dangerous objects, that can lead to various accidents in the form of “a domino” effect. It is also necessary to describe the sources of data used for impact assessment; determine the significance of the negative impacts for each disaster and enterprise.

Detailed impact assessments can confuse managers and divert attention and resources from major impacts. Consequently, impacts must have priorities.

As objects of impact are used a enterprises, buildings, vessels, ports and whole industries such as maritime transport, fishing, agriculture, as well as activities for transportation of people, loading and unloading materials, transmission of energy. In the situations under consideration, objects are exposing to unintentional (spontaneous) impacts of disasters. Impacts can manifest themselves in different economy sectors and at different levels of government. Specific objects can be dividing by industry: mining and processing of gas and oil; transport; construction, etc. In each industry the division is already carried out on the basis of traditional classification, for example, in the “marine transport” emit “port facilities”, “passenger transport”.

The object of impact can be not only a material object or the branch of material activity in general, but also a functional process. Therefore, together with the system of material objects, it is necessary to take into account the system of functional impacts on production processes: generation of electric current, emission of pollutants; breeding of fish and mariculture; cleaning the air and water; ore dressing; unloading / loading materials, products; energy transfer; transportation of people, substances, materials; storage of goods.

A sufficient condition for detecting disasters is the excess of the danger level (power, speed, force or any other value of the threshold of the object's sensitivity) above the level of the object's resistance to impacts (seismic resistance of buildings, wind resistance, noise immunity of communication lines). Each specific object can be characterizing by the level of resistance in relation to the impacts of disasters. Just like the thresholds values, characteristics of the stability of objects in relation to external influences (moisture resistance, frost resistance, drought resistance, fire disasters, wind resistance, wave resistance, comfort or severity of climate) may be a properties of object. You can talk about the universal property of objects "resistance to external influences". The value of this property plays the same role as the threshold value of the disasters indicators (wind speed, number of precipitation, water level).

Impacts are considered in the context of changes in the state of objects (the condition of roads deteriorates, the availability of settlements decreases); of destruction of the objects themselves (bridges are destroyed, vessels are killed); of damage component element of objects (demolished roof, broken water pipes); occurrence a negative processes (soaking crops, which leads to the death of the crop); changing an object properties (reduced strength materials deteriorated operational characteristics of equipment), and the properties of the processes (reduces fishing, increases cruises duration).The prerequisites of impacts are called impacts conditions. They should be sought not only in the environment, but also in the places of impacts manifestation (for example, landslides often arise as a result of construction work on the hillsides). The reasons for the landslide here are the condition of the soil after heavy rainfall, and the prerequisite is the construction work. Prerequisites for the occurrence of impacts can be current and remote in time. Knowledge of the prerequisites of the occurrence of impacts may allow preventive actions to be taking in order to prevent impacts of disasters or for reduce of their impacts. Therefore, for example, knowing remote prerequisites of impacts that lead to aging or corrosion of materials of structures and reduction of their physic-mechanical characteristics allows improving these structures in period the design or increasing the number of preventive actions during operation. Design defects, materials of structures exposed to natural processes (rain, snow) that reduce the reliability of structures, may also be prerequisites of impacts.

When assessing impacts it is important to know, what information (climatic, prognostic or observed) are used an impact assessment. When forecasting the impact of disasters on the population and enterprises, the tendency of changes in the values of disasters indicators is determined. Need continuously to identify trends, since trend values indicate the possibility of increasing or decreasing disaster impacts.

Delayed consequences may be associated with several situations. For example, if one of the situations of the consequences of natural disasters "Loss of ship management" is a delayed consequence: "Shipwreck" and "Oil spill".

According to the time of exposure can be:

- long-term (with a long delay) negative impacts, which is taken into account when designing and decommissioning business facilities;

- possible impacts in the near future (related to disasters prediction), which are taken into account in the construction and operation of enterprises;
- direct impacts (with the passage of the disaster), which are taken into account in the construction and operation of facilities;
- after the passage of a natural disaster – impacts taken into account in search and rescue and rescue operations.

Depending from the time lag of impacts manifestation and the category of information used (climate, prognostic, observed), the record of impacts should reflect future, present and past impacts. Future possible distant impacts are recording in an indefinite form of the verb (violates, makes difficult, limits, excludes, creates, violates). The nearest predicted impacts are recorded with a touch of probability (may be destroyed) or in the future tense. Impacts occurring at the moment of disasters passing are recorded in the present tense. After past disasters the impacts are writing in the past tense (washed away, destroyed).

3.4 Preparation of Recommendations

For some disasters, existing forecasting methods provide insufficiently accurate results, and the user faces a dilemma: to apply or not to apply protective actions past receiving forecasting of disaster is. He has three possible strategies: never take protective actions; always take protective actions; apply protective actions selectively, focusing on intuition or additional information.

Recommendations that do not lead to substantial economic damage are giving lower priority for their implementation. It is determined what level of costs exceeds the benefits of a corresponding reduction in losses for the enterprise.

When creating a knowledge base, it should use all kinds of unstructured knowledge – textbooks, guidelines, practical guides, and even news. This information is processed and converted into formalized information. During the period of normal conditions, instructions are being developing that allow managers to prepare for disasters. The manager’s actions and responsibilities for inaction should be defining in normative documents. After the acquisition legislative powers recommendations must be strictly adhered to.

In preparing recommendations, it is necessary to take into account, both existing and develop technical regulations, the provisions on interaction [14, 15]. They developed as rule based on international and national standards adopted to protect the life and health of citizens, the property of individuals and legal entities, state and municipal property, environmental protection, animals. They must contain exhaustive lists of regulatory objects for which recommendations are establishing taking into account all categories of the population and industrial facilities, and contain rules of conduct to ensure the safety of the public and industrial enterprises in the case of disasters. With the help of the regulations we can speed up the process of filling the knowledge base. Recommendations are creating separately for different levels of enterprise management. An example of impacts and recommendations for the “Amateur fishermen on ice” situation is presented below.

Indicators: ice thickness <30 cm, wind speed >15 m/s, ice drift >0.5 m/s.

Sources: 1) Guidelines for the development of a Safety plan for water objects of the Russian Federation in the winter period. – Approved the EMERCOM of Russia. 01.07.2013 N 2-4-87-15-14.

2) Regulations on the interaction of rescue services of ministries, departments and organizations at sea and water basins of Russia. – M.: The EMERCOM of Russia. – 1995. Approved 21.06.1995.

Type of information: forecast.

Impacts for *amateur fishermen*:

Possible separation of ice from the shore.

Recommendations for *amateur fishermen*:

Do not depart from the coast further 300 m.
Leave the car on the beach.

Type of information: at the time of disaster.

Impacts for *people are on the ice*:

A crack of ice is heard.
There are rustling sounds - snow and ice fall into the cracks.
Crack width increases.
Carries an ice floe with fishermen in the sea.

Recommendations for *amateur fishermen*:

Inform the local authorities of the Ministry of Emergencies about what happened (indicate the coordinates of the place, the number of people on the ice).
To use for transition of cracks auxiliary materials – long boards, poles, logs, etc.
Count the available products. Divide the products for 2–3 days.
Prohibit the try to look for opportunities to reach the shore in single.
Wait for the rescue service.

4 Results

The collected materials on the impacts of disasters formalized in the form of 3,000 situations for 108 disasters, 30 typical objects, 100 types of activity, 3 dangerous levels, four situations (future climate change, forecast disaster, real-time data, after the disaster) with a total volume of >10 thousand impacts and recommendations are presented in the PostgreSQL database management system. An application, created for access the database (http://test.shpirat.net/), get out formalized information about dangerous situations, as well as organize the replenishment and editing of information about impacts and recommendations.

As a result of the study, a demo version of the DSS created to transfer information about disasters to the public on mobile Internet devices with the ability to provide information on the impacts of disasters and recommendations to reduce these impacts.

For receiving recommendations, may use the following variants of implementation. The user independently obtained from any official sources (radio, TV, EMERCOM of Russia, Roshydromet) or even from an unofficial foreign source information about a possible disaster, and selects the appropriate disaster in application on base their classification and ordering in alphabetical order. In the future, the search will organizes for situations related to the type of object and activities for which recommendations are necessary; the level of dangerous, the type of data used (observed, predictive, climate), the level of making decision.

A more promising implementation is associated with the automatic detection of dangerous situations based on threshold values of disasters indicators using integrated data from the Unified State System of Information on the Situation in the World Ocean [17], http://portal.esimo.ru. The system integrates the observed operational data received via the global telecommunication system of the World Meteorological Organization; forecast data in the regular grid, coming from the Hydrometeorological Center of Russia and other forecasting organizations; climate data obtained on the basis of generalization of historical data RIHMI-WDC. At the same time, dangerous situations for each object are identifying separately based on threshold values of the indicators of disasters and are automatically delivering directly to the MeteoAgent program is running on Internet device of the decision makers. After receiving a message about disasters, the MeteoAgent program is initializing on a mobile Internet device and, if necessary, the decision-makers will receive information about the possible impacts of the disaster detected and recommendations for decisions making. At this stage, it is possible to connect economic models that allow one to assess the possible damage and calculate the cost of preventive measures before the onset of the disaster.

In more detail with the existing demonstration version of the implementation and the prospects for the development of DSS can be found in the articles [1, 5, 18].

5 Conclusions

As a result of the work done for the first time in field hydrometeorology: an information model has been developed for describing information about the impacts of disasters and recommendations for taking preventive measures; methods of formalizing information about the impacts and recommendations for decision-making are tested; an experimental database of threshold values of disaster indicators, characterizing type objects, type of activities depending on the year season, geographic area, climatic zone has been prepared; materials on the manifestation of various disasters are collected; demo variant DSS for disasters created.

Acknowledgements

The work financially supported by the Ministry of Science and Higher Education of the Russian Federation, a unique project identifier RFMEFI61618X0103.

References

1. Chunaev, N. and Viazilov, E.: About changing paradigm of hydrometeorological services for a natural disasters. Proceedings of Hydrometcentre of Russia **362**, 224–235 (2016).
2. Order of the State Committee for Ecology of the Russian Federation No. 372 dated May 16, 2000 “On Approval of the Regulation on Environmental Impact Assessment of the Planned Economic and Other Activities in the Russian Federation” (2000).

3. The role of hydrometeorological services in disaster risk management. Coordinated by the World Bank, WMO. Washington, DC Columbia. 68 p. (2012).
4. Managing Extreme and Disaster Risk to Facilitate Adaptation to Climate Change: A Summary for Policymakers. Special Report of the Intergovernmental Panel on Climate Change. WMO. UNEP. Report of Working Groups I and II of the IPCC. 32 p. (2012).
5. Viazilov, E.: Development of hydrometeorological services to support the decision of enterprises of Russian Federation. United Nations Office for Disaster Risk Reduction (UNISDR). The UN Global Assessment report on disaster risk reduction in 2019 (GAR19), (2019). https://www.preventionweb.net/publications/view/66441
6. The WMO Guidelines for the maintenance of forecasts of hazards with impacts and warnings. WMO, Commission for Basic Systems, 27 p. (2014).
7. Kobysheva, N.V., Akenteva, L.M, and Galyuk, L.P.: Climate change and variability in the technical field. SPb: “Publishing Cyrillic”, 256 p. (2015).
8. Korshunov, A.: Risks and climate to weather and their impact on the economy and population. Obninsk: RIHMI-WDC (20013). http://meteo.ru/component/content/article/9-uncategorised/271-pogodno-klimaticheskie-riski-i-ikh-vozdejstvie-na-ekonomiku-i-aselineie
9. Vyazilov, E.D.: On the use of databases of hydrometeorological data and knowledge bases when making decisions on objects of the national economy. Gidrometeoizdat, Proceedings RIHMI-WDC (151), 24–41 (1990).
10. Vyazilov, E.D., Britkov, V.B., and Bashlykov, A.A.: Creation of decision support systems in hydrometeorology. SPb.: Proceedings of RIHMI-WDC (160), 124–135 (1996).
11. Gelovani, V., Britkov, V., Bashlykov, A., and Viazilov, E.: Intellectual decision support systems in emergencies using information on the state of the environment. Moscow: Institute for Systems Analysis of RAS. URSS. 304 p. (2001).
12. Bashlykov, A.A.: Computer information systems for intellectual support of AS operators. Moscow: RIENG, 520 p. (2016).
13. RD 52.88.699. Regulations on the procedure for the actions of institutions and organizations in the event of the threat of occurrence and occurrence of disasters. – Moscow: Roshydromet, 31 p. (2008).
14. Guidelines for the development of a safety plan for water bodies of the Russian Federation for the winter. Approved by the EMERCOM of Russia 01.07.2013 N2-4-87-15-14 (2013).
15. Regulation on the interaction of emergency services of ministries, departments and organizations in the sea and water basins of Russia. – Moscow: EMERCOM of Russia (1995).
16. Federal Law of the Russian Federation dated 11.23.1995 No. 174-FZ “On Ecological Expertise” (1995).
17. Viazilov, E. and Mikhailov, N.: Data integration for marine environment and activities. Infrasructure of geoinformation Resources and integration. Proceedings of scientific papers. Editer Dr. M.A. Popov, Dr. E.B. Kudashev. Kiev. Karbon-Service, 174 –181 (2013).
18. Viazilov, E.: Development of hydrometeorological support for consumers using modern IT. Conference “CITES-2019”, 3–6 June 2019, Moscow, Russia, 231–235 (2019).

DATA INTEGRATION, ONTOLOGIES AND APPLICATIONS

Applied Ontologies for Managing Graphic Resources in Spectroscopy

N.A. Lavrentev, A.I. Privezentsev, and A.Z. Fazliev[0000-0003-2625-3156]
Institute of Atmospheric Optics SB, RAS, Tomsk 634055, Russia
lnick@iao.ru, remake@iao.ru, faz@iao.ru

Abstract The report presents the tasks on graphical resources management thoroughly describing applied ontologies of GrafOnto research graphics collection used for solving problems of spectroscopy. The problems of ontology modularity and automatic classes` generation are being discussed. Examples of solving reduction problem as well as applied ontologies metrics are presented.

Keywords Research Graphical Resources Classification, Spectroscopic Graphical Resources Ontology

1 Introduction

In the middle of 2000s the emergence of digital scientific libraries with publications as well as Semantic Web approach oriented on semantic description of information resources induced the work on decomposition of resources into smaller parts that require the creation of semantic annotations oriented on the description of domains and various data representations used in them. Various forms of data representation are always used in scientific publications (text, tables, graphics, symbols (for example, formulas), etc …). On the other side researcher got the facilities for storing and presenting large amounts of information, although published data and information was needed for the control of this information quality. Virtual data centers in various domains appeared in the second half of the 2000-th. These data centers usually contained the published data represented in publications in tabular form. In the end of 2000s publications on scientific graphical resources' systematization started to appear Ref. [1–4]. An example of an approach to creating a collection of graphical resources in High Energy Physics is presented in Ref. [5].

The report presents the results of the final stage of scientific plots' systematization in three disciplines of spectroscopy. At the first stage we formed GrafOnto collection of graphical resources [6–10] describing the results of studies on the problems of a water molecule spectral lines' continuum and on spectral properties of weakly bounded complexes and absorption cross-sections used for the photochemical reactions rates' calculation. At the second stage the typification of plots and figures as well as the first version of GrafOnto resources ontology was done (see Ref. [11–13]).

In order to upload new datasets into GrafOnto system and support them one has to solve the tasks on managing graphical resources. These are such tasks as specification of informational resources' structure for spectroscopy problems and analysis of resources' validity, control of data completeness and trust estimation. The decision support system which used in management of the collection GrafOnto is based on ontologies describing the primitive and composite plots and figures. Description of these ontologies is the aim of this report.

2 GrafOnto Collection of Scientific Graphics

The collection is based on a digital library, containing more than a thousand articles. These articles are dedicated to spectroscopy research such as spectral lines' continuum, weakly bound complexes' properties and spectral functions in near and far ultraviolet range. A distinctive feature of the above problems of spectroscopy is that the major part of published data is represented in a form of plots, figures and images.

In order to create a collection, graphical objects should be manually extracted and converted into a digital form. Software used to upload, storage, view, search and integrate graphical resources into collection is original. At present, the collection contains about 3000 primitive plots included into 625 composite plots and 104 composite figures as well as about 4000 primitive plots ready for the upload. The uploaded plots describe properties of 19 molecules, 25 complexes and 50 mixtures. Almost a half of primitive plots characterize properties of a water molecule. Collections' plots are related to dozens of physical quantities (functions) and a dozen of physical quantities (arguments). Table 1 illustrates spectral lines' collections and a number of primitive plots related to these functions for substance groups. It is worth noting that, at present, only a part of the plots from the publication chosen by experts is uploaded into the collection. Other plots will be processed automatically after the software for machine processing of graphical resources is developed. The collection of plots that has already been created will be used as a data set for training a neural network aimed at automatic recognition of scientific graphics.

Table 1. Number of plots with the functions that are most frequently used in the collection

Functions	The number of the primitive plots			
	Mixture	Molecule	Complex	Total
Absorption Coefficient (cm^{-1})	20	103	17	140
Absorption Coefficient (cm^{-1}atm^{-1})	14	38		52
Absorption Coefficient (cm^{2}mol^{-1}atm^{-1})	95	567	27	689
Absorption Coefficient (dB/km)	14	128	11	153

3 Managing Graphical Resources in GrafOnto Collection

The principal tasks of graphical resources management are to control resources structure and data quality. An ontology knowledge base accumulating all computer-generated information on collection components is used for making decisions during the management.

Resources structure contains plots of various types, their description, substances, functions and their arguments, physical quantities' units, units table as well as coordinate systems and level of detail of their description, etc. Control of plots and figures validity is based on the analysis of calculated values of paired relations between cited plots and original plots related to them. Such a relation is characterized by a reference to publication, figure number and an identifier of a curve. Note that, at present, the collection of cited plots contains 693 primitive and 248 composite plots. The ontology describing the present state of the collection resources is presented below.

4 Applied Ontologies of Scientific Plots and Figures in Spectroscopy

Taxonomy of some of the most important artifacts of research publishing [5] includes concepts: figure (composite figure, plot (exclusion area plot, GenericFunctionPlot, histogram), diagram, picture. In our work we defined additional concepts characterized by the methods of acquiring physical quantities (FTP, Cell, etc …) as well as their types (Theoretical, Experimental, Fitting, Asymptotic), slang names of physical quantities, etc. and declared them as subclasses o GenericFunctionPlot class. These definitions are oriented on physical quantities used as plots' axes.

We defined the following hierarchy for forming ontologies in spectroscopy domain. Basic ontology of spectroscopy graphical resources contains three parts and each part is related to one of the three problems of spectroscopy. These problems are the following: problems of continuum absorption, weakly bound complexes as well as the specific task of spectral functions related to photochemical reactions in the atmosphere.

4.1 Basic Ontology and Applied Ontologies of Domain Problems

Basic ontology contains some classes and properties, which are used in applied ontologies of domain problems. In our case, these problems are weakly related to each other and are represented by the following independent modules: graphical resources of continuum absorption, weakly bound complexes and absorption profiles, defining rate of photochemical reaction. Each of these modules is split into three parts: the first part characterizes coordinate systems used in GrafOnto collection, the second one characterizes physical quantities, while the third one characterizes the substances, the properties of which are presented in the collection.

4.2 Main Classes

In ontologies classes define many resources presented in our work in a form of plots and figures from GrafOnto collection as well as in a form of description of their properties. All the classes are explicitly defined in OWL 2 syntax with the use of Manchester syntax for their definition.

Basic ontology classes

In the framework of the chosen model the main entities in spectroscopy are substances (**Substance** class) – molecules as well as complexes and mixtures, and methods of acquiring physical quantities' values (**Method**). Graphical representation is related to graphical system entity (**GraphicalSystem**). The components of a graphical system are, for example, the coordinate axes of plots representing physical entities. In GrafOnto each published plot or figure is related to the description of its properties (**Description**, **ResearchPlotDescription** classes). One of such properties is a bibliographic reference to a publication (**Reference**). The **Problem** class contains three individuals (Continuum, Complex and CrossSection) each identifying a problem related to a graphical resource.

Classes related to domain problems' ontologies

Domain problems are closely related to the tasks for their solution. GrafOnto collection contains graphical resources related to the problems mentioned in the introductory abstract of this paragraph. Classes of spectroscopy problems' ontologies contain numerous resources and their description. **PhysicalQuantity** class consists of two non-adjacent subclasses named **SystemPhysicalQuantityDepended** and **SystemPhysicalQuantityIndepended**. The first class contains physical quantities the dependency of which on other physical quantities is presented in plots and figures, while the second one contains physical quantities the dependency of which is presented in plots and figures.

In order to understand the names of ontology classes we have to describe the etymology first. A name may consist of several words. These words correspond to the names of individuals in the corresponding classes **MethodType**, **SystemPhysicalQuantityDepended** and **Substance**. Fig. 1 presents examples of schemes for creating subclasses names in A classes (**Physical quantity and related substances**) and B classes (**Substance and related physical quantities**) presented in simplified syntax.

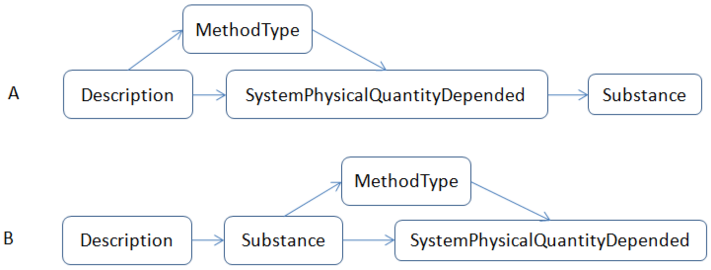


Fig. 1. Word order in the names of A and B groups' classes

For example, a class named **Description_Experimental_Absorption_Coefficient__cm2mol_1atm_1_** contains all the descriptions of measured absorption coefficients with cm^{2}molecule^{-1}atm^{-1} dimension for a series of substances being a subclass of **Physical quantity and related substances**. The third group of classes related to subclasses of **GraphicalSystem** class is not presented in this work.

4.3 Main Properties

Comments for all the properties used in natural language are presented in OWL 2 ontologies code. Here we present a simplified classification of some properties related to physical quantities and descriptions of plots and figures. Description of properties related to Description and CoordinateSystem classes as well as to Temperature and Pressure quantitative characteristics is omitted.

Table 2 lists ontology properties defining their domains and ranges. The last column of the table shows abbreviations of properties used in the scheme of individual presented in Fig. 2.

Qualitative properties characterized physical quantities are *hasOriginType*, *hasSourceType* and *hasMethodType*. The values of *hasOriginType* property indicate the origin of dataset related to the plot: it should be original and should be obtained by digitizing the curve of a primitive plot. The values of *hasSourceType* property can describe primary data, i.e. the data obtained by the authors of the publication as well as the previously published curves (i.e. cited) and commonly known curves (i.e. expert). The values of *hasMethodType* property characterize qualitative acquisition of datasets of primitive plot: Theoretical is a calculation using physical or mathematical model, Experimental is measurement, Fitting is a continuous curve creation using the method of fitting to experimental values.

The relations between plots and figures are defined by 6 properties (*has*{*OPPD, CPPD, OCPD, CCPD, MCPD*}, *hasPrototype*) . First five mereological properties describe composition of composite plots (*OPPD, CPPD*) and figures (*OCPD, CCPD, MCPD*). The value of *hasPrototype* property used in the description of cited primitive plot is the corresponding original plot. This property defines the descriptions that contain datasets with closely related values.

4.4 Main Types of Individuals

Being equivalents of figures and plots from published graphical resources on the above problems images generated in GrafOnto system are related to the description of their metadata making the most significant part of ontology individuals included in A-box. Typification of figures and plots given in Ref. [14] is defined by the property values. Abbreviation of corresponding values is used in the names of such individuals (for example, OCP – Original Composite Plot). Fig. 2 illustrates the structure of one of such plot types, i.e. original primitive plot. Ovals stand for ontology individuals, rectangles stand for literals and directed arcs stand for objective (OP) and determined (datatype – DTP) properties. Cited primitive plot have a similar structure with an addition of observations with hasPrototype, hasChild and hasParent properties. Special cases of individuals characterizing properties of coordinate system and its axes are shown in the lower part of Fig. 2. A series of individuals are related to the classes defined by enumeration of its individuals.

Table 2. Main properties of base ontology and spectroscopy problems ontology

Primitive Plot Description (PPD)			
Domain	Property	Range	Abbr
Description	hasReference	Reference	OP1
PrimitivePlotDescription	hasSubstance	Substance	OP2
PrimitivePlotDescription	hasSourceType	{Primary, Expert, Cited}	OP3
PrimitivePlotDescription	hasOriginType	{Digitized, Original}	OP4
PrimitivePlotDescription	hasCurveType	{Line, Point}	OP5
PrimitivePlotDescription	hasCS	CoordinateSystem	OP6
CitedPrimitivePlotDescription	hasCitedReference	Reference	OP7
CoordinateSystem	hasCSType	{2D-Decartes}	OP8
CoordinateSystem	hasX-axis	X-axis	OP9
CoordinateSystem	hasY-axis	Y-axis	OP10
Y-axis	hasMethod	Method	OP11
Y-axis	hasMethodType	{Theory, Experiment, Fitting}	OP12
Y-axis	hasPY-axis	PubPhysQuanDepended	OP13
Y-axis	hasSY-axis	SysPhysQuanDepended	OP14
X-axis or Y-axis	hasAxisScale	{Linear, Logarithmic}	OP15
X-axis	hasPX-axis	PubPhysQuanIndep	OP16
X-axis	hasSX-axis	SysPhysQuanIndep	OP17
CitedPrimitivePlotDescription	hasPrototype	OriginalPrimitive-PlotDescription	OP18
PrimitivePlotDescription	hasTemperature	float	DT1
PrimitivePlotDescription	hasPressure	float	DT2
PrimitivePlotDescription	hasSystemFigureNumber	integer	DT3
ResearchFigureDescription	hasOriginalImageOfPlot	URI	DT4
ResearchFigureDescription	hasOriginalPlotInformation	URI	DT5
FigureDescription	hasFigureCaption	string	DT6
FigureDescription	isPartOfFigureNumber	integer	DT7
ResearchFigureDescription	hasNumberOf Points	integer	DT9
ResearchFigureDescription	hasPlotCaption	string	DT12
Original Composite Plot Description (OCPD)			
OriginalCompositePlotDescription	hasOPPD	OriginalPrimitivePlotDescription	Op19
Cited Composite Plot Description (CCPD)			
CitedCompositePlotDescription	hasCPPD	PrimitivePlotDescription	Op20
Composite Figure Description (CFD)			
CompositeFigureDescription	hasOCPD	OriginalCompositePlotDescription	Op21
CompositeFigureDescription	hasCCPD	CitedCompositePlotDescription	Op22
CompositeFigureDescription	hasMCPD	MultipaperCompositePlotDescription	Op23

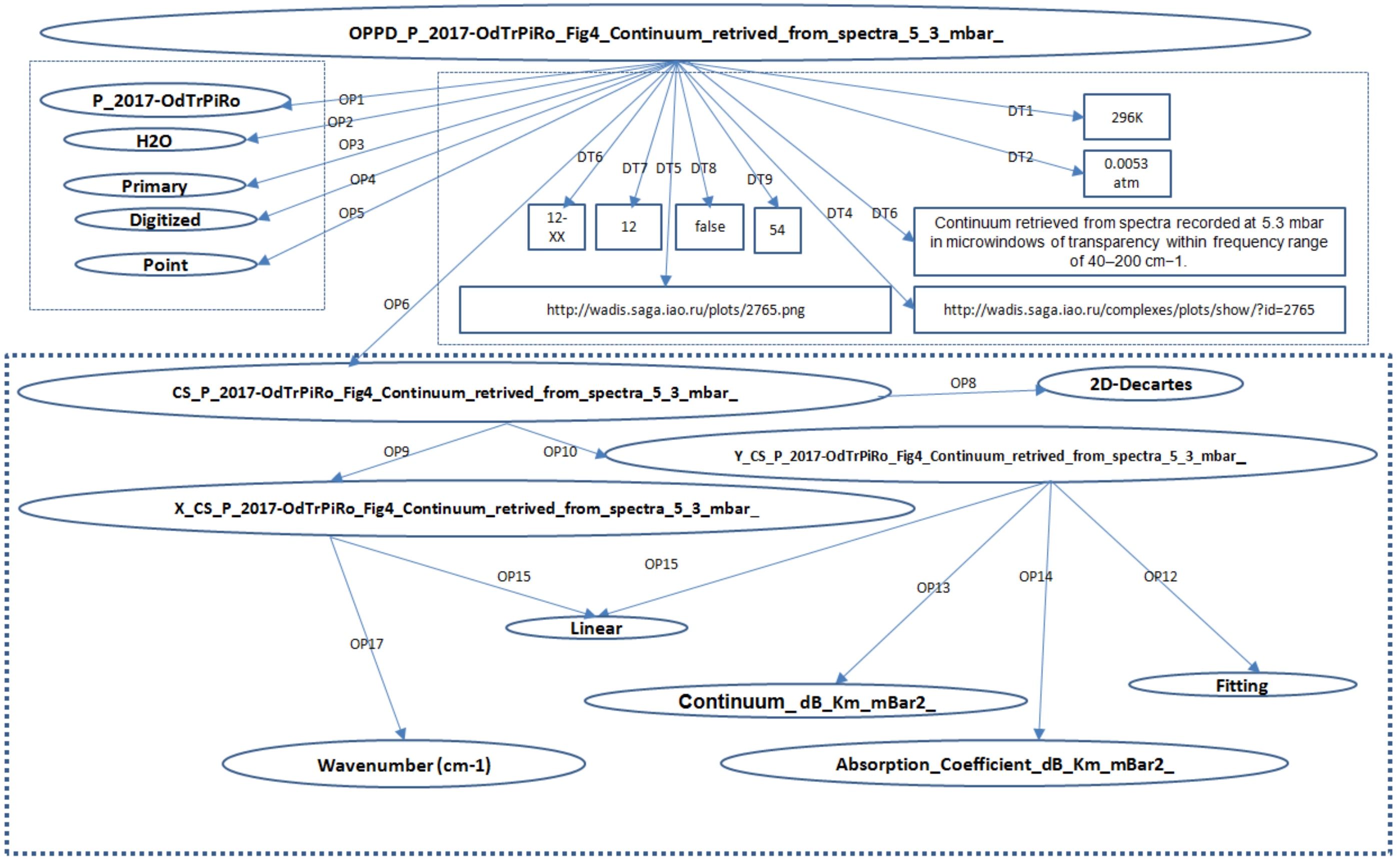

OPPD_P_2017-OdTrPiRo_Fig4_Continuum_retrived_from_spectra_5_3_mbar_
P_2017-OdTrPiRo
H2O
Primary
Digitized
Point
OP1
OP2
OP3
OP4
OP5
OP6
DT6
DT7
DT5
DT8
DT9
DT4
DT6
DT1
DT2
296K
0.0053 atm
12-XX
12
false
54
Continuum retrieved from spectra recorded at 5.3 mbar in microwindows of transparency within frequency range of 40–200 cm−1.
http://wadis.saga.iao.ru/plots/2765.png
http://wadis.saga.iao.ru/complexes/plots/show/?id=2765
CS_P_2017-OdTrPiRo_Fig4_Continuum_retrived_from_spectra_5_3_mbar_
OP8
2D-Decartes
OP9
OP10
Y_CS_P_2017-OdTrPiRo_Fig4_Continuum_retrived_from_spectra_5_3_mbar_
X_CS_P_2017-OdTrPiRo_Fig4_Continuum_retrived_from_spectra_5_3_mbar_
OP15
OP15
OP13
OP14
OP12
OP17
Linear
Continuum_dB_Km_mBar2_
Fitting
Wavenumber (cm-1)
Absorption_Coefficient_dB_Km_mBar2_

Fig. 2. Structure of individual characterizing original primitive plot from Fig. 4 in Ref [14]

4.5 Ontologies Metrics

Ontologies metrics are used for comparing ontologies of different parts of a domain or of different domains, characterizing quantitative and qualitative peculiarities of ontological description. In OWL ontologies the number of object properties characterizes the number of paired relations between individuals. Some of these individuals may have quantitative estimation. The estimated relations are described by certain (datatype) properties.

Table 3 contains metrics for applied ontologies of three spectroscopy problems as well as the unification of these ontologies (Σ Ontology). As for GrafOnto resources collection the equality of numbers characterizing the number of properties, their domains and ranges means that they are characterized by identical properties. However, the difference in classes` numbers indicates the use of a greater number of spectral functions in Continuum problem in comparison with Complex and Cross Section problems. As individuals of one and the same group of types are used applied ontologies we may conclude that ontology on Continuum problem describe the highest number of primitive plots.

Table 3. Ontology metrics, characterizing the spectroscopy graphical resources collection

	Continuum	Complex	Cross Section	Σ Ontology
Metrics				
Axiom	53279	12737	10704	84662
Logical axiom count	43180	10171	8634	68587
Declaration axiom count	8187	2134	1715	13135
Class count	295	205	55	552
Object properties count	24	24	24	24
Datatype properties count	14	14	14	14
Individual count	7910	1984	1641	12578
DL expressivity	ALCO(D)	ALCO(D)	ALCO(D)	ALCO(D)
Object property axioms				
Object properties domains	24	24	24	24
Object properties ranges	24	24	24	24
Datatype property axioms				
Datatype properties domains	14	14	14	14
Datatype properties ranges	14	14	14	14
Individual axioms				
Class assertion	426	204	206	854
Object properties assertion	27971	6231	5590	44152
Datatype properties assertion	14204	3314	2683	22516
Annotation axioms				
Annotation assertions	1912	432	355	2940

Metrics comparison of Σ Ontology with ontologies of tabular information resources Ref. [14] reveals that in our work on graphical resources ontology we managed to significantly increase the number of classes in one year. It clearly indicates that Σ Ontology contains the highest number of obvious answers on typical user requests.

5 Conclusion

The report presents applied ontologies of scientific plots and figures used for managing graphical resources in three problems of spectroscopy. Ontologies describe a collection of plots and figures published in the period from 1918 till 2018. Ontologies are created for managing structure of collection resources as well as for making decisions on such tasks as development, storage and systematization of plots and figures for solving such problems as continuum absorption and research of properties of weakly related complexes and cross sections absorption. Ontology as well as its individuals and classes are automatically generated with the enlargement of the collection.

The future of GrafOnto collection is related to automatic recognition of plots and figures used in spectroscopy as well as to the generation of applied ontologies characterizing validity analysis and confidence estimation of its resources.

References

1. Halpin, H. and Presutti, V.: An ontology of resources for linked data. Linked Data on the Web 2009, Madrid, Spain. ACM 978-1-60558-487-4/09/04
2. Thorsen, H. and Pattuelli, C.M.: Ontologies in the time of linked data. In Smiraglia, Richard P., ed. Proceedings from North American Symposium on Knowledge Organization **5**, 1–15 (2015).
3. Niknam, M. and Kemke, C., Modeling shapes and graphics concepts in an ontology. https://pdfs.semanticscholar.org/c20b/3b819ce253715bbfa9c2151a10ea87f718e4.pdf
4. Kalogerakis, E., Christodoulakis, S., and Moumoutzis, N.: coupling ontologies with graphics content for knowledge driven visualization. https://people.cs.umass.edu/~kalo/papers/graphicsOntologies/graphicsOntologies.pdf
5. Praczyk, P.A.: Management of scientific images: an approach to the extraction, annotation and retrieval of figures in the field of High Energy Physics. Thesis Doctoral, Universidad de Zaragoza (2013). ISSN 2254-7606
6. Voronina, Yu.V., Lavrentiev, N.A., Privezentzev, A.I., and Fazliev, A.Z: Collection of published plots on water vapor absorption cross sections. Proc. SPIE **10833** (2018). doi: 10.1117/12.2504586s
7. Lavrentiev, N.A., Rodimova, O.B., Fazliev, A.Z., and Vigasin, A.A.: Systematization of published research plots in spectroscopy of weakly bounded complexes of molecular oxygen and nitrogen. Proc. SPIE **10833** (2018). doi: 10.1117/12.2504327
8. Lavrentiev, N.A., Rodimova, O.B., and Fazliev, A.Z.: Systematization of published scientific graphics characterizing the water vapor continuum absorption: I. Publications of 1898–1980. Proc. SPIE **10833** (2018). doi: 10.1117/12.2504325

9. Lavrentiev, N.A., Rodimova, O.B., Fazliev, A.Z., and Vigasin A.A.: Systematization of published research graphics characterizing weakly bound molecular complexes with carbon dioxide. Proc. SPIE 104660E (2017). doi: 10.1117/12.2289932
10. Lavrentiev, N.A., Rodimova, O.B., and Fazliev, A.Z.: Systematization of graphically plotted published spectral functions of weakly bound water complexes. Proc. SPIE **10035** (2016). doi: 10.1117/12.2249159
11. Lavrentiev, N.A., Privezentsev, A.I., and Fazliev, A.Z.: Tabular and Graphic Resources in Quantitative Spectroscopy. In: L. Kalinichenko et al. (eds.) DAMDID/RCDL 2018, CCIS
12. Lavrentiev, N.A., Privezentsev, A.I., and Fazliev, A.Z.: Systematization of Tabular and Graphical Resources in Quantitative Spectroscopy. CEUR Workshop Proceedings, Selected Papers of the XX International Conference on Data Analytics and Management in Data Intensive Domains. Edited by Leonid Kalinichenko, Yannis Manolopoulos, Sergey Stupnikov, Nikolay Skvortsov, Vladimir Sukhomlin **2277**, 25–32 (2018).
13. Lavrentiev, N.A., Privezentsev, A.I., and Fazliev, A.Z.: Applied Ontology of Molecule Spectroscopy Scientific Plots. Proc. of Conference "Knowledge, Ontologies, Theories", DigitPro **2**, 36–40 (2017).
14. Odintsova, T.A., Tretyakov, M.Yu., Pirali, O., and Roy, P.: Water vapor continuum in the range of rotational spectrum of H_2O molecule: New experimental data and their comparative analysis. Journal of Quantitative Spectroscopy and Radiative Transfer **187**, 116–123 (2017). doi: 10.1016/j.jqsrt.2016.09.00

Ontology-based Data Integration

Manuk Manukyan

Yerevan State University, Yerevan 0025, Armenia,
mgm@ysu.am

Abstract. The data integration concept formalization issues have been considered within an XML-oriented data model. An ontology for data integration concept is proposed. Three kinds mechanisms are used to formalize the data integration concept: content dictionary, signature file and reasoning file (collections of reasoning rules). The reasoning rules are based on an algebra of integrable data and formalized by an XML DTD. The data translation mechanisms are non-sensitive to extension of the considered algebra. It is important that the considered data model is extensible and we use a computationally complete language to support the data integration concept.

Keywords: Data Integration, Data Warehouse, Mediator, Data Cube, Ontological Modeling, XML, OPENMath.

1 Introduction

We have published a number of papers that are devoted to investigation of data integration problems (for instance, see [12, 13, 15, 16]). Within of these works an approach to virtual and materialized integration of data has been developed. In [12] we considered the existence issues of reversible mapping of an arbitrary source data model into a target data model. The considered approach in [12] is based on the method of commutative mapping of data models of L. A. Kalinichenko [10]. According to this method, each data model is defined by syntax and semantics of two languages, data definition language (DDL) and data manipulation language (DML). The main principle of mapping of an arbitrary resource data model into the target one could be reached under the condition, that the diagram of DDL (schemas) mapping and the diagram of DML (operators) mapping are commutative. A new dynamic indexing structure for multidimensional data has been developed in [15] to support data materialized integration. The problems to support OLAP-queries are considered in [13, 16].

In this paper we will consider an approach to ontology-based data integration. An ontology is a formal, explicit specification of a conceptualization of a shared knowledge domain. In other words, ontologies offer means to represent high level concepts, their properties, and their interrelationships. Such representations are used for reasoning about entities of the subject domains, as well as for the domains description. In the frame of our approach to ontology-based data integration we have developed an XML-oriented data model by strengthening the

XML data model by means of the OPENMath concept [5]. OPENMath is a standard to represent mathematical concepts with their semantics on the Web. Usage of OPENMath concept allows to extend the XML language with computational and ontological constructs. We have certain experience in OPENMath usage in our research in this context (for instance see [14]. The proposed ontology is based on the OPENMath formalism and the so-called algebra of integrated data which also has been developed by us. Three kinds of mechanisms of the OPENMath are used to formalize the data integration concept: content dictionary, signature file and reasoning file (collections of reasoning rules). The reasoning rules are based on the algebra of integrable data and formalized by an XML DTD. It is essential that the considered data model is extensible and we use a computationally complete language to support the data integration concept.

The paper is organized as follows: the formal bases of the data integration concept formalization are considered briefly in Section 2. An algebra of integrable data and an ontology for data integration concept are proposed in Section 3 and Section 4 correspondingly. Related work is presented in Section 5. The conclusion is provided in Section 6.

2 Formal Bases

In this section we will briefly consider the OPENMath concept. Namely, the formalism and the constructions on which this concept is based. OPENMath is an extensible formalism and we use it to formalize the ontology-based data integration concept. This Section is based on the following works [13, 14].

2.1 The OPENMath Concept

OpenMath is a standard for representation of the mathematical objects, allowing them to be exchanged between computer programs, stored in databases, or published on the Web. The considered formalism is oriented to represent semantic information and is not intended to be used directly for presentation. Any mathematical concept or fact is an example of mathematical object. OpenMath objects are such representation of mathematical objects which assumes an XML interpretation.

Formally, an OpenMath object is a labeled tree whose leaves are basic OpenMath objects. The compound objects are defined in terms of *binding* and *application* of the λ-calculus [9]. The type system is built on the basis of types that are defined by themselves and certain recursive rules, whereby the compound types are built from simpler types. The basis consists of the conventional atomic types (for example, *integer*, *string*, *boolean*, etc.). To build compound types the following type constructors are used:

- *Attribution.* If v is a basic object variable and t is a typed object, then **attribution**(v, *type* t) is typed object. It denotes a variable with type t.

- *Abstraction.* If v is a basic object variable and t, A are typed objects, then **binding**(*lambda*, **attribution**(v, *type* t), A) is typed object.

- *Application.* If F and A are typed objects, then $\mathbf{application}(F, A)$ is typed object.

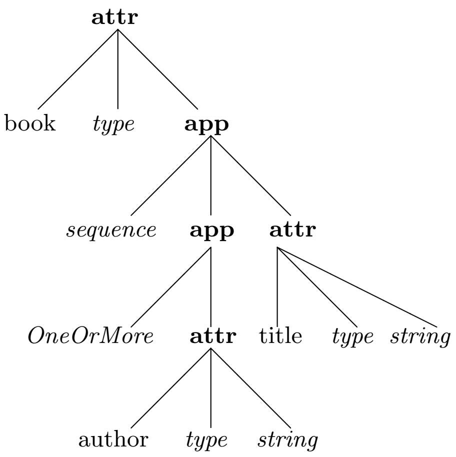


Fig. 1. An example of compound object

Semantic Level. OPENMath is implemented as an XML application. Its syntax is defined by syntactical rules of XML, its grammar is partially defined by its own DTD. Only syntactical validity of the OPENMath objects representation can be provided on the DTD level. To check semantics, in addition to general rules inherited by XML applications, the considered application defines new syntactical rules. This is achieved by means of introduction of *signature files* concept, in which these rules are defined. Signature files contain the signatures of basic concepts defined in some content dictionary and are used to check the semantic validity of their representations. A content dictionary is the most important component of OPENMath concept on preservation of mathematical information. In other words, content dictionaries are used to assign formal and informal semantics to all symbols (concepts) used in the OPENMath objects. A content dictionary is a collection of related symbols, encoded in XML format and fixing the "meaning" of concepts independently of the application.

2.2 Data Integration Model

The weakness of XML data model is the absence of data types concept in conventional sense. To eliminate this shortcoming and to support ontological dependencies on the XML data model level, we expand the XML data model by means of the OPENMath concept. The result of such extension is a data model which coincides with XML data model and which was strengthened with computational and ontological constructs of OPENMath. In the frame of this model we proposed a minor extension of OPENMath to support the built-in data types concept of the XML Schema. Namely, to model the constants of built-in data types of the XML Schema the corresponding basic objects were introduced. In the context of the considered data model we consider three kinds of mechanisms to formalize the data integration concept:

- content dictionaries to define basic concepts (integrable data, operations, types, etc.);
- signature files to define signatures of basic concepts to check the semantic validity of their representations;
- reasoning file to define knowledge in the frame of data integration concept. Defining a concept in terms of known ones we introduce a new concept (knowledge) in this area. Thus, the considered file is collections of reasoning rules, which are defining the new concepts in terms of known ones.

Extension Principle. Our concept to data integration assumes that the data integration model must be extensible. The extension of the data integration model is formed during consideration of each new data model by adding new concept(s) to its DDL to define logical data dependencies of the source model in terms of the target model if necessary. Thus, the data integration model extension assumes defining new symbols. The extension result must be equivalent to the source data model. For applying a *symbol* on the data integration model level the following rule is proposed:
Concept ← *symbol* ContextDefinition.

For example, to support the concepts of *key* of relational data model, we have expanded the data integration model with the symbol *key*. Let us consider a relational schema example: S={Snumber, Sname, Status, City}. The equivalent definition of this schema by means of extended data integration model is considered below:

S ← *attribution*(S, *type* TypeContext, *constraint* ConstraintContext)

TypeContext ← *application*(*sequence*, ApplicationContext)

ApplicationContext ← *attribution*(Snumber, *type int*),
attribution(Sname, *type string*), *attribution*(Status,
type int), *attribution*(City, *type string*)

ConstraintContext ← *attribution*(ConstraintName, *key* Snumber)

It is essential that we use a computationally complete language to define the context [11]. As a result of such approach usage of new symbols in the DDL does not lead to any changes in DDL parser. According to this approach, the data integration model is synthesized as a union of extensions. A schema of the integrated databases is an instance of the XML DTD for modeling reasoning rules.

3 Algebra of Integrable Data

In the frame of the data integration concept we differentiate one kind of data – integrable data.

3.1 Formalization of Integrable Data

Definition 1. *An integrable data schema X is an attribution object and is interpreted by a finite set of attribution objects $\{A_1, A_2, ..., A_n\}$. Corresponding to each attribution object A_i is a set D_i (a finite, non-empty set), $1 \leq i \leq n$, called the domain of A_i.*

Definition 2. *Let $D = D_1 \cup D_2 \cup ... \cup D_n$. An integrable data x on integrable data schema X is a finite set of mappings $\{e_1, e_2, \ldots, e_k\}$ from X to D with the restriction that for each mapping $e \in x$, $e[A_i]$ must be in D_i, $1 \leq i \leq n$. The mappings are called elements.)*

Definition 3. *A key of integrable data x is a minimal subset K of X such that for any distinct elements $e_1, e_2 \in x$, $e_1[K] \neq e_2[K]$.*

We introduce a symbol d to denote the set of all integrable data. It is assumed that the schema of each integrable data is a subset of the set of all attribution objects.

3.2 Operations

Virtual and materialization integration of data assumes introduction of special operations, such as filtering, joining, aggregating, etc. The proposed operations are similar to the realtional algebra operations.

To support n-ary associative operations union and joining, we introduced the symbols *union* and *join* correspondingly. The symbol *union* is used to denote the n-ary union of sets (integrable data). It takes sets as arguments, and denotes the set that contains all the elements that occur in any of them: $union : x^{*assoc} \rightarrow d$.

The symbol *join* is used to denote the n-ary join of sets. It takes sets as arguments, denotes a set of elements, and is interpreted analogously to the operation natural join of the relational algebra in general case (joins of many relations): $join : x^{*assoc} \rightarrow d$.

To support a filtering operation, we introduced the symbol σ. This symbol is used to denote a select operation on the set. It takes a set and a predicate as arguments, and denotes the set which contains all the elements for which the predicate is satisfied:

$\sigma : \{x \rightarrow \{p : \{element\} \rightarrow boolean\}\} \rightarrow d.$

Here p is a predicate which is applied to *element*.

To support a projection operation, we introduced the symbol π. This symbol is used to denote a unary operation on the set. It takes a set and a list of *attribution* object names as argument, denotes a set of elements, and is interpreted analogously to the operation *project* of the relational algebra:

$\pi : x[name^*] \to d.$

Here *name* denotes the name of an attribution object and is defined as follows:

$name : \{Attribution\} \to string.$

For integrating data, aggregating functions play a significant role. We introduced the *count*, *sum* and *avg* symbols to support the corresponding aggregate functions of the relational algebra. Let $f \in \{avg, sum, count\}$, then

$f : x[name] \to numericalvalue.$

Often, we needs to consider the elements of an integrable data in groups. For this purpose, we introduced a grouping symbol γ. This symbol is used to denote a unary operation on the set. It takes a set, a list of *attribution* object names and aggregate functions as arguments, denotes a set of elements, and is interpreted analogously to the operation *grouping* of the relational algebra:

$\gamma : x[name^*, (f : (element[name^*])^* \to numericalvalue)^*] \to d.$

4 An Ontology for Data Integration Concept

Formalization of the data integration concept assumes developing new content dictionaries to model the algebra of integrable data and data types concept of the XML Schema. Also we should define signatures of the introduced symbols (basic concepts) and reasoning rules of the data integration concept.

4.1 The *dic* Content Dictionary File

A content dictionary which contains representation of basic concepts of the data integration concept contains two types of information: one which is common to all content dictionaries, and one which is restricted to a particular basic concept definition. Definition of a new basic concept includes name and description of the basic concept, and also some optional information about this concept (analogously the *xts* content dictionary for modeling the type system concept of the XML Schema is defined). Below an example of a basic concept definition is considered:

```
<CDDefinition>
    <Name> X < /Name>
    <Description>
    To support the concept of integrable data schema we introduce
    the symbol X. Below we are using the Attribution symbol which has
    been defined in the OPENMath.
    < /Description>
    <CMP> X : Attribution* → {Attribution} < /CMP>
< /CDDefinition>
```

The above used XML elements have obvious interpretations. Only note, that the element ”CMP” contains the commented mathematical property of the defined algebraic concept. Specific information pertaining to the basic concept like the signature and the defining of a concept in terms of known ones is defined in additional files associated with content dictionaries. Content dictionaries contain just one part of the information that can be associated with a basic concept in order to stepwise define its meaning and its functionality. Signature files and files of reasoning are used to formalize the different aspects of the data integration concept. Namely, to formalize the basic concepts formats, and to define reasoning rules to formalize knowledge in this area.

4.2 The *dic* Signature File

As is mentioned above, to check semantic validity of the basic concepts representations we associate extra information with content dictionaries, namely signature files. A signature file contains the definitions of all the basic concept signatures of the considered content dictionary. We use Small Type System [4] to formalize the basic concept signatures. Below the definition of the signature of the above considered symbol X is provided :

```
<Signature name = “X”>
    <OMOB>
      <OMA>
        <OMS name = ”mapsto” cd = ”sts”/ >
        <OMA>
          <OMS name = ”nary” cd = ”sts”/ >
          <OMS name = ”attribution” cd = ”sts”/ >
        < /OMA>
        <OMS name = ”attribution” cd = ”sts”/ >
      < /OMA>
    < /OMOB>
< /Signature>
```

The above considered symbols *mapsto* and *nary* were defined in the OPENMath. The symbol *mapsto* represents the construction of a function type. The first n-1 children denote the types of the arguments, the last denotes the return type. The symbol *nary* constructs a child of *mapsto* which denotes an arbitrary number of copies of the argument of *nary*. The operator is associative on these arguments which means that repeated uses may be flattened/unflattened.

4.3 Reasoning File

We propose an XML DTD to define reasoning rules to support an ontology for data integration concept. The proposed ontology is based on the OPENMath formalism and the algebra of integrable data. As mentioned above, within our approach to ontology-based data integration we consider issues of virtual as well as materialized data integration. Therefore we should formalize the concepts of this subject area such as integrable data, mediator, data warehouse, data cube, etc.

Let the symbols *msch* and *wrapper* correspondingly denote the set of all mediator schemas and the set of all subsets of the wrappers which are defined on source data schemas to support the mediator concept, and let the symbol *med* denote the set of all mediators, then

$med \subseteq msch \times wrapper.$

The *msch* symbol is based on the OPENMath *attribution* concept. By means of this concept we can model source data schemas. The *wrapper* symbol is based on the OPENMath *application* concept and is presented by an algebraic program of the integrable data.

Let the symbols *wsch* and *extractor* correspondingly denote the set of all data warehouse schemas and the set of all subsets of the extractors which are defined on source data schemas to support the data warehouse concept, and let the symbol *whse* denote the set of all data warehouses, then

$wshe \subseteq wsch \times extractor.$

The symbols *wsch* and *extractor* are interpreted analougsly as in the mediator case.

Materialized integration of data assumes the creation of data warehouses. Our approach to create data warehouses is mainly oriented to support data cubes. Using data warehousing technologies in OLAP applications is very important [7]. Firstly, the data warehouse is a necessary tool to organize and centralize corporate information in order to support OLAP queries (source data are often distributed in heterogeneous sources). Secondly, significant is the fact that OLAP queries, which are very complex in nature and involve large amounts of data, require too much time to perform in a traditional transaction processing environment.

In typical OLAP applications, some collection of data called *fact table* which represent events or objects of interest are used [7]. Usually, fact table contains several attributes representing dimensions, and one or more dependent attributes that represent properties for the point as a whole. The creation of the data cube requires generation of the power set (set of all subset) of the aggregation attributes. To implement the formal data cube concept in literature the CUBE operator is considered [8]. In addition to the CUBE operator in [8] the operator ROLLUP is produced as a special variety of the CUBE operator which produces the additional aggregated information only if they aggregate over a tail of the sequence of grouping attributes. In this context, it is assumed that all independent attributes are grouping attributes. For some dimensions there are many degrees of granularity that could be chosen for a grouping on that dimension. When the number of choices for grouping along each dimension grows, it becomes non-effective to store the results of aggregating based on all the subsets of groupings.

Thus, it becomes reasonable to introduce materialized views. A materialized view is the result of some query which is stored in the database, and which does not contain all aggregated values. The materialized view is interpreted by the OPENMath *application* concept.

Let the symbols *ssch* and *mview* correspondingly denote the set of all fact table schemas which are defined on source data schemas and the set of all materialized views to support the data cube concept, and let the symbol *cube* denote the set of all data cubes in this context, then

$cube \subseteq ssch \times mview.$

As we noted above, the reasoning rules to support ontology for the data integration concept are based on the OPENMath formalism and the algebra of integrable data. Let the symbol *source* denote the set of all integrable data schemas and let the symbol *dir* denote the set of all data integration rules, then

$dir \subseteq source \times (med \cup whse \cup cube).$

In Appendix A an XML DTD for modeling the reasoning rules of the data integration concept is presented. Below, an example of a mediator for an automobile company database is adduced [7] which is an instance of the XML DTD of the data integration concept. It is assumed that the mediator with schema AutosMed = {SerialNo, Model, Color} integrates two relational sources: Cars = {SerialNo, Model, Color} and Autos = {Serial, Model}, Colors = {Serial, Color}.

```
<dir>
<! − − Source schemas definitions − − >
 <med>
  <msch>
  AutosMed: schema for mediator is defined
  < /msch>
  <wrapper>
   <OMA>
    <OMS name=”union” cd=”dic”/ >
    <OMV name=”cars”/ >
    <OMA>
     <OMS name=”join” cd=”dic”/ >
     <OMV name=”Autos”/ >
     <OMV name=”Colors”/ >
    < /OMA>
   < /OMA>
  < /wrapper>
 < /med>
< /dir>
```

5 Related Work

Using ontologies to support the data integration concept, it is explained that they provide an explicit and machine- understandable conceptualization of a subject domain. The use of ontologies for data integration is discussed in [3]. Examples of ontological modeling languages are XML Schema, RDFS, OWL, etc. There are the following variants of data integration based on the ontologies [19]:

- *Single-ontology.* All source schemas are directly related to a shared global ontology that provide a uniform interface to the user [2]. Single-ontology approach assumes that data sources are semantically close. SIMS [1] is a system which is based on such approach.

- *Multiple-ontology.* Each data source is defined independently using a local ontology. Such approach assumes developing a formalism for defining the inter-ontology mappings. The OBSERVER system [17] is based on this approach.

- *Hybrid-ontology.* This approach combines the two preceding approaches. In other words, for each source schema a local ontology is built alongside a global shared ontology. [2] is an example of this approach.

An important challenge in data integration is the construction of mappings from the source data models into the target one. As rules in well-known works, the mapping from the source models into the target one is constructed semi-automatically. The exception is the work [18] in which the mappings between relational databases and ontologies are generated automatically. Our approach to data integration allows to automatically generate mappings from arbitrary data models into the target one. A more detailed analysis of approaches to ontology-based data integration can be found in [6, 19].

6 Conclusions

In the frame of a data integration model, the data integration concept was formalized. The proposed ontology is based on the OPENMath formalism and the so-called algebra of integrated data which has also been developed by us. The considered data integration model is oriented to XML and is distinguished by its computational capabilities. Such data integration model is an extension of XML by means of the OPENMath concept. In the result of such extension, the XML data model has been strengthened with ontological and computional constructions. Three kinds of mechanisms of the OPENMath are used to formalize the data integration concept: content dictionary, signature file and reasoning file. By these mechanisms we formalize the different aspects of the data integration concept. Namely, we formalize basic concepts (integrable data and operations on it), their signatures and reasoning rules (to model the data integration concepts). The reasoning rules are based on the algebra of integrable data and are formalized by an XML DTD. If necessary, we can extend the algebra of integrable data by adding new algebraic operations. It should be noted that the different mechanisms of data translation (wrapper, extractor) are non-sensitive to such extension. It is essential that the data integration model is extensible and we use a computationally complete language to support the data integration concept. Finally, based on the proposed ontology, algorithms can be developed to generate the schemas of integrable data and transformers from high level concepts which are represented as an instance of the XML DTD.

Acknowledgments

This work was supported by the RA MES State Committee of Science, in the frames of the research project No. 18T-1B341.

References

1. Arens, Y., Knoblock, C.A., and Hsu, C.: Query processing in the sims information mediator. In *ARPI 1996*. AAAI Press, 61–99 (1996).
2. Cruz, I.F. and Xiao, H.: Using a layered approach for interoperability on the semantic web. In *WISE 2003*, 221–232 (2003).
3. Cruz, I.F. and Xiao, H.: The role of ontologies in data integration. *International Journal of Engineering Intelligent Systems for Electrical Engineering and Communications* **14** (4), 1–18 (2005).
4. Davenport, J.H.: A small openmath type system. *ACM SIGSAM Bulletin* **34** (2), 16–21 (2000).
5. Drawar, M.: Openmath: An overview. *ACM SIGSAM Bulletin* **34** (2), 2–5 (2000).
6. Ekaputra, F.J., Sabou, M., Serral, E., Kiesling, E., and Biffl, S.: Ontology-based integration in multi-disciplinary engineering environments: A review. *Open Journal of Information Systems* **4** (1), 1–26 (2017).
7. Garcia-Molina, H., Ullman, J., and Widom, J.: *Database Systems: The Complete Book.* Prentice Hall, USA, 2009.
8. Gray, J., Bosworth, A., Layman, A., and Pirahesh, H.: Data cube: A relational aggregation operator generalizing group-by, cross-tab, and sub-tab. In *ICDE*, 152–159. USA, 1996.
9. Hindley, J.R. and Seldin, J.P.: *Introduction to Combinators and λ-Calculus.* Cambridge University Pressl, Great Britain, 1986.
10. Kalinichenko, L.A.: Methods and tools for equivalent data model mapping construction. In *Advances in Database Technology-EDBT'90*, 92–119. Italy, Springer, March 1990.
11. Manukyan, M.G.: Extensible data model. In *Advances in Databases and Information Systems*, 42–57. Finland, 2008.
12. Manukyan, M.G.: Canonical model: Construction principles. In *iiWAS2014*, 320–329. Vietnam, ACM, December 2014.
13. Manukyan, M.G.: On an approach to data integration: Concept, formal foundations and data model. In *CEUR-WS* **2022**, 206–213 (2017).
14. Manukyan, M.G.: On an ontological modeling language by a non-formal example. In *CEUR-WS* **2277**, 41–48 (2018).
15. Manukyan, M.G. and Georgyan, G.R.: A dynamic indexing scheme for multidimensional data. *Modern Information Technologies and IT-Education* **14** (1), 111–125 (2018).
16. Manukyan, M.G. and Georgyan, G.R.: Canonical data model for data warehouse. In *Communications in Computer and Information Science* **637**, 72–79 (2016).

17. Mena, E., Kashyap, V., Sheth, A.P., and Illarramendi, A.: Observer: An approach for query processing in global information systems based on interoperation across pre-existing ontologies. In *CoopIS 1996*, 14–25 (1996).
18. Pinkel, C., Binnig, C., Jimenez-Ruiz, E., Kharlamov, E., Nikolov, A., Schwarte, A., Heupel, C., and Kraska, T.: Incmap: A journey towards ontology-based integration. In *BTW 2017*, 145–164. Lecture Notes in Informatics, 2017.
19. Wache, H., Vogele, T., Visser, U., Stuckensmidht, H., Schuster, G., Neumann, H., and Hubner, S.: Ontology-based integration of information-a survey of existing approaches. In *CEUR-WS*, 1–10 (2001).

APPENDIX A. An XML DTD for Modeling the Reasoning Rules of the Data Integration Concept

```
<! −− include dtd for extended OPENManth objects −− >
<!ELEMENT dir (source+, (med | whse | cube))>
<!ELEMENT med (msch, wrapper)>
<!ELEMENT msch (OMATTR)+ >
<!ELEMENT wrapper (OMA)>
<!ELEMENT whse (wsch, extractor)>
<!ELEMENT wsch (OMATTR)>
<!ELEMENT extractor (OMA)>
<!ELEMENT cube (ssch, mview)>
<!ELEMENT ssch (OMATTR)+ >
<!ELEMENT mview (view+, granularity+) >
<!ELEMENT view (OMA)>
<!ELEMENT granularity (partition)+ >
<!ELEMENT partition EMPTY>
<!ELEMENT source (OMATTR)+>
<!ATTLIST source name CDATA #REQUIRED>
<!ATTLIST granularity name CDATA #REQUIRED>
<!ATTLIST partition name CDATA #REQUIRED>
<!ATTLIST view name CDATA #REQUIRED>
```

Applying of Machine Learning Techniques to Combine String-based, Language-based and Structure-based Similarity Measures for Ontology Matching

Lev Bulygin[1][0000-0003-3244-1217] and Sergey Stupnikov[2][0000-0003-4720-8215]

[1] Lomonosov Moscow State University, Moscow, Russia
buliginleo@yandex.ru
[2] Institute of Informatics Problems, Federal Research Center “Computer Science and Control” of the Russian Academy of Sciences, Moscow, Russia
sstupnikov@ipiran.ru

Abstract. In the areas of Semantic Web and data integration, ontology matching is one of the important steps to resolve semantic heterogeneity. Manual ontology matching is very labor-intensive, time-consuming and prone to errors. So development of automatic or semi-automatic ontology matching methods and tools is quite important. This paper applies machine learning with different similarity measures between ontology elements as features for ontology matching. An approach to combine string-based, language-based and structure-based similarity measures with machine learning techniques is proposed. Logistic Regression, Random Forest classifier and Gradient Boosting are used as machine learning methods. The approach is evaluated on two datasets of Ontology Alignment Evaluation Initiative (OAEI).

Keywords: ontology matching, machine learning, similarity measures

1 Introduction

An *ontology* is “a formal, explicit specification of shared conceptualization” [1], where conceptualisation is an abstract model of some phenomenon in the world. Ontologies were created to facilitate the sharing of knowledge and its reuse [2]. They are used for organization of knowledge and for communication between computing systems, people, computing systems and people [3]. Ontologies deal with the following kinds of e*ntities:* classes, properties and individuals. A *class* (concept) of ontology is a collection of objects, i.e., “Person” (the class of all people) or “Car” (the class of all cars). *Property* (attribute) describes characteristics of a class or relations between classes, i.e., “has as name” or “is created by”. *Individual* (instance) is a particular instance or object represented by a concept, i.e., “a human cytochrome C” is an instance of the concept “Protein” [4].

Ontology matching is a process of establishing correspondences between semantically related entities in different ontologies [4]. A set of correspondences (equivalence, subsumption, disjointness) between ontologies elements is called an *alignment*. Ontology matching can be applied in many different subject areas: Semantic Web, Peer-to-Peer (P2P) systems, learning systems, multi-agent systems [5, 6].

In the *Semantic Web*, ontologies are used to extract logical conclusions from data. Many ontologies on the same subject areas have been created recently. These ontologies have a different format and they cannot exchange information, so it is necessary to apply ontology matching [4]. In *P2P systems* ontology matching is used to reduce the semantical heterogeneity (differences in the interpretation of the meaning) between the queries of the users to system [7]. In *learning systems* ontology matching is a way to ease the knowledge share and reuse [8]. In *multi-agent systems*, ontology matching is used for interaction of different agents [9].

Ontology matching can be used also for schema mapping during data integration [10]. *Data integration* is a process of combining the heterogeneous data sources into a unified view. *Schema mapping* is a process of establishing correspondences between elements of two different semantically related schemas (i.e. database schemas) [11]. Ontology matching can help to resolve semantical heterogeneity during schema mapping, for instance, if the schemas have ontologies as metadata or external domain knowledge [12, 13].

The classification of ontology matching approaches is very similar to the classification of schema matching approaches [11, 10, 4]. Matchers can be *individual* (use single matcher criterion) or *combining* (combination of individual matchers). An individual matcher can be *schema-based* (uses information about classes, properties and their relationships) or *instance-based* (uses information about instances/content). Schema-based matchers are divided into *element-level* (uses information about element without its relationships) and *structure-level* (uses information about structure and hierarchy). Combining matchers can be *hybrid* (creates alignment using several matching criteria in sequentially) or *composite* (combines several independent matching results). Composite matchers are divided into *manual composition* matchers and *automatic composition* matchers [11].

This paper proposes an approach for combining individual element-level and structure-level matchers into an automatic composition matcher based on machine learning. Individual matchers produce similarity measures between ontology elements. In terms of machine learning, similarity measures are used as *features* [14, 15]. The composite matcher is a machine learning model trained on these features. Logistic Regression, Random Forest and Gradient Boosting are used as the machine learning methods in the paper. The idea of the approach is as follows: to combine a large number of different similarity measures from other papers [32–34] in hope of increasing the universality of the approach, that is, applicability to different subject areas.

The paper is structured as follows. In Section 2 related works on application of machine learning for ontology and schema matching are reviewed. Section 3 describes the formal problem statement and an evaluation metric. In Section 4, similarity measures and machine learning techniques applied are listed. Section 5 considers implementation and evaluation issues.

2 Related Work

This section reviews related works on schema matching and ontology matching because they often applies similar techniques.

In [11], a classification of approaches for schema matching is introduced and a review of existing systems for schema matching is conducted. In [4], the authors described the classification of ontology matching approaches based on [11], existing matching systems, evaluation methods, similarity measures and matching strategies. The most promising option is the apply combining matcher because it uses much more information than an individual matcher. Many papers showed that the combining matchers are more accurate than individual ones [16–18]. Most approaches use string-based similarity measures, i.e., N-gram [17, 19, 20], Soundex [17, 21, 22], Levenshtein distance [17, 21, 23], Jaro measure [24–26] and others. Language-based similarity measures are also used, i.e., information from lexical database WordNet [19] [27] [28] or vector representation of words from Word2vec models [29] [21] [30] [31]. Some articles described structure-based similarity measures: differences between numbers of properties [15], similarity measures based on subclasses or parents [15] [19] and graph-based similarity [32].

Supervised machine learning for combining similarity measures is used in [14]. The authors describe an approach matching only concepts of ontologies. They used the string-based similarity measures (prefix, suffix, Edit distance, n-gram), language-based similarity measures (WordNet, Wu&Palmer, description, Lin), similarity measures between lists of words (for instance, name “socialNetwork” is divided into list of words [“Social”, “Network”]), and structure similarity measures (string-based and language-based similarity measures between parents). Support Vector Machine (SVM) is used as a machine learning method. The authors conducted experiments with data from Ontology Alignment Evaluation Initiative 2007 (OAEI). The data is constructed from three Internet directories (Google, Yahoo and Looksmart) and contains 4639 pairs of ontologies defined using OWL language. The authors used 10-fold cross validation and got 56.1% accuracy, 52.5% precision and 92.5% recall on average.

In [55] the authors combined the various similarity measures into a input sample for the first time. String-based, linguistic-based and structure-based similarity measures are used.

In [15] language, structural and web similarity measures are used. Web similarity measure “Web-dice” is the difference between the count of pages in a search engine when searching for an entity.` An SVM method is selected for training. The dataset used is OAEI benchmark tests ontologies. The authors trained two models “SVM-Class” and “SVM-Property” for matching classes and properties respectively.

In [54] 10 string-based, linguistic-based and instance-based similarity measures are used as features. Decision Tree (DT) and Naive Bayes are used for classification. The authors achieved 0.845 F-measure value.

In [33] SVM, K-Nearest Neighbours (KNN), SVM, DT and AdaBoost are used as the machine learning methods. The authors choose OAEI ontologies #301, #102 and #103 as train dataset and ontologies #302, #303, #304 as test dataset and achieved 0.99 F-measure value.

In [34] Stoilos, Soft Jaccard and Lin similarity measures are used for names, labels and comments of entities. The authors also used information on abbreviations. Samples from “Conference” track and benchmarks from OAEI are used as datasets. Multilayer perceptron, Decision Trees and M5Rules are selected as machine learning methods. The authors achieved 0.67 F-measure value.

In [31] the authors used string-based similarity measures, measures related with parents and children of entities and chose “Conference” track from OAEI and EuroVoc dataset.

Note that known works use different datasets for their experiments and it is very hard to compare them with each other.

3 Ontology Matching as a Machine Learning Problem

3.1 Formal Problem Statement of Ontology Matching

Let ontology be a tuple *(C, P, H)*. Here *C* is a set of classes, *P* is a set of properties. *H* define the hierarchical relationships between classes. Other components of ontologies like axioms and instances are not applied for ontology matching in the paper. The objective of ontology matching is to find an alignment between classes and properties of a source ontology O_1 and a target ontology O_2. An *alignment* is a set of tuples (e_1, e_2, con), where e_1 is an entity of O_1, e_2 is an entity of O_2, and con is the confidence of the correspondence. *A predicted alignment* is the alignment obtained by ontology matching. *A true alignment* is a manual alignment conducted by an domain expert.

3.2 Ontology Matching Problem as a Machine Learning Problem

Entity pairs are extracted from source and target ontologies. Each pair of entities is assigned with a label “0” or “1”, where “0” means that entities do not match, “1” means that entities match. Thus, the problem is reduced to a machine learning binary classification problem. The authors of most of the reviewed papers and OAEI used F-measure for the evaluation of their approaches [4, 33].

4 An Approach for Ontology Matching Applying Machine Learning Models Trained on Similarity Measures

This section describes machine learning techniques applied (subsection 4.1), similarity measures used (subsection 4.2) and algorithms constituting the approach (subsection 4.3).

4.1 Machine Learning Techniques

The following machine learning methods are applied in this paper: Logistic Regression, Random Forest and Gradient Boosting. In [31, 47] it is shown that a powerful ensemble method Random Forest [48] and relatively simple and interpretable Logistic Regression outperformed other machine learning algorithms like Gaussian Naive Bayes, K-nearest Neighbors Algorithm, Classification and Regression Trees for ontology matching. Gradient boosting has proven itself in many machine learning contests [49, 50], so it was also selected as a machine learning method to be applied. In the future, we also want to test neural network (multilayer perceptron) as a machine learning method and an approach based on automatic machine learning[1].

4.2 Similarity Measures

String-based. We used all string-based similarity measures listed in our previous work [21]. The listed metrics are aimed at handling various sorts of scenarios. *N-gram* consider similarity of substrings and it is efficient when some characters are missing [4]. *Dice coefficient* is defined as twice the number of common words of compared strings over the total number of words in both strings [35]. *Jaccard* and *Generalized Jaccard* similarity are defined as the size of the intersection divided by the size of the union of the sample sets of words **[**24]. *Levenshtein distance* between two strings is the minimum number of single-character edits required to change one word into the other [36]. *Jaro* and *Jaro–Winkler* measures is edit distance measure designed for short strings [37]. *Monge–Elkan* is a type of hybrid similarity measure that combines the benefits of sequence-based and set-based methods [38]. The *Smith-Waterman* measure determine similar regions between two strings [35]. The *Needleman–Wunsh* distance is computed by assigning a score to each alignment between the two input strings and choosing the score of the best alignment [39]. The *Affine gap distance* is an extension of the Needleman–Wunsch measure that handles the longer gaps more gracefully [40]. The *Bag distance* is edit distance for sets of words [52]. *Cosine similarity* transforms a string into vector so Euclidean cosine rule is used to determine similarity [24]. *Fuzzy Wuzzy Partial Ratio* finds the similarity measure between the shorter string and every substring of length *m* of the longer string, and returns the maximum of those similarity measures [41]. *Soft TF-IDF* and *TF-IDF* are numerical statistics that are intended to reflect how important a word is to a document in a collection or corpus [39]. *Partial Token Sort*[2] and *Token Sort* are obtained by splitting the two strings into tokens and then sorting the tokens. The score is the fuzzy wuzzy partial ratio raw score of the transformed strings. *Fuzzy Wuzzy Ratio* is the ratio of the number of matching characters to the total number of characters of two strings [41]. *Editex* [42] and *Soundex*[3] are phonetic matching measures. *Tversky Index* is an asymmetric similarity measure on sets that compares a

[1] https://github.com/automl/auto-sklearn

[2] https://anhaidgroup.github.io/py_stringmatching/v0.3.x/PartialTokenSort.html

[3] http://anhaidgroup.github.io/py_stringmatching/v0.4.1/Soundex.html

variant to a prototype [43]. *Overlap coefficient* is defined as the size of the intersection divided by the smaller of the size of the two sets [44].

Language-based. It is possible that words differ but are close in meaning, i.e., "car" and "auto". *WordNet* can solve this problem. *Wu and Palmer similarity* are used for handling this scenario [45]. If the strings consist of several words then the maximum similarity measure of all possible pairs of sets of words is taken. But the weakness of WordNet is that it contains only a part of all words of the language. Usage of vector representations of words from *Word2vec* models [46] facilitates this problem. Cosine similarity between two vector representations of words is calculated. If the strings consists of several words then *Sentence2vec* algorithm from [30] is used.

Structure-based. Additionally, structure-based similarity measures are used: all listed string-based and language-based similarity measures between parents of entities and between paths of entities. These similarity measures embrace the hypothesis that matched entities have similar parents and a similar place in hierarchy.

Since we used the same model for the match of classes and properties, we added feature "Type", in which label "1" means class and label "0" means property.

Such an extensive selection of similarity measures is aimed to get as much information as possible so that a machine learning model is able to select the best factors for prediction. Finally, we chose for each pair of entities 88 similarity measures (29 for names, 29 for parents, 29 for paths, 1 for type), which are described in Table 1.

Table 1. Similarity measures

String-based	N-gram 1, N-gram 2, N-gram 3, N-gram 4, Dice coefficient, Jaccard similarity, Jaro measure, Monge-Elkan, Smith-Waterman, Needleman-Wunsh, Affine gap, Bag distance, Cosine similarity, Partial Ratio, Soft TF-IDF, Editex, Generalized Jaccard, Jaro-Winkler, Levenshtein distance, Partial Token Sort, Fuzzy Wuzzy Ratio, Soundex, TF-IDF, Token Sort, Tversky Index, Overlap coefficient, Longest common subsequence
Language-based	Wu and Palmer similarity Word2vec and Sentence2vec similarity
Structure-based	All string-based and language-based similarity measures between parents of entities All string-based and language-based similarity measures between paths of entities

4.3 Training and Matching Algorithms

The approach is restricted with the following limitations: entities are matched only by equivalence relation, classes are matched only with classes, properties are matched only with properties, instances of ontologies are not used. The approach includes two main algorithms: training of a machine learning model (training phase) and using it to predict alignment (testing phase).

The ontology matching algorithm using the trained model is described as follows:

Algorithm 1 Matching algorithm

Input:
ontology1, *ontology2* – input ontologies,
THRESHOLD – threshold for create matching between entities
Auxiliary functions:
get_classes – get list of classes,
get_properties – get list of properties,
create_alignment – Algorithm 2
Output: *final_alignment* – output alignment for *ontology1* and *ontology2*

```
classes1 ← get_classes(ontology1)
classes2 ← get_classes(ontology2)
alignment_classes ← create_alignment(classes1, classes2, THRESHOLD)
properties1 ← get_properties(ontology1)
properties2 ← get_properties(ontology2)
alignment_properties ← create_alignment(properties1, properties2,
THRESHOLD)
final_alignment ← alignment_classes ∪ alignment_properties
return final_alignment
```

Here ← denotes an assignment operation, and ∪ – the operation of merging lists. The input data of the algorithm are two ontologies and a matching probability threshold for filtering pairs of entities. If the probability is higher than the threshold, then the pair is added to the alignment. A list of classes is extracted from each ontology. Next, two lists of classes are fed to the input of Algorithm 2:

Algorithm 2 Creating predicted alignment from two lists of entities – *create_alignment(entities1, entities2, THRESHOLD)*

Input:
entities1, *entities2* – input lists of entities (classes or properties),
THRESHOLD – threshold for create matching between entities
Auxiliary functions:
calculate_all_sim_measures – Algorithm 3,
predict_match – predict confidence based on similarity measures
Output: *alignment* – output alignment for *entities1* and *entities2*

```
for entity1 ∈ entities1 do
    for entity2 ∈ entities2 do
        sim_measures ← calculate_all_sim_measures(entity1, entity2)
        match ← predict_match(sim_measures)
        if match > THRESHOLD then
            alignment ← alignment ∪ (entity1, entity2)
        end if
    end for
end for
return alignment
```

Then, each class from the first ontology is matched with each class from the second ontology. For example, if in the first ontology includes 10 classes and in the second ontology includes 12 classes, then 120 pairs are matched. Each pair is fed to the input of a machine learning model, which calculates the probability (confidence) of matching for each pair. Then the threshold is set: if the probability is above the threshold, then the pair is added to the final alignment. Similar actions are performed for properties. The similarity measures for each pair are calculated in Algorithm 3:

Algorithm 3 Calculating similarity measures algorithm –
calculate_all_sim_measures(entity1, entity2)
Input:
entity1, *entity2* – input entities (classes or properties)
Auxiliary functions:
get_name – get name of entity,
get_parent – get parent of entity,
get_path – get full path of entity,
calculate_sim_measures – calculates string-based and linguistic-based similarity measures listed in 4.2 and returns a list of 88 values
concat – merge lists
Output: *sim_measures* – output list of calculated similarity measures for *entity1* and *entity2*

```
name1 ← get_name(entity1)
name2 ← get_name(entity2)
parent1 ← get_parent(entity1)
parent2 ← get_parent(entity2)
path1 ← get_path(entity1)
path2 ← get_path(entity2)
name_sim_measures ← calculate_sim_measures(name1, name2)
parent_sim_measures ← calculate_sim_measures(parent1, parent2)
path_sim_measures ← calculate_sim_measures(path1, path2)
sim_measures ← concat(name_sum_measures, parent_sim_measures,
                      path_sim_measures)
return sim_measures
```

Name, parent name, and the full hierarchical path are retrieved from each class. The parent of a class is its super class. The full path is a string that describe the entire hierarchy of classes: from the most general class to the current class. For example, the class “Book” has the name “Book”, the parent name “Publication” and the full path “Thing/Publication/Book”. Thus, a list of pairs for matching is generated. For properties, the parent is the class that it describes. And the full path is a string describing the complete hierarchy up to the class that describes the property. For each pair, all similarity measures listed in Section 4.2 are calculated. Then all similarity measures are combined into a list.

The algorithm of model training is described as follows:

Algorithm 4 Creating dataset and training a machine learning model

Input:
train_pairs_ontologies – set of tuples *(ontology1, ontology2, true_alignment)*
model_name – name of machine learning method (logistic regression, random forest, gradient boosting),
model_params – set of parameters of machine learning model,
create_dataset – Algorithm 5
Auxiliary functions: *train_model* – train machine learning model on training dataset
Output: *model* – trained model for predicting matching

```
for ontology1, ontology2, true_alignment in train_pairs_ontologies do
    classes1 ← get_classes(ontology1)
    classes2 ← get_classes(ontology2)
    train_dataset_classes ← create_dataset(classes1, classes2, true_alignment,
                                           'Class')
    properties1 ← get_properties(ontology1)
    properties2 ← get_properties(ontology2)
    train_dataset_properties ← create_dataset(properties1, properties2,
                                              true_alignment, 'Property')
    train_dataset ← train_dataset ∪ train_dataset_classes ∪
                        train_dataset_properties
end for
model ← train_model(train_dataset, model_name, model_params)
return model
```

The input data is a list of ontology pairs and the true alignment between them. A model from Section 2.5 and its parameters are also selected. The process is similar to the first algorithm: the names of objects, the names of parents and full paths are retrieved, and the similarity measures are calculated. First, a dataset is created for the classes, then for properties, and after that the datasets are combined.

The algorithm for creating a dataset is described in Algorithm 5:

Algorithm 5 Creating dataset from two lists of entities – *create_dataset(entities1, entities2, true_alignment, type_entity)*

Input:
true_alignment – set of matched pairs of entities,
entities1, *entities2* – input lists of entities (classes or properties),
type_entity – type of input entities (class or property)
Auxiliary functions: *train_model* – train machine learning model on training dataset
Output: *train_dataset* – output list of tuples with pairs of entities, their matchings and similarity measures

```
for entity1 $\in$ entities1 do
    for entity2 $\in$ entities2 do
        sim_measures $\leftarrow$ calculate_all_sim_measures(entity1, entity2)
        if (entity1, entity2) $\in$ true_alignment then
            train_dataset $\leftarrow$ train_dataset $\cup$ (entity1, entity2, 1, type_entity,
                                                          sim_measures)
        else
            train_dataset $\leftarrow$ train_dataset $\cup$ (entity1, entity2, 0, type_entity,
                                                          sim_measures)
        end if
    end for
end for
return train_dataset
```

The input is a true alignment, two lists of entities and the type of input entities. Each entity from the first list is mapped to each entity from the second list. Then, if a pair of entities is contained in the true alignment, then the pair is assigned label “1”, otherwise – label “0”. Also, each pair indicates the type of entity (either “Class” or “Property”) because the same model was used to map classes and properties. Then all pairs are combined into one dataset. Further, the model is trained on the created dataset with the selected parameters.

5 Implementation and Evaluation Results

5.1 Datasets

Two datasets are selected for evaluation experiments (called as Dataset #1 and Dataset #2 below). These datasets are sets of ontologies and their true alignments taken from Ontology Alignment Evaluation Initiative (OAEI). Some pairs of ontologies and their true alignments are selected for training the machine learning models and their testing. This selection is called a *partition*. Ontologies from OAEI are used in many papers. These papers include [33] and [34]; [33] presents an approach to combining similarity measures without instances of ontologies and user feedback. KNN, SVM, DT and AdaBoost were used as machine learning models. The authors achieve on some alignments the value of F-measure 0.99; [34] proposed a new ontology matching approach. The authors used five different similarity measures: syntactic, semantic, abbreviation and context similarity. Multilayer Perceptron, REPTree, M5Rules are used as the machine learning models. Average F-measure is 0.67. Dataset #1 is a partition from third experiment of [33]. Dataset #2 is a partition from [34]. The used pairs of ontologies and their true alignments are described in Tables 3 and 4. All ontologies are defined using OWL-DL[4] language in the RDF and XML format.

Dataset #1 is a set of ontologies about Bibliographic references from Benchmark test library. Ontology #101 is the reference ontology. Other ontologies (#102–#103, #301–#304) are compared with the reference ontology. Dataset has 7 ontologies and 6 true alignments: 3 alignments for training and 3 alignments for testing.

Dataset #2 consists several ontologies from Benchmark test library and all ontologies from Conference track of OAEI. Conference track contains 16 ontologies, which dealing with conference organization, and 21 true alignments. Dataset has 27 ontologies and 26 alignments: 8 alignments for training and 18 alignments for testing.

The pairs of entities from each pair of ontologies and their alignments are extracted. Dataset #1 has 14148 training samples (156 positive and 13992 negative samples) and 14940 testing samples (172 positive and 14768 negative samples) and Dataset #2 has 55348 training samples (284 positive and 55064 negative samples) and 114045 testing samples (253 positive and 113792 negative samples). A positive sample is a pair of entities which are matching, and a negative example is a pair of non-matching entities. Note that the datasets are very unbalanced.

Table 2. Example part of true alignment 101–302

Entity from Ontology #101	Entity from Ontology #302
Collection	Book
TechReport	TechReport
Report	Publication
Reference	Resource
date	publishedOn

[4] https://www.w3.org/TR/owl-features/

5.2 Implementation

The approach was implemented using Python 3.5. This language is widely used for implementation of machine learning workflows and possesses a lot of useful program libraries.

Ontologies are represented as RDF/OWL files. The *owlready2*[5] library was used for syntactic parsing of ontologies. Alignments are defined in RDF format. For parsing alignments, the *BeautifulSoup*[6] library was used. As implementation of logistic regression and random forest machine learning techniques *sklearn*[7] library is used. As a gradient boosting implementation the XGBoost[8] library is used. Dataset is formed as a dataframe of the *pandas*[9] library. To evaluate F-measure, the Alignment API[10] library was used. Computation experiments: training of machine learning models and the selection of their parameters were performed at the Hybrid high-performance computing cluster [51]. WordNet dictionary is taken from the *nltk*[11] library. Word2vec model was trained on GoogleNews[12] news. N-gram implementation is taken from the *ngram*[13] library. Similarity measures based on edit distance implementation is taken from the *editdistance*[14] library.

5.3 Experiments

The best parameters for the models were selected by the brute force method (a grid of values was created for each parameter): the models were trained on all combinations of parameters and the model with the best F-measure value using threshold 0.5 was selected.

For logistic regression, the following parameters were selected: inverse of regularization strength, weights of classes and norm used in the penalization. For random forest the number of trees in the forest, the maximum depth of the tree, the number of features to consider when looking for the best split and class weights were selected. For XGBoost, minimum sum of instance weight, minimum loss reduction requited to make a further partition, subsample ratio for training instances, subsample ratio of columns when constructing each tree and maximum depth of a tree were selected.

After training and searching for the best parameters, a threshold was selected with the highest F-measure value for each alignment. For each machine learning model, a grid of values for parameters was created manually. For numerical parameters, a grid of 3–5 values with different steps was created, i.e. for numbers of estimators in random

[5] https://owlready2.readthedocs.io/en/latest/

[6] https://pypi.org/project/beautifulsoup4/

[7] https://scikit-learn.org/stable/

[8] https://xgboost.readthedocs.io

[9] https://pandas.pydata.org

[10] http://alignapi.gforge.inria.fr

[11] https://www.nltk.org

[12] https://github.com/mmihaltz/word2vec-GoogleNews-vectors

[13] https://pythonhosted.org/ngram/

[14] https://pypi.org/project/editdistance/

forest: 10, 100, 200, 500, 1000. For parameters with options, all possible options were taken (2–4 options).

The values of F-measure for each alignment are presented in Tables 3, 4 and 5. The best models for the Dataset #1 are logistic regression and random forest. Gradient boosting is a bit less accurate. However, the gradient boosting is the best model on average on Dataset #2. It is more accurate than logistic regression at 0.02 and than random forest at 0.01. The values of F-measure on Dataset #1 are comparable with the classical methods [37, 38, 54, 55] but lower than [33]. This may be associated with a specific set of training and test datasets, and it is also possible that the metrics that were not implemented in this work have an impact. In [33] the importance of each similarity measures is not described, but there is a hypothesis that the main contribution comes from similarity measures associated with comments to entities, and two structural measures from [55]. The study of this issue is future work. The values of F-measure on Dataset #2 are comparable with [34].

Table 3. F-measure values for Dataset #1 with best thresholds

Alignment	Logistic Regression	Random Forest	XGBoost	Best results from [33]	FOAM [37]	DT [54]	OMAP [38]	OLA [55]
101–302	0.72	0.71	0.72	0.92	0.77	0.759	0.74	0.34
101–303	0.82	0.82	0.75	0.90	0.84	0.816	0.84	0.44
101–304	0.90	0.91	0.91	0.97	0.95	0.96	0.91	0.69
Average	0.81	0.81	0.79	**0.93**	0.85	0.845	0.83	0.49

The computational complexity of the approach is $O(n_1n_2+m_1m_2)$, where n_1, n_2 are the number of classes in the ontology O_1 and O_2, and m_1 and m_2 are the number of properties. The computation time and the used memory depending on the size of the ontology are showed on Figs. 1 and 2. The calculations were performed on MacBook Air 1.8 GHz 8GB RAM. The dependence of training and testing time on the ontology size is showed on Fig. 1. 20 points (evenly distributed between 10 and 1000) were used to build the figure. Training of random forest is longer than training of logistic regression and gradient boosting. The reason for the jumps on the figure is that the dataset is sampled randomly. Unfortunately, the used machine did not have enough capacity to calculate the time for training and testing gradient boosting with an ontology size of more than 600. The testing time of logistic regression and gradient boosting is much less than a random forest, therefore in the figure the graphs are close to zero. In general, there is a quadratic dependence. The dependence of memory usage on the ontology size is showed on Fig. 2. 20 points were also used to build the figure. Memory was measured using the memory profiler[15] package: the amount of used memory was measured when

[15] https://pypi.org/project/memory-profiler/

running the training and testing script. It is hard to understand why increasing the size of the ontology does not increase the amount of memory used, perhaps this is due to the internal work of the Python language. It is noticeable that the most memory is used by gradient boosting.

Table 4. F-measure values for Dataset #2 with best thresholds

Alignment	Logistic Regression	Random Forest	XGBoost
conference-edas	0.53	0.5	0.55
cmt-sigkdd	0.73	0.8	0.73
edas-sigkdd	0.53	0.63	0.63
ekaw-sigkdd	0.77	0.77	0.77
cmt-edas	0.72	0.76	0.63
conference-sigkdd	0.64	0.54	0.58
confof-edas	0.62	0.62	0.62
confof-iasted	0.71	0.61	0.66
conference-confof	0.61	0.54	0.57
cmt-confof	0.44	0.41	0.48
conference-ekaw	0.43	0.40	0.47
cmt-ekaw	0.58	0.62	0.70
confof-ekaw	0.58	0.68	0.64
iasted-sigkdd	0.75	0.81	0.81
cmt-iasted	0.88	0.88	0.88
edas-iasted	0.42	0.57	0.57
ekaw-iasted	0.58	0.75	0.70
confof-sigkdd	0.72	0.72	0.72
Average	0.62	0.64	0.65

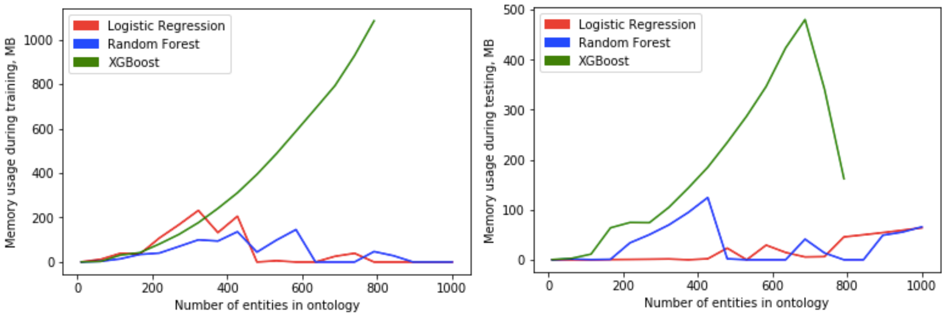
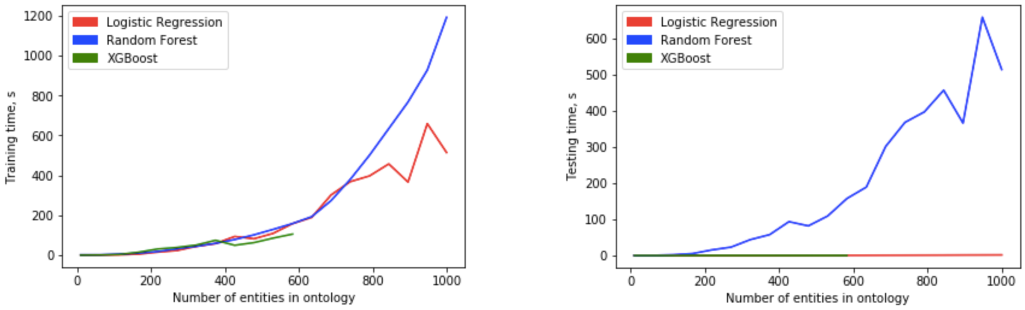


Fig. 1. Training and testing time of approach depending on the size of ontology

Table 5. Comparison of F-measure values



Alignment	Logistic Regression	Random Forest	XGBoost	Multi-layer perceptron [34]	REPTree [34]	M5 Rules [34]
Average	0.62	0.64	0.65	**0.67**	0.65	0.65

Fig. 2. Memory usage during training and testing

6 Conclusions and Future Work

We combined string-based, language-based, and structural-based similarity measures using three different machine learning models and apply them for ontology matching problem. The approach is implemented and evaluated using datasets selected from Ontology Alignment Evaluation Initiative (OAEI).

Due to the large number of similarity measures, there is hope that there is a potential for a more universal use of the approach. Universality refers to the applicability of the different subject areas. It is necessary to test the approach on ontologies with other subject areas. As a future work we would like to add similarity measures based on comments of entities, more structure-based similarity measures, such as a path length, a number of children, a number of properties of a class. It is also necessary to test the similarity measure from [53]. Neural network (multilayer perceptron) is planned to be used as a machine learning model. Evaluation issues to be resolved are checking the effectiveness of learning two different models separately for classes and properties and testing different strategies to resolve a problem of the strong imbalance of classes as well as strategies for significant reduce of a number of pairs of entities for matching.

Acknowledgements

The research is financially supported by Russian Foundation for Basic Research, projects 18-07-01434, 18-29-22096. The calculations were performed by Hybrid high-performance computing cluster of FRC CS RAS [51].

References

1. Gruber, T.: A Translation Approach to Portable Ontology Specifications. In: Knowledge Acquisition – Special issue: Current issues in knowledge modeling, vol. 5, issue 2 (1993). doi: 10.1006/knac.1993.1008
2. Fensel, D.: Ontologies: Silver Bullet for Knowledge Management and Electronic Commerce. doi: 10.1007/978-3-662-09083-1
3. Gruninger, M. and Lee, J.: Ontology Applications and Design – Introduction. In: Communications of the ACM (2002). doi: 10.1145/503124.503146
4. Euzenat, J. and Shvaiko, P.: Ontology Matching. Springer-Verlag Berlin Heidelberg, Berlin (2007). doi: 10.1007/978-3-642-38721-0
5. Otero-Cerdeira, L., Rodríguez-Martínez, F., and Gómez-Rodríguez, A.: Ontology Matching: A Literature Review. In: Expert Systems with Applications **42** (2), 949–971 (2015). doi: 10.1016/j.eswa.2014.08.032
6. Shvaiko, P. and Euzenat, J.: A Survey of Schema-Based Matching Approaches. In: Journal on Data Semantics IV, 146–171 (2005). doi: 10.1007/11603412_5
7. Atencia, M., Euzenat, J., Pirro, and G., Rousset, M.: Alignment-Based Trust for Resource Finding in Semantic P2P Networks. In: The Semantic Web – ISWC 2011: 10th International Semantic Web Conference, pp. 51–66 (2011). doi: 10.1007/978 -3-642-25073-6_4
8. Arch-int, N. and Arch-int, S.: Semantic Ontology Mapping for Interoperability of Learning Resource Systems using a rule-based reasoning approach. In: Expert Systems with Applications **40** (18), 7428–7443 (2013). doi: https://doi.org/10.1016/j.eswa.2013.07.027
9. Mascardi, V., Ancona, D., Bordini, R., and Ricci, A.: CooL-AgentSpeak: Enhancing AgentSpeak-DL Agents with Plan Exchange and Ontology Services. In: WI-IAT '11 Proceedings of the 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology, vol. 2, pp. 109–116 (2011). doi: 10.1109/WI-IAT.2011.255
10. Dong, X. and Srivastava, D.: Big Data Integration. In: 2013 IEEE 29th International Conference on (2015). doi: 10.1109/ICDE.2013.6544914

11. Rahm, E. and Bernstein, P.: A survey of approaches to automatic schema matching. In: The International Journal on Very Large Data Bases **10** (4), 334–350 (2001). doi: 10.1007/s007780100057
12. Hlaing, S.: Ontology based schema matching and mapping approach for structured databases. In: Proceedings of the 2nd International Conference on Interaction Sciences: Information Technology, Culture and Human, pp. 853–859 (2009). doi: 10.1145/1655925.1656080
13. Nathalie, A.: Schema Matching Based on Attribute Values and Background Ontology. In: 12th AGILE International Conference on Geographic Information Science (2009).
14. Ichise, R.: Machine Learning Approach for Ontology Mapping using Multiple Concept Similarity Measures. In: Seventh IEEE/ACIS International Conference on Computer and Information Science (2008). doi: 10.1109/ICIS.2008.51
15. Mao, M., Peng, Y., and Spring, M.: Neural Network based Constraint Satisfaction in Ontology Mapping. In: Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence, vol. 2, pp 1207–1212 (2008). http://www.dit.unitn.it/~p2p/RelatedWork/Matching/AAAI10-MaoM.pdf
16. Do, H., Melnik, S., and Rahm, E.: Comparison of Schema Matching Evaluations. In: Revised Papers from the NODe 2002 Web and Database-Related Workshops on Web, Web-Services, and Database Systems, pp. 221–237 (2002). doi: 10.1007/3-540-36560-5_17
17. Do, H. and Rahm, E.: COMA: a system for flexible combination of schema matching approaches. In: VLDB '02 Proceedings of the 28th international conference on Very Large Data Bases, pp. 610–621 (2002). doi: 10.1016/B978-155860869-6/50060-3
18. Xu, L. and Embley, D.: Automating Schema Mapping for Data Integration. (2003). http://www.deg.byu.edu/papers/AutomatingSchemaMatching.journal.pdf
19. Lambrix, P. and Tan, H.: SAMBO – A system for aligning and merging biomedical ontologies. In: Journal of Web Semantics **4** (3), 196–206 (2006). doi: 10.1016/j.websem.2006.05.003
20. Ngo, D.: Enhancing Ontology Matching by Using Machine Learning, Graph Matching and Information Retrieval Techniques. In: University Montpellier II – Sciences et Techniques du Languedoc (2012). doi: 10.1.1.302.587
21. Bulygin, L.: Combining Lexical and Semantic Similarity Measures with Machine Learning Approach for Ontology and Schema Matching Problem. In: Selected Papers of the XX International Conference on Data Analytics and Management in Data Intensive Domains, pp. 245–249 (2018).
22. Gal, A., Modica, G., Jamil, H., and Eyal, A.: Automatic Ontology Matching Using Application Semantics. In: AI Magazine – Special issue on semantic integration **26** (1), 21–31 (2005).
23. Hariri, B., Sayyadi, H., and Abolhassani, H.: Combining Ontology Alignment Metrics Using the Data Mining Techniques. In: Proceedings of the 2nd International Workshop on Contexts and Ontologies: Theory, Practice and Applications (2006).
24. Stoilos, G., Stamou, G., and Kolias, S.: A String Metric for Ontology Alignment. In: The Semantic Web – ISWC 2005, pp. 624–637 (2005). doi: 10.1007/11574620_45
25. Cheatham, M. and Hitzler, P.: String Similarity Metrics for Ontology Alignment. In: The Semantic Web – ISWC 2013, pp. 294–309 (2013). doi: 10.1007/978-3-642-41338-4_19
26. Saruladha, K., Aghila, and G., Sathiya, B.: A Comparative Analysis of Ontology and Schema Matching Systems. In: International Journal of Computer Applications **34** (8), 14–21 (2011).
27. Jean-Mary, R., Shironoshita, and P., Kabuka, M.: Ontology Matching with Semantic Verification. In: Web Semant. **7** (3), 235–251 (2009). doi: 10.1016/j.websem.2009.04.001

28. Seddiqui, H. and Aono, M.: Anchor-flood: Results for OAEI 2009. In: Proceedings of the 4th International Workshop on Ontology Matching collocated with the 8th International Semantic Web Conference (2009).
29. Kolyvakis, P., Kalousis, A., and Kiritsis, D.: DeepAlignment: Unsupervised Ontology Matching with Refined Word Vectors. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, vol. 1 (2018). doi: 10.18653/v1/N18-1072
30. Zhang, Y., Wang, X., Lai, S., He, S., Liu, K., Zhao, J., and Lv, X.: Ontology Matching with Word Embeddings. In: Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data, pp. 34–45 (2014). doi: 10.1007/978-3-319-12277-9_4
31. Nagy, M., Vargas-Vera, M., and Motta, E.: DSSim-ontology mapping with uncertainty. In: 1st International Workshop on Ontology Matching (2006).
32. Nkisi-Orji, I., Wiratunga, N., Massie, S., Hui, K., and Heaven, R.: Ontology alignment based on word embedding and random forest classification. In: Energy Transfer Processes in Polynuclear Lanthanide Complexes, pp. 557–572 (2018). doi: 10.1007/978-3-030-10925-7_34
33. Nezhadi, A., Shadgar, B., and Osareh, A.: Ontology Alignment Using Machine Learning Techniques. In: International Journal of Computer Science & Information Technology **3**, 39–150 (2011). doi: 10.5121/ijcsit.2011.3210
34. Alboukaey, N. and Joukhadar, A.: Ontology Matching as Regression Problem. In: Journal of Digital Information Management **16** (1) (2018). http://dline.info/fpaper/jdim/v16i1/jdimv16i1_4.pdf
35. Cohen, W., Ravikumar, P., and Fienberg, S.: A Comparison of String Metrics for Matching Names and Records.
36. Euzenat, J.: An API for ontology alignment. In: The Semantic Web – ISWC 2004: Third International Semantic Web Conference (2004). doi: 10.1007/978-3-540-30475-3_48
37. David, J., Guillet, F., and Briand, H.: Association Rule Ontology Matching Approach. In: International Journal on Semantic Web and information systems **3** (2), 27–49 (2007).
38. Straccia, U. and Troncy, R.: oMAP: Combining Classifiers for Aligning Automatically OWL Ontologies. In: Web Information Systems Engineering, pp. 133–147 (2005). doi: 10.1007/11581062_11
39. Needleman, S. and Wunsch, C.: A General Method Applicable to Search for Similarities in Amino Acid Sequence of 2 Proteins. In: Journal of Molecular Biology **48** (3), 443–453 (1970). doi: 10.1016/0022-2836(70)90057-4
40. Doan, A., Halevy, and A., Ives, Z.: Principles of Data Integration. (2012). doi: 10.1016/C2011-0-06130-6
41. Appa Rao, G., Srinivas, G., Venkata Rao, K., and Prasad Reddy, P.: A partial ratio and ratio based fuzzy-wuzzy procedure for characteristic mining of mathematical formulas from documents. (2018). doi: 10.21917/ijsc.2018.0242
42. Zobel, J. and Dart, P.: Phonetic String Matching: Lessons from Information Retrieval. In: SIGIR '96 Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 166–172 (1996). doi: 10.1145/243199.243258
43. Tversky, A.: Features of Similarity. In: Psychological Review **84** (4), 327–352 (1977). doi: 10.1037/0033-295X.84.4.327
44. Vijaymeena, M. and Kavitha, K.: A Survey on Similarity Measures in Text Mining. (2016). doi: 10.5121/mlaij.2016.3103
45. Wu, Z. and Palmer, M.: Verbs Semantics and Lexical Selection. In: Proceedings of the 32nd annual meeting on Association for Computational Linguistics (1994).

doi: 10.3115/981732.981751

46. Mikolov, T., Corrado, G., Chen, K., and Dean, J.: Efficient Estimation of Word Representations in Vector Space. In: Proceedings of the International Conference on Learning Representations (2013).
47. Jurisch, M. and Igler, B.: RDF2Vec-based Classification of Ontology Alignment Changes. (2018).
48. Breiman, L.: Random Forests. In: Machine Learning, vol. 45, issue 1, pp. 5–32. doi: 10.1023/A:1010933404324
49. Volkovs, M., Wei Yu, G., and Poutanen, T.: Content-based Neighbor Models for Cold Start in Recommender Systems. In: Proceedings of the Recommender Systems Challenge (2017). doi: 10.1145/3124791.3124792
50. Sandulescu, V. and Chiru, M.: Predicting the future relevance of research institutions – The winning solution of the KDD Cup 2016. (2016).
51. Federal Research Center Computer Science and Control of Russian Academy of Sciences. Available at: http://hhpcc.frccsc.ru (accessed 09/12/2018)
52. Nobarian, M. and Derakhshi, M.: The Review of Fields Similarity Estimation Methods. In: International Journal of Machine Learning and Computing **2** (2012). doi: 10.7763/IJMLC.2012.V2.200
53. Znamenskij, S.: Stable assessment of the quality of similarity algorithms of character strings and their normalizations. In: Program systems: theory and applications **9** (39), 561–578 (2018). doi: 10.25209/2079-3316 -2018-9-4-561-578
54. Eckert, K., Meilicke, C., and Stuckenschmidt, H.: Improving Ontology Matching using Meta-level Learning. In: The Semantic Web: Research and Applications, pp. 158–172 (2009). doi: 10.1007/978-3-642-02121-3_15
55. Euzenat, J., Guégan, P., and Valtchev, P.: OLA in the OAEI 2005 alignment contest. In: Proceedings of the K-CAP 2005 Workshop on Integrating Ontologies (2005).

Mathematical Physics Problems: Thesaurus and Ontology

O.M. Ataeva[1], V.A. Serebryakov[1], N.P. Tuchkova[1]

[1]Dorodnicyn Computing Centre, Federal Research Centre “Computer Science and Control” of the Russian Academy of Sciences, 40 Vavilov St., Moscow

Abstract. The work is devoted to the study of knowledge representation in the subject area “Equations of mixed type in the sections of mathematical physics”. Such comprehensive resources as Wikipedia, claim to be encyclopedic knowledge, but cannot provide informational support for in-depth research. This is a field of activity for specialists in specific areas of knowledge.

The thesaurus and ontology are considered as a formal means of describing the subject area. This approach takes into account the peculiarities of knowledge representation in the domain, namely the presence and use of formulas as independent objects. Consideration of the domain features leads to a reduction in search noise and a reduction in search time within the framework of the constructed library. The use of thesaurus and ontology in the design of a digital semantic library is considered.

Keywords: mathematical physics, thesaurus, ontology, formulas and texts, digital libraries

1 Introduction

The issues with information representation of mathematical knowledge in digital space are directly related to the logical framework organization set within the mathematical subject domains. The research is prompted by spreading digital information representation in mathematical sciences and explained by the effect mathematics now have on the developed countries’ economy [1]. Many papers highlight that the way mathematical branches are presented on the Internet is significant, for the future of science in general as well [2]. The analysis of digital mathematical resources shows that it is essential to generate relevant thesauri to build digital information images of mathematical subject domains relying on scientific knowledge ontologies. Such comprehensive resources as Wikipedia, Wiktionary claim to be encyclopedic knowledge, but cannot provide informational support for in-depth research. This is a field of activity for specialists in specific areas of knowledge.

They do not reveal the essence and mathematical meanings of concepts, in contrast to a special tool – the thesaurus of the subject area, where emphasis is placed on semantics and the use of the term in linked publications. The originality of the proposed work is that the bibliographic base is associated with persons, and there is an opportunity for authors and experts to replenish information by linking it to existing terms and complement the list of terms, including in Russian.

The main condition for research information support is the provision of up-to-date information on achievements, confirmed by publications of professionals. Naturally, the mathematical resources on the Internet need such professional information support. Known databases of scientometric data of publishing houses partially perform this function in the section of actual publications. However, exponentially growing number of publications, it complicates the search, requires expensive and time consuming. In this case, every professional is interested in having a collection on a certain topic. It is possible to organize the creation of such a collection technologically by using thematic thesauri and the mechanism of their replenishment, as proposed in this paper. Feature of modern digital representation of the data makes it possible to move the centuries created thesauri and ontologies in thematic databases and thereby ensure their search for the completion and updating.

Another feature of mathematical subject domains is explained by the fact that mathematical statements in natural language are better expressed as mathematical equations. When building information images for branches of mathematical physics it is crucial to consider the listed aspects, specifically, rely on representative dictionaries which define terms and formulas as a background for information retrieval thesaurus for the subject domain.

In [12], a thesaurus was presented for the subject domain “ordinary differential equations” and now its extension to the domain of “partial differential equations” is being developed as part of a common mathematical resource on “equations of mathematical physics”. Numerous studies of specialists, such as V. A. Steklov [13], V. S. Vladimirov [14], R. Curant, D. Gilbert [15], A. N. Tihonov, A. A. Samarsky [19], A. G. Sveshnikov, A. N. Bogolyubov [20], M. M. Smirnov [16], A. V. Bitsadze [17], V. A. Ilyin, E. I. Moiseev [18] and other classics of mathematical analysis and differential calculus allow to establish paradigmatic connections of concepts and formulas to use them as lexico-semantic data arrays for presentation and search in mathematical information resources.

This paper describes the creation of the ontological model of a thesaurus for some problems of mathematical physics within the framework of the terminology of the semantic library LibMeta and its use in the tasks of searching and navigating through its resources. At the first stage, a series of related dictionaries for individual equations is combined into a thesaurus. It is incomplete and therefore a means is proposed for replenishing it with the inheritance of previous knowledge from available data from open sources. It turns out a non-standard resource, but it reflects the state of modern research.

2 Thesaurus Description

Mathematical physics deals with mathematical models of physical phenomena [3]. It relies on mathematical methods to build and study the models [3–5]. The methods of mathematical modelling enable us to solve mathematical problems applying equations of differential calculus [3], [4]. Each equation establishes a correlation between mathematical model and physical phenomena. The topics to be described are as follows: *problems of mathematical physics, modeling methods, equations, methods of solutions, solutions and their analysis*.

The thesaurus was formed by analyzing the original works of classics of mathematical analysis and differential calculus, and a representative list of articles was organized for that purpose. The problem of defining paradigmatic relations between definitions of certain fields in mathematical physics is brought to attention along with outlining the hierarchical relations between the terms that can be used when searching on the mathematical resources along with additional classification parameters set in secondary documents.

It is possible to study a streamlined scheme (Fig. 1) to offer a step-by-step description to problems of mathematical physics starting with the name of physical/technical process and ending with solutions to develop data layout in this domain.

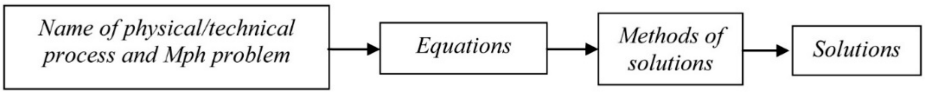


Fig. 1. The relation scheme in mathematical physics subject domain

Given that equations of mathematical physics, as a subject domain, cover a huge amount of research, the paper focuses on *physical processes identification*, as the pillar of mathematical models, and *partial differential equations* terminology with examples of *mixed type equations*.

2.1 MPh Problems

When describing the mathematical set of MPh problems, we consider it a hierarchy as provided in Fig. 2 that follow the logic of the domain. The graphical representation is one of the ways to describe hierarchical relation of problems in mathematical physics [6].

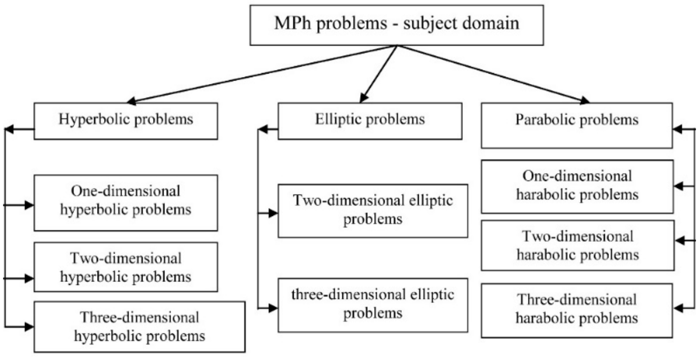

MPh problems - subject domain
Hyperbolic problems
Elliptic problems
Parabolic problems
One-dimensional hyperbolic problems
Two-dimensional hyperbolic problems
Three-dimensional hyperbolic problems
Two-dimensional elliptic problems
three-dimensional elliptic problems
One-dimensional harabolic problems
Two-dimensional harabolic problems
Three-dimensional harabolic problems

Fig. 2. The MPh problems relation scheme

Such structure provides a topic-related distribution within the section describing the problems of the MPh.

2.2 Partial Differential Equation

Let us note the features related to partial differential equations that should be added to the thesaurus:

- equation scope, as well as the material object of the physical process;
- summary on equation properties;
- researchers’ surnames, authorship, named equations;
- specific and associative equation formulas;
- synonyms for the terms.

In different domains mixed type equations can be classified as hyperbolic, parabolic and elliptic [7]. Fig. 3 graphically shows hierarchical links of second order PDE with two independent variables.

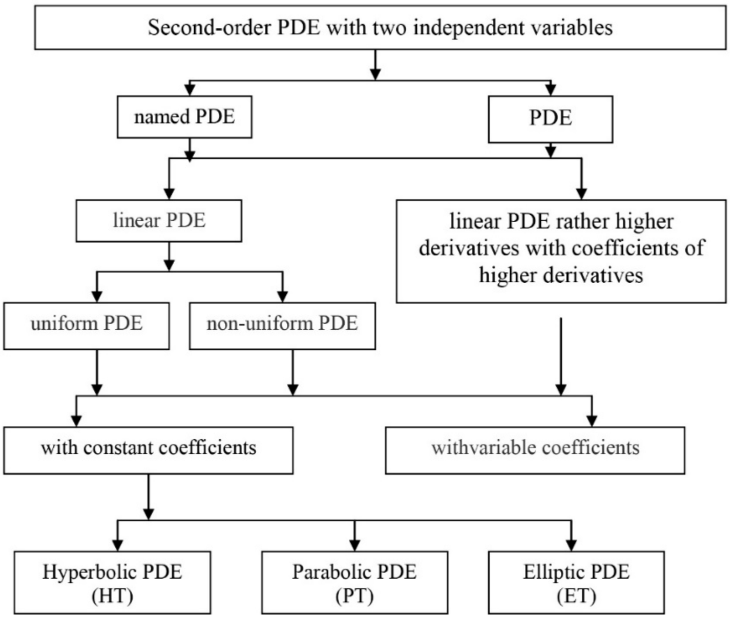

Second-order PDE with two independent variables
named PDE
PDE
linear PDE
linear PDE rather higher derivatives with coefficients of higher derivatives
uniform PDE
non-uniform PDE
with constant coefficients
withvariable coefficients
Hyperbolic PDE (HT)
Parabolic PDE (PT)
Elliptic PDE (ET)

Fig. 3. Graphic chart representing linear/linear rather higher derivatives second order PDE with two independent variables

2.3 Thesaurus Structure

While developing the thesaurus one of the main objectives is to develop its structure considering the characteristics of the domain. The structure consists of thematic sections, sets of links between the elements of the thesaurus, the structure of the thesaurus articles.

Thus, the basic version of the thesaurus includes the following main thematic sections:

- problems of mathematical physics
- equations of mathematical physics,
  - partial differential equations,
  - equations of mixed type.

The analysis of the domain/subject area revealed the need to allocate the following categories for mentioned sections:

- Type of problem (elliptic, hyperbolic, parabolic);
- Dimension of the problem (one-dimensional, two-dimensional, three-dimensional);
- Type of equations (named, nominal);
- Homogeneity of equations (linear uniform, linear non-uniform, …);
- Types of equation coefficients (with variable coefficients, with constant coefficients, …);
- Types of equations (elliptic, hyperbolic, parabolic).

Based on these categories, the following link are established:

- Task type – task dimension;
- Type of equations – type of equations;
- Uniformity of equations – types of coefficients in equations.

The following thesaurus terms are also reflected:

- Hierarchical: genus, species;
- Horizontal: synonyms, associations.

In addition to the main term categories in the thesaurus, it is necessary to introduce additional categories that support generation of various links tied to objects that are not explicitly reflected in the thesaurus but are necessary for the completeness of description. Such objects include Authors and References. In order to implement these features, thesaurus conceptual structure provides relevant set of links to describe references, authors, etc.

- References – introduced to describe references to literature that contains in-depth information about a concept of the thesaurus;
- Author – introduced for designating the author and the author's term for a concept.

The mentioned hierarchically and horizontally linked categories form a conceptual model of the domain.

The conceptual model of the thesaurus is thus reflecting the following:

- means to define concepts;
- method for defining concept synonyms;
- a list of conceptual properties and attributes;

- object category;
- composition of objects of each category.

Thus, structurally, the concept of the thesaurus includes the following elements:

- alphanumeric code of the concept;
- concept descriptor;
- non-descriptor – concepts' synonyms;
- thematic section of the concept;
- symbolic representation of the concept formula;
- list of links to other concepts;
- text additions (comments, notes, help);
- a list of references for the concept;
- authors of the concept.

Given the essence of the structural description there is a need to include various lexico-semantic categories as follows:

- type of equation: one-dimensional, two-dimensional, three-dimensional;
- type of equation: hyperbolic, parabolic, elliptic;
- types of coefficients: variables, constants;
- etc.

Thesaurus Ontological Model

Basics of LibMeta Information Model

The paper develops the resource for the topic in math on "mathematical physics problems", index of "MPh mixed type equations" and its integration in LibMeta [8].

LibMeta is an information system that implements a set of features that are necessary to work with the content of a prospective semantic library. LibMeta is a special electronic library management system (ELMS). LibMeta library is a storage of structured and diverse data with the ability to integrate it with other data sources that meet the requirements for sources within Linked Open Data [21]. Also the possibility of specifying its content by defining the subject area is presented.

The versatility of the system's content is based on the set of concepts which represents the LibMeta informational content model: *information resource* and *information object* that define a resource instance. The *information resource* is the basic descriptive unit of library content, and the *information object* represents instance of *information resources*. Each of them has its own unique LOD identifier. In fact, the semantic meaning of the *information resource* is equivalent to the concept of *ontology class* with discrepancies in description. The structure of the description of information objects is determined by the concepts of an *attribute* and a *set of attributes* that are defined in the description of the corresponding resource. The *attribute* is an element of a resource property description, and the *set of attributes* is defined as a collection of different attributes. Attribute types are as follows: *file*, *object*, *numeric*, *text*, *string*.

The model of the thesaurus at full compliance with ISO 25964 [9]. The model described by this standard supports multilingual thesauri and other types of dictionaries.

The standard contains recommendations for establishing and maintaining mutual correspondence between several thesauri or between thesauruses and other types of dictionaries used in information retrieval. This standard is also compatible with the SKOS model, which offers a way to present thesauri in the Internet.

The standard suggest rationale for using the following concepts: hierarchical relationship, horizontal relationship, term, thesaurus, concept, thematic group, terms, descriptor (or preferred term for a concept), non-descriptor (a set of terms that are synonymous with a descriptor). Earlier in [10] we provide analysis of the basic entities of the ontology of the model, which forms the base of the ontological thesaurus model [10], presented in current publication.

The base model of thesaurus is designed such way that the concepts of this thesaurus are related to the concept of an information object from the LibMeta content description model and allow for concept association with any type of resource present in the library. Description of the ontological content model of the library allows you to describe additional types of resources like *Authors* and *References* and link them with the thesaurus to support them.

2.4 Basic Thesaurus Model Expansion

In order to add the basic thesaurus version description [10], namely the structure of the thesaurus concepts, the system supports a class hierarchy for additional concept attributes. It includes subclasses of the ResourceAttribute superclass that add the description of the concept structure that corresponds to a certain thesaurus with the following concept values:

- ThesaurusAttributeText – presented as a text;
- ThesaurusAttributeTaxonomy – presented in the form of item of a particular classifier or dictionary;
- ThesaurusAttributeString – presented as a string;
- ThesaurusAttributeObject – presented as a certain content library information object;
- ThesaurusAttributeNumber –presented as a number;
- ThesaurusAttributeHref – presented as a link;
- ThesaurusAttributeFile – presented as a file;
- ThesaurusAttributeConcept – presented as other thesaurus concepts (it defines relationships between concepts implicitly supported in the system).

Each of these classes is compliant with OWL-supported inheritance paradigm, that contains the properties ascribed to the ThesaurusAttribute superclass.

The ThesaurusAttributeSet class contains the thesaurusAttributes property, which in turn contains many instances of the above listed classes that define the *Concept* class structure of a thesaurus.

The *Thesaurus* class is linked to ThesaurusAttributeSet through properties mediator thesaurusAttributeSet.

Such standardized modeling allows LibMeta to easily adjust the system to any subject area.

The description of the thesaurus “The MPh Mixed Type Problems” based on initial ontology version can be broadened by terms of an extended model, in order to further extend the article structure of this thesaurus, provided the following attributes are added:

- *comment;*
- *note;*
- *help;*
- *references;*
- *authors.*

2.5 Three Levels of Thesaurus Model

The *Comment* and *Note* attributes represent **ThesaurusAttributeText** class attributes, with Help being **ThesaurusAttributeString** class attribute, while References and Authors are instances of the **ThesaurusAttributeObject** class. Combined they are a set of thesaurus attributes.

Next, we analyze a three-level representation of the subject area thesaurus within the LibMeta library.

In order to be able to use the thesaurus of a specific subject area, the following sequence should be followed when constructing a semantic library within LibMeta system.

1. Based on the introduced model, a set of information resources used in the library is given. It is necessary to provide descriptions of the content of the future library in terms of the proposed model.
2. The structure of the thesaurus is finally set up. On the basis of certain classes the respective links between terms are set, the term description is expanded if necessary, the links with the system resources are determined as well.
3. According to the definition of collections, a module is implemented, within which collections are created and filled.

After completing these steps, we form a domain information model described in terms of the ontology of the semantic library introduced above. At the same time, if the newly introduced concepts are instances of the designated resources at the first level, then when filling the library, we use them as classes to describe the data. Dividing instances into classes is *metamodeling*. Even though the semantics of OWL 2 ontologies that is used to describe ontologies does not allow such metamodeling, this language limitation is bypassed with a syntax trick known as *punning*. This means that when an instance identifier is present in a class axiom, it is seen as a class, and when the same identifier seen in a separate statement, it is treated as an instance.

While describing a specific subject area in terms of the proposed semantic library ontology, we, in fact, construct a three-level ontology, in which instances of the first level are high-level concepts, with the second level containing concepts of a specific subject area. When uploading data to the ontology we use the first level terms to define the third level classes.

3 Searching for Mathematical Publications through Thesaurus Links

The use of mathematical formulas is novelty, yet due to the recent software development progress they see their use. The LibMeta system [8, 10] implements a comparison of formulas starting with its denotation. This renders possible to include symbolic expressions in search queries.

Using the domain/subject area of mathematical physics and related fields as an example, we can see how expanding a thesaurus-based query can improve search results. Let us consider Tricomi Equation as an example and highlight the advantages of adjusting the query with formulas.

The concept of “Tricomi Equation” is associated with synonymously tied to similar equations of different kind. For each of these equations there is a symbolic notation and a “Tricomi Equation” TEX formula notation. The formulas in this example are actually synonymous. For each of the formulas are also references to the source materials with a specific mathematical record. These “Tricomi Equations” are used for indexing and searching publications in databases on mathematical subject areas. Thus, if you select one of the entries you will find works about the “Tricomi Equation” for the other variants of different scientific schools and paradigms.

If necessary, you can expand the references section and expand the LibMeta library thesaurus. The formulae-synonyms links will provide references and respective authors data, search for which in this case was not directly undergone. Thus, the process of updating the thesaurus for the subject area is realized through the links. As a result, the LibMeta library will have new data on publications, and a user will receive a new list of publications for the “Tricomi Equations”. By requesting this topic, the user will also receive complete information about the semantic links for the searched formula, which will include links to the formula-synonyms, which is especially important for the experts. Thus, we are talking about how far the search result corresponds to the information need of the user (pertinence property). The scope of the search and query can be quite narrow, but can be supplemented with a formula and the result will be complete in the sense of pertinence. This is one of the necessary properties of the usefulness of the information system and the success of utilizing it, which was noted by the founder of the term “information retrieval” Calvin Mooers [11].

4 Conclusion

Specific developments in the mathematical subject areas representation still in the focus modern interdisciplinary research papers. The proposed information resource implements the combining specified semantic relations, symbolic language of formulas and systematic representation of mathematical resources (publications with secondary information, indexed in accordance with the thesaurus) in the digital library. This approach corresponds to the current trends in the development of information technology and allows for expansion of the mathematical ontological domain in the digital space. The choice of the subject area with the lack of representation in literature and descriptor dictionaries, and the digital library approach renders this study relevant. Also a tool is being created for the formation of Russian-language content on this topic. The discovered solution for mixed-type problems does not limit the other MPh problems ontology generality. The functions of the digital library LibMeta allows for updating this subject area as the development of relevant resources (thesaurus, dictionaries, term lists) goes on.

It takes into account all the properties of a mathematical text and a combination of symbolic (formula) and textual information. And this is the main feature of the proposed ontological model of a thesaurus for equations of mathematical physics. As a result, all properties of a mathematical text are taken into account when searching and indexing publications. Now the thesaurus of mathematical physics is presented in terms of the ontology LibMeta includes about 100 concepts and more than 200 terms. The ontology language is OWL, the syntax is RDF/XML.

Acknowledgements

The study was supported by Russian Foundation for Basic Research, projects Nos 17-07-00217, 18-00-00297, 17-07-00214.

References

1. Bond, Ph.: The Era of Mathematics. An Independent Review of Knowledge Exchange in the Mathematical Sciences. Professor Philip Bond. https://epsrc.ukri.org/newsevents/pubs/era-of-maths/, 02.12.2018
2. Mayans, R.: The Future of Mathematical Text: A Proposal for a New Internet Hypertext for Mathematics. Journal of Digital Information [S.l.] **5** (1) (2006). https://journals.tdl.org/jodi/index.php/jodi/article/view/128/126, 02.12.2018
3. Mathematical encyclopedia. Chief Editor. I.M. Vinogradov. Moscow: Soviet encyclopedia. 1979. 1104 p.
4. Tihonov, A.N. and Samarskij, A.A. Equations of mathematical physics. M.: Publisher Nauka, 1972. 736 p. (ru)
5. Sveshnikov, A.G., Bogolyubov, A.N., and Kravcov, V.V.: Lectures on mathematical physics: Textbook. Moscow: Publisher MSU, 1993. 352 p. (ru).
6. Moiseev, E.I., Muromskij, A.A., and Tuchkova, N.P.: Internet and mathematical knowledge: representation of equations of mathematical physics in the information retrieval environment. M: Publisher MAKS Press, 2008. 80 p. (ru).
7. Smirnov, M.M.: Equations of mixed type. M: Publisher Nauka, 1970. 296 p. (ru).
8. Serebryakov, V.A. and Ataeva, O.M.: Information model of the open personal semantic library LibMeta. Proceedings of the XVIII Russian Scientific Conference "Scientific Service on the Internet". Novorossijsk, 19–24 September 2016. Keldysh Institute of Applied Mathematics (Russian Academy of Sciences). P. 304–313 (ru).
9. ISO 25964 thesaurus schemas. http://www.niso.org/schemas/iso25964

10. Ataeva, O.M. and Serebryakov, V.A.: Ontology of the digital semantic library LibMeta. Informatics and its applications **12**, 2–10 (2018) (ru).
11. Muromskij, A.A. and Tuchkova, N.P.: On the ontology of the addressee in the mathematical subject field. Electronic Libraries **21** (6), 506–533 (2018) (ru).
12. Moiseev E.I., Muromskij A.A., Tuchkova N.P. Thesaurus information retrieval in the subject area: ordinary differential equations. M.: Publisher MAKS Press, 2005. 116 p. (ru)
13. Steklov, V.A.: The main tasks of mathematical physics. Moscow: Publisher Nauka, 1983. 433 p. (ru).
14. Vladimirov, V.S.: Equations of mathematical physics. Ed 4. Moscow: Publisher Nauka, Main editorial office of the physical and mathematical literature, 1981. 512 p. (ru).
15. Kurant, R. and Gil'bert, D.: Methods of mathematical physics, v. I, II, M.: Publisher Gostekhizdat, 1951 (ru).
16. Smirnov, M.M. Equations of mixed type. Moscow: Publisher Nauka, 1970. 296 p. (ru).
17. Bicadze, A.V.: Equations of mathematical physics. Moscow: Publisher Nauka, 1982. 336 p. (ru).
18. Il'in, V.A. and Moiseev, E.I.: Nonlocal boundary-value problem of the second kind for the Sturm–Liouville operator. Differ. Equations **23** (8), 1422–1431 (1987) (ru).
19. Tihonov, A.N. and Samarskij, A.A.: Equations of mathematical physics. Moscow: Publisher Nauka, 1972. 736 p. (ru).
20. Sveshnikov, A.G., Bogolyubov, A.N., and Kravcov, V.V.: Lectures on mathematical physics: Textbook. Moscow: Publisher MSU, 1993. 352 p. (ru).
21. Bizer, C., Heath, T., and Berners-Lee, T.: Linked data: The story so far. Semantic services, interoperability and web applications: emerging concepts. IGI Global, 205–227 (2011).

Thesaurus and Ontology Building for Semantic Library Based on Mathematical Encyclopedia

O.M. Ataeva[1], V.A. Serebryakov[1], and E.K. Sinelnikova[2]

[1]Dorodnicyn Computing Centre, Federal Research Centre “Computer Science and Control”
of the Russian Academy of Sciences, 40 Vavilov St., Moscow
[2]Moscow State University, 1 Leninskie Gory St., Moscow

Abstract. The paper focuses on the task of constructing the ontology of a thesaurus and its filling for the mathematics-related sources. The data of the Mathematical Encyclopedia, the Soviet encyclopedic edition, is used to fill the thesaurus of the subject domain and outlines it terminologically. The encyclopedia was preprocessed to structure its content and isolate semantic links automatically. The results were incorporated into the semantic e-library and used as a thesaurus for its subject domain.

Keywords: mathematical enciclopedia, ontology, thesaurus, digital library, semantic library

1 Introduction

This paper covers the task of constructing an ontology of a thesaurus and its filling of the library for the resources of the semantic library devoted to mathematics. The Mathematical Encyclopedia makes the thesaurus of the subject domain and outlines it terminologically. The Mathematical Encyclopedia is a five-volume Soviet encyclopedic issue that covers mathematical subjects. This fundamental illustrated edition on all main branches of mathematics comprising more than 6,000 articles was published in 1977–1985 [1]. Later the encyclopedia was digitalized. The e-version entries pose as an unstructured text with formulas in the form of images, lack any references to the related articles of the encyclopedia or other sources, and do not indicate the branch of mathematics. The listed drawbacks make the encyclopedia unfit for usage by Internet-users within the framework of an e-library.

To successfully integrate the data from the Mathematical Encyclopedia and make it available for e-library users, it is essential to ensure high level article structuring. Initially, the articles featured only text without any metadata such as references to the related articles of the encyclopedia, articles from other knowledge bases, or indication of a particular branch of mathematics. Such marking is important as the e-library user values not only the articles, but also options to navigate the library, to search relevant materials and related data.

The Mathematical Encyclopedia was translated into English in 1987 with about 2,000 new articles added. As of today, the e-version of the translated encyclopedia is supported by the international publisher Springer (Luxembourg) and is available on the Internet [2]. The Encyclopedia of Mathematics entries feature TEX-format formulas which can be machine-processed, and references to the related articles of the encyclopedia. Each article has a matching MSC (Mathematics Subject Classification) index [3] which is used for classification based on the branches of mathematics. Together, these metadata provide the user with a variety of options for searching relevant articles and studying related topics.

The data from the Encyclopedia of Mathematics is significant for adding new metadata to the Mathematical Encyclopedia and then making it available to the Internet-users as an electronic reference resource. This resource can be incorporated into a semantic e-library or provided with its own web-interface. In this paper, we have opted for LibMeta information system as such a library.

2 Overview

The overview of the related sources in Russian included online resources comprising articles from Mathematical Encyclopedia or any other mathematical knowledge.

Mathematics Library

This is a resource, available at www.MathemLib.ru, covers the knowledge accumulated throughout the Soviet period in a form of books published in the USSR and was updated with current news articles [4]. Moreover, this resource allows users to access articles from the Mathematical Encyclopedia listed alphabetically. It features the letters A-E, С, У from Russian alphabet and letters L, N, P, S from the Latin alphabet. The articles pose as a simple text with formulas in the form of images. It still lacks the references to other encyclopedia articles, articles from the library or any other sources.

Math-Net.Ru

It is the project of Steklov Mathematical Institute of Russian Academy of Science. The authors describe the project as a modern information system which provides Russian and foreign mathematicians with various options for searching information on mathematical life in Russia [5]. This information resource deals with mathematical journals and users have access to the full-version issues. Certain journals require paid subscription. The articles posted on Math-Net.Ru website also lack references to the related materials as they are available for downloading in PDF format.

World Digital Mathematical Library (WDML)

The researchers from Kazan Federal University cover the idea to create the mentioned resource in details in their papers [6]. The key purpose of WDML is to combine the digital versions of the entire corpus of mathematical academic literature within the distributed system of interrelated repositories, including both current resources and historical ones. Still, the ultimate goals of the project have not been attained.

The overview of the aforementioned and some other open mathematical resources in Russian has shown that there is an issue related to data structuring. The majority of the information resources do not allow users to build queries to the database, and observe relations between mathematical concepts, articles and authors.

3 Mathematical Encyclopedia Arrangement

To build information and reference resource based on the Mathematical Encyclopedia data, we have worked on including information on the branches of mathematics the articles belong to, placing cross-references between the articles, defining the article-related machine-readable formulas. The designed resource allows building queries to its contents as well as further integration with other databases within Linked Open Data space.

To attain the goal, we have addressed the data comprised in the English version of the encyclopedia – Encyclopedia of Mathematics. In particular, MSC section indexes and TEX-format formulas from the articles were employed. To employ them it is essential to match the Mathematical Encyclopedia articles with their translations from the Encyclopedia of Mathematics. Cross-references between the articles were generated by applying semantic annotation methods [7, 8]. A data model have been developed which helps build queries and ensure establishing links with other open sources.

Thus, the objective involved the following steps:

1. Develop the information resource data model based on the Mathematical Encyclopedia.
2. Match the Mathematical Encyclopedia articles with their translations in Ecyclopedia of Mathematics.
3. Add annotations, i.e. references to other encyclopedia entries, to the Mathematical Encyclopedia articles.
4. Attach MSC indexes to encyclopedia entries as in Encyclopedia of Mathematics.
5. Match articles with TEX-format formulas derived from the following Encyclopedia of Mathematics entries.
6. Correlate the entries with the list of similar articles.
7. Ensure building queries to the Mathematical Encyclopedia.
8. Ensure further integration of the developed resource with other sources.

3.1 Mathematical Encyclopedia Data Model

We have developed a data model that shows the relations between Mathematical Encyclopedia articles, terms, formulas, annotations and MSC sections. Fig. 1 illustrates the ontology classes and their relations.

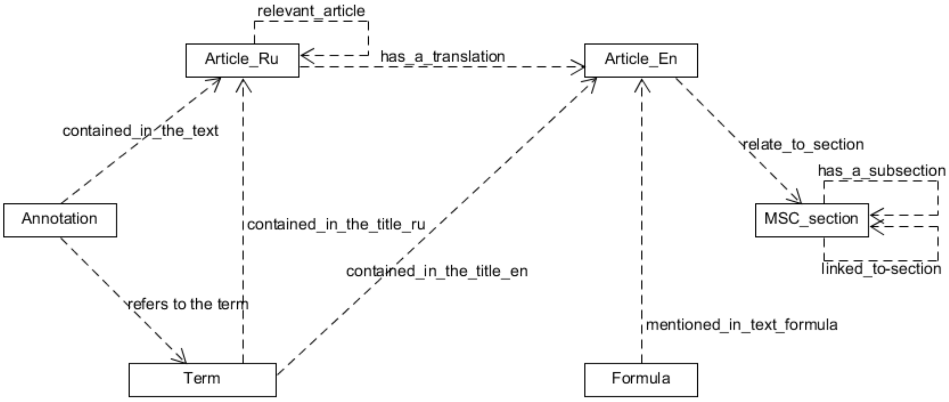


Fig. 1. Data Model

The ontology involves the following classes and relations:

1. ***Article_En* –** Encyclopedia of Mathematics article in English.
2. Object Properties: *relate_to_section* – indicates the MSC rubricator section which the article is related to.
3. ***Article_Ru*** – Mathematical Encyclopedia (ME) article in Russian.
4. Object Properties: *has_a_translation* – a reference to the item on the ***Article_En*** class which holds the translation of the Russian article; *relevant_article* – indicates the related article from ME.
5. ***Term*** – a mathematical concept which poses as a title to a certain article from ME or Encyclopedia of Mathematics.
6. Object Properties: *contained_in_the_title_ru* – indicates the ME article the title of which features the term; *contained_in_the_title_en* – indicates the Encyclopedia of Mathematics article the title of which features the term.
7. ***Annotation*** – an annotation found in the ME article text.
8. Object Properties: *contained_in_the_text Содержится_в_тексте_аннотация* – a reference to the ME article, the text of which features the annotation; *refers_to_the_term* – indicates the term being annotated.
9. Data Properties: *beginning_of_the_link* – the number of the word in the article which marks the beginning of the annotation*; end_of_the_link* – the number of the last word in the article included in the annotation.
10. ***Formula*** – a TEX-format formula included in the Encyclopedia of Mathematics article.
11. Object Properties: *mentioned_in_text_formula* – a reference to the Encyclopedia of Mathematics article which features the formula.
12. ***MSC_section*** – the section of the MSC mathematical rubricator.
13. Object Properties: *relate_to_section* – indicates the related MSC section, *has_a_section* indicates the close MSC section, *linked_to_section* – indicates the “parent” MSC section.

The developed model allows building queries to the Mathematical Encyclopedia. Let us consider some of these queries.

1. Find *ME Articles* that feature the *Term*.

The following is an example of a SPARQL query for the russian term “Multi-connected domain”:

```
SELECT  ?article_name
WHERE {
   ?annotation math_enc: contained_in_the_text> ?ru_article.
   ?ru_article rdfs:label ?article_name.
   ?annotation math_enc: refers_to_the_term> ?term.
   ?term rdfs:label ?label.
   filter contains(?label,"Многосвязная область")}
```

2. Find *ME* Articles that feature the *Formula*.

The following is an example of a SPARQL query for the formula $f\,(x_1, ..., x_n)$=0, presented in TeX notation

```
SELECT ?article_title
WHERE {
?ru_article math_enc: has_a_translation> ?en_article.
?ru_article rdfs:label ?article_title.
?formula math_enc: mentioned_in_text_formula> ?en_article.
?formula rdfs:isDefinedBy ?tex.
filter contains(?tex,"f(x_1,\\ldots,x_n)=0,\\tag{*}") }
```

3. Find *ME Articles* that relate to the *MSC Section*.

The following is an example of a SPARQL query for the *MSC – 60-XX* section – “Probability theory and stochastic processes”

```
SELECT  ?article_title
WHERE {
?ru_article math_enc: has_a_translation> ?en_article.
?ru_article rdfs:label ?article_title.
?en_article math_enc: relate_to_section> math_enc:60-XX>.
```

4. Show *ME Articles* that are relevant to the selected *ME article*.

The following is an example of a SPARQL query for a selection of articles relevant to the article “*Multi-connected domain*”

```
SELECT  ?article_title
WHERE {
math_enc:Article_Ru_Multi_connected_domain    relevant_article
   ?see_also_article.
?see_also_article rdfs:label ?article_title.}
```

5. Show *ME Articles* that the annotations in the selected article reference to.

Below is an example of a SPARQL query for a selection of article links to which are found in the article “*Multi-connected domain*” in Russian version

```
SELECT  ?related_article_title
WHERE {
?annotation  math_enc:  contained_in_the_text  math_enc:  Arti-
   cle_Ru_Multi_connected_domain.
?annotation math_enc: refers_to_the_term ?termin.
?termin math_enc: contained_in_the_title_ru ?related_article.
?related_article rdfs:label ?related_article_title.}
```

6. Show *Formulas* related to the *Term*.

Below is an example of a SPARQL query for selecting formulas related to the term “*Pole*” in russian version

```
SELECT ?tex
WHERE {
?term  math_enc:Соде contained_in_the_title_ru ?ru_article.
?ru_article math_enc: has_a_translation ?en_article.
?formula math_enc: mentioned_in_text_formula ?en_article.
?formula rdfs:isDefinedBy ?tex.
?term rdfs:label ?label.
 filter contains(?label, «Полюс») }
```

7. Show *Formulas* related to the *MSC section*.

Below is an example of a SPARQL query for selecting formulas related to the *MSC section – 41-XX – "Approximations and expansions*"

```
SELECT ?tex
WHERE {
?formula math_enc: mentioned_in_text_formula ?en_article.
?en_article  math_enc: relate_to_section math_enc:41-XX.
?formula rdfs:isDefinedBy ?tex.}
```

Queries 1, 2 and 3 ensure the search through the Mathematical Encyclopedia articles based on such criteria as mathematical term, formula, and branch of mathematics. Query 4 and 5 show the relation between Mathematical Encyclopedia articles. Queries 6 and 7 is meant for observing the relations between formulas and mathematical terms, branches of mathematics.

Still, the listed queries do not explain the advantages of the ontological data model over other models concerning the current objective. One of the key distinctions of the data model from, for instance, the relational models is that the ontology can be processed by an inference engine which helps find out the relations between any two subjects within a model. Thus, the number of queries to the developed model is not limited to the aforementioned ones. Consider examples of queries that allow us to define new information templates based on existing information. Queries of this kind are also called rules.

8. Display formulas for the Russian article.

```
CONSTRUCT {?formula math_enc: mentioned_in_text_formula ?ru_article}
WHERE {
?ru_article math_enc: has_a_translation ?en_article.
?formula math_enc: mentioned_in_text_formula ?en_article}
```

9. Display MSC sections for Russian articles

```
CONSTRUCT {?ru_article math_enc: relate_to_section ?msc}
WHERE {
?ru_article math_enc: has_a_translation ?en_article.
?en_article math_enc: relate_to_section ?msc. }
```

Moreover, one of the objectives of the present paper is to ensure further linking of the developed data model to other knowledge sources within Linked Open Data. The structure of the relational models is rigidly fixed which makes the linking of two deferent models sophisticated and irresolvable in general. At the same time, ontologies can be linked relatively simply with the use of such OWL language properties as *owl:sameAs*, *owl:equivalentTo*.

Thus, the designed data model can serve as a frame for further information resource development based on the Mathematical Encyclopedia.

4 Thesaurus and Ontology Building for Semantic Library

Addressing the concept model described in the paper [9], as well as the ideas of Semantic Web and Linked Open Data, we have developed LibMeta *personal open semantic digital library* which supports the users' work with libraries' digital resources and collections within a particular scientific subject domain that is terminologically outlined by a thesaurus.

Apart from this, the key requirements to the system content, specifically *versatility, structure, adaptability* ensure support of the custom metadata repository for the objects as well as expanding information resource set. *Versatility* allows describing types of the system's resources and its objects regardless the subject domain or the users' scope of interest. The description *structure* supports relations between different external and internal resources relying on the LOD principles. The resource description *adaptability* allows adding new properties and links within the system development process and ensures user interface customization to reflect perspective changes. In fact, LibMeta makes the design of scientific knowledge space functional within the framework of a library.

The mathematical articles served as the subjects of the developed library. *Authors* and *Publications* were taken as examples of the resource types respectively. We have defined the set of attributes for each resource type within the minimum property set based on Dublin Core for publications and FOAF to describe the authors.

In fact, the concepts Authors and Publications serve as the items of the Information Resource class, which is defined as the basic unit of semantic library content. As each resource has a set of attributes, each of these items has its own assigned set of attributes that are described in the system. The set of attributes consists of the following elements: *title in the original language, title in Russian, surname, name, patronymic name, email address, date of birth, abstract, ID, author, occupation, publication type, place of birth, biography, description, additional title, language.*

4.1 Ontology of "Mathematical Encyclopedia" Thesaurus

The LibMeta thesaurus model is build to meet the standard, the ISO 25964 standard in particular [10]. This standard defines the thesaurus as a set of terms that are related by their respective links (relations).

The description of the mathematical encyclopedia in terms of the concepts of the basic version includes such concepts as *Thesaurus, Concept, Term, HierarchicalRelation, FamilyRelation.* The Mathematical Encyclopedia description in the terms of Lib-

Meta ontology concepts can be additionally expanded. The attributes added are as follows: *formula, person, UDC code, MSC code, reference (to the English version of the concept)*.

The attributes *reference* serves as the items of ***ThesaurusAttributeHref*** class, *formula, person* serve as the items of ***ThesaurusAttributeObject*** class. At the same time they make up the attribute set for the thesaurus, *UDC code, MSC code* are the items of ***ThesaurusAttributeTaxonomy*** class.

The concept structure of the Mathematical Encyclopedia lacks hierarchy as it is, still, due to the use of MSC codes related to the concepts we have managed to highlight related terms from certain branches of mathematics. We have derived the mentioned persons from the articles and linked concepts to the persons. Formulas were separately indexed and each concept, if possible, was matched to the set of respective formulas.

4.2 Mathematical Subject Area Features

To support the formula search within the sub-system, the concept *Formula* has been introduced and it helps store the original formula line from the resource it was derived from. The line might be featured in the following formats: Content MathML, Presentation MathML, LATeX. If needed, the number of formula representation types within different notations can be easily expanded. The concept *Formula* is related to *Information Objects* and *concepts* of the thesaurus. Thus, we can always build a network of formula relations with other system information objects and thesaurus concepts. Each formula can be updated with key words. Key words might be placed either by a system expert, or be added when they are derived automatically along with a formula from its resource, as well as be filled with key words from related objects.

5 Conclusion

The developed information resource allows studying Mathematical Encyclopedia articles, their relations, ensures their categorization regarding the branches of mathematics. The resource has the property of replenishment: the developed mechanisms can be applied to the new data to include it in the encyclopedia.

The further studies might address the development of the semantic article annotation. In particular, the researches might dwell on pseudonym support for the concepts. Another possible direction involves studying the correlation between MSC sections and other rubricators, for instance, UDC. If successful, it would be sufficient to link the ontology class that matches MSC sections, to a new class for UDC sections, so that we could categorize encyclopedia articles and related resources using a new rubricator.

The present study allowed us to employ a considerable amount of knowledge stored in the Mathematical Encyclopedia and then pass in on to the wide spectrum of amateur-users and experts in the mathematical field, which is particularly significant in the lack of open access to the similar resources.

The means of LibMeta system helped define the relations to the terms of Mathematical Encyclopedia for each publication based on its title, abstract and key words. This allowed us to carry out an additional thematic division of publications within the subject domain. To some degree, such linkage helped find the articles related to different branches of mathematics and arrange them in collections based on the thesaurus and placed MSC links. The study employed about 5,000 publications.

Acknowledgements

The study was supported by Russian Foundation for Basic Research, project 17-07-00214.

References

1. Mathematical Encyclopedia. https://ru.wikipedia.org/wiki/ Математическая_энциклопедия (ru)
2. Encyclopedia of Mathematics. https://www.encyclopediaofmath.org/index.php/Main_Page (21.11.2018).
3. Mathematics Subject Classification. http://msc2010.org/mediawiki/index.php?title=MSC2010 (04.12.2018)
4. Math Library. http://www.mathemlib.ru (04.12.2018)
5. Math-Net.Ru. http://www.mathnet.ru
6. Elizarov, A.M., Kirillovich, A.V., Lipachev, E.K., and Nevzorova, O.A.: Management of mathematical knowledge: ontological models and digital technologies. DAMDID/RCDL'2016, Ershovo, 11–14 oktyabrya 2016. P. 44–50 (ru).
7. Oren, E., Hinnerk Moller, K., Scerri, S., Handschuh, S., and Sintek, M.: What are Semantic Annotations? http://www.siegfriedhandschuh.net/pub/2006/whatissemannot2006.pdf (12.12.2018)
8. Le Hoaj, Tuzovskij, A.F.: Semantic annotation of documents in electronic libraries. News of Tomsk Polytechnic University **322** (5), 157–164 (2013) (ru).
9. Serebryakov, V.A. and Ataeva, O.M.: Information model of the open personal semantic library LibMeta. Proceedings of the XVIII Russian Scientific Conference "Scientific Service on the Internet". Novorossijsk, 19–24 September 2016. Keldysh Institute of Applied Mathematics (Russian Academy of Sciences), 304-313 (2016) (ru).
10. ISO 25964 thesaurus schemas. http://www.niso.org/schemas/iso25964

DATA AND KNOWLEDGE MANAGEMENT

Intra-page Indexing in Generalized Search Trees of PostgreSQL

Andrey Borodin$^{1[0000-0002-1231-7959]}$, Sergey Mirvoda$^{2[0000-0002-4615-7164]}$, and Sergey Porshnev$^{2,3[0000-0001-6884-9033]}$

[1] Yandex, Khokhryakova str. 10, Yekaterinburg, Russia
amborodin@acm.org

[2] Ural Federal University, Mira str. 19,Yekaterinburg, Russia
{s.g.mirvoda,s.v.porshnev}@urfu.ru

[3] N.N. Krasovskii Institute of Mathematics and Mechanics of the Ural Branch of the Russian Academy of Sciences, S. Kovalevskaja str. 16, Yekaterinburg, Russia

Abstract. The Generalized Search Tree (GiST) is a framework for creating balanced tree access methods for data types, which can be provided as a database extension. This framework offers a big part of the access method's code but places some algorithmic limitations. One of these limitations is the concept that one tree node is a single page. In this paper, we propose changes to this limitation with additional intra-page indexing, based on the concept of skip tuples. This approach allows to increase of insert and update performance by the factor of 1.5 and opens new ways towards GiST API advancement. We implemented the proposed approach as a PostgreSQL core patch.

Keywords: Database · Indexing · GiST · Performance · PostgreSQL.

1 Introduction

PostgreSQL is one of the most advanced open source relational databases. And one of the prominent features of PostgreSQL is extensibility, the database is designed to be "hackable". PostgreSQL allows hackers to implement functions, describe data types, implement joins, hook internals and change functionality and features. Additionally, PostgreSQL has some levels of generalization to avoid writing boilerplate code. One of such parts is GiST.

Generalized index search tree (GiST) is an access method (AM) technique, which allows to abstract significant parts of data access methods structured as a balanced tree. Use of the GiST allows AM developer to concentrate on his own case-specific details of AM and skip common work on the tree structure implementation within the database engine, a query language integration, a query planner support, concurrency, recovery, etc.

The GiST was first proposed by J. Hellerstein in [12], further researches were undertaken by M. Kornacker [14, 13]. Later GiST was implemented in PostgreSQL with a large contribution by O. Bartunov and T. Sigaev [9]. Current

PostgreSQL GiST implementation accepts different trees as a datatype (and so-called operator class or opclass which identifies the operators to be used for the given index type). Opclass developer must specify 4 core operations to make a type GiST-indexable:

1. Split: a function to split a set of datatype instances into two parts.
2. Penalty calculation: a function to measure the penalty for the unification of two keys.
3. Collision check: a function that determines whether two keys may have overlap or are not intersecting.
4. Unification: a function to combine two keys into one so that combined key collides with both input keys.

Operations 1, 2 and 4 are responsible for the construction of the index tree, while operation 3 is used to query the constructed tree. In general case, tree construction can be seen as the serial insertion of a dataset into an index tree. Insertion is a recursive algorithm, executed for every node starting from the root of the tree. This algorithm searches within a node for an entry (also called downlink) with a minimal penalty of insertion of an item being inserted. For a chosen downlink the key is updated with operation 4, the algorithm is invoked recursively. If the algorithm is invoked on a leaf page it just places the item being inserted, if the node is overflown then the upward sequence of splits with operation 1 is started.

In terms of PostgreSQL operation 3 is called "consistency check" since it allows many different search strategies (intersection, inclusion, exclusion, adjacency, etc.). Also, PostgreSQL GiST requires one more operation – equality comparison called "same function". The data type can specify three optional operations: compress\decompress for compacting storage and distance for generalized kNN searches.

For example, if for operations 1–4 we pick rectilinear rectangles, we get regular R-tree [11], though many different indexing schemes are possible.

There are some differences between the original GiST and PostgreSQL implementation. For example, PostgreSQL implementation does not use stack recursion in algorithms and allows us to use in one index multiple different data types with different opclasses.

Currently, GiST is used by many geoinformation systems due to the PostGIS extension for PostgreSQL. PostGIS provides a versatile toolbox for map applications and other GIS functions. PostGIS uses GiST as the main indexing engine. The GiST is used in astronomic databases via pg_sphere extension. This extension is used to search for new stars by comparing observed signals using spatial join, implemented in GiST. Also, GiST is used for full-text search, for search in a set of ranges, for image similarity search and so on.

The PostgreSQL GiST implementation assumes one node of the generalized tree is a single page. It allows manipulating data that is larger than RAM: while some pages reside in RAM buffer, others reside in persistent memory. Our team is working on different improvements for spatial indexing [8]. Research of the GiST implementation showed us that there is a room for insert\update performance improvements by rethinking "one node is one page" concept.

In this paper we will review current GiST implementation concepts, this review will cover two main GiST procedures: inserts and scans. We do not consider tuple deletion since it is irrelevant in current PostgreSQL MVCC implementation. After this review, we will propose algorithmic and structural improvements and describe necessary code changes. The 3rd section will be devoted to a technique called skip tuples, which allows faster inserts into an index and skipping of tuple groups during scans. Also, we will describe motivation and details of Advanced Generalized Search – a framework for a showcase of GiST advances which can be deployed by current PostgreSQL users in their databases and is production ready. Then we will cover proposed changes with relevant experiments and their analysis. In this section, we also choose parameters necessary to tune skip tuples technique. The next section will cover the limitation of current algorithms and implementations. Later we will discuss related work focused on tackling similar problems with different approaches by other teams.

2 Current Implementation

PostgreSQL typically uses 8 Kb pages. The concept "one node is one page" means that fan-out of index tree typically varies between 100 and 1000 (from 80 to 8 bytes per tuple) with practically observable fan-outs around 250. This rough estimate shows that the GiST tree is usually low and guarantees path from the root to leaf in few disk reads. But in practical usage scenarios, GiST performance is not constrained by block device throughput. But the CPU operations often are the bottleneck.

The PostgreSQL GiST has two major parts of the functionality, dependent on page layout: index construction and index scan. In turn, index construction consists of insertion into an index and buffered build for an existing table. This work is focused on the insertion part. Despite buffered insertion also benefits from described technique and code changes, it's performance rarely is a bottleneck. Due to PostgreSQL MVCC implementation GiST updates are also represented by GiST inserts. Index scan also contains two interleaving parts: generalized search (regular scan) and generalized kNN ordering. In this work, we focused on a regular index scan. GiST concurrency is based on page-level latches and is not affected by this work. Recovery in GiST is updated, but its changes are straightforward and will not be discussed in detail.

2.1 Index Insertion

The insertion algorithm for any given index tuple searches the suited leaf page by descending from root page to leaf. Each step on the internal page invokes penalty calculation for each tuple on the page to find the best fitting subtree.

The only case when the insertion algorithm does not need to deal with every tuple is zero-penalty tuples (for a given inserted tuple) in the middle of a search.

But the existence of many zero-penalty tuples means overlapping of keyspace and inefficient index from the search point of view.

Overflown pages are divided into many parts by the so-called split algorithm. Most of the split-algorithms have algorithmic complexities are at least $\Omega(n)$.

2.2 Index Scan

The GiST scan maintains a stack of pages that may contain tuples, satisfying query conditions. Initially, this stack contains only a root page. During a scan, each internal page on top of the stack is replaced by referenced child pages with index tuples relevant to the query search condition. If the top of the stack contains a leaf page – each tuple of this leaf page is examined on matching query search condition and, if passed, outputted as the scan result. The index scan is complete when the stack is empty.

This algorithm ensures the depth-first scanning order of the index in a regular scan. But GiST also supports the k-nearest neighbor (kNN) index scan. kNN type of scan ensures that tree paths are not searched when it is known that they cannot contain tuples with a distance less than that have already been found by the scan. In a regular scan and in a kNN scan the inner page is deconstructed into its child references by $O(n)$ algorithm: checking all contents against search conditions (consistency operation), where n is the fan-out number of the page.

2.3 Index Scan Thought Experiment

GiST itself has no usable performance prediction model. But there are performance prediction models for indexes implementable over GiST. For example, there is a quite accurate cost model for R-Tree [16]. But this model has some limitations. First, the output of this model is accurate for "optimum, not implemented yet method" to build an R-tree. But R-tree-over-GiST is far from optimum. Second, this model outputs the number of disk accesses as a function of data properties. But actual index performance is itself a function of a number of disk accesses. When the index fits into main memory, disk accesses are not a bottleneck.

We can apply some theoretical reasoning to estimate index scan performance in terms of key collision checks (consistency function calls). We observed that these calls dominate in the CPU profile during the execution of lasting GiST index scans. If we have an index with N leaf tuples and each GiST node has a fan-out factor f, the height h of the tree will be $h \approx log_f N$. If the scan will find exactly 1 tuple within a tree without overlapping subtree keyspace, number of calls to consistency function CN will be

$$CN \approx f, \quad h \approx f \log_f N. \tag{1}$$

Here the approximation of CN is minimal when $f = e$, where e is the basis of the natural logarithm. And e is far smaller than usual $f \approx 250$.

Of cause, this reasoning has limitations. First, we assumed that on each level GiST scan did not encounter keyspace overlap of subtrees. Access methods are designed with overlap minimization as a goal [2]. But it is not always achievable. Second, single cache-missed disk access will dominate thousands of collision check calls. Multiple cache-missed disk accesses will render CN optimization useless.

In-memory GiST insertion, according to our observations, is dominated by penalty function calls. In previous work [8] we were focusing on enhancing the penalty function of *cube* and PostGIS extensions to achieve better index properties during the spatial search. Reasoning about CN can be applied to a number of *penalty* function calls intact: lower fan-out should yield fewer *penalty* calls.

3 Proposed Changes

3.1 Initial Design Considerations

How to reduce tree node fan-out? Initially, we proposed multi-level intra-page tree at PostgreSQL hackers mailing list [5].

While we can't fill a page with just 3 tuples, we can install a small tree-like structure inside one page. General GiST index has a root page. But a page tree should have a "root" layer of tuples. Let's consider the concept of private tuples (or internal, intermediate, auxiliary, we have to distinguish them from initial internal\leaf dichotomy) without links to other pages. These private tuples could have only keys and a fixed-size array of underlying records offsets (with size f). Each layer is a linked-list. After the page has just been allocated there is only "ground" level of regular tuples. Eventually, record count reaches $f - 1$ and we create a new root layer with two private tuples. Each new tuple references half of the preexisting records. Placement of new "ground" tuples on the page eventually will cause private tuple to split. If there is not enough space to split private tuple, we mark the page for the whole page-split during the next iteration of the insertion algorithms of owning the GiST tree. That is why tuple-split happens on $f - 1$ tuples, not on f: if we have no space for splitting, we just adding a reference to the last slot. In this algorithm, page split will cause major page defragmentation: we take the root layer, halve it and place halves on different pages. When half of a data is gone to another page, restructuration should tend to place records in such a fashion that accessed together tuples are placed together. Let's look how page grows with fan-out factor $f = 5$.

When we added 3 ground tuples it's just a ground layer, here `RLS` is root layer start, `G` is ground tuple, $\mathtt{I_x}$ is internal tuple of level x:

`RLS=0|G G G`, then we place one more tuples and layer splits:

$\mathtt{RLS{=}4|G\ G\ G\ G\ I_0\ I_0}$, each $\mathtt{I_0}$ tuple now references two `G` tuples.

We keep placing `G` tuples:

$\mathtt{RLS{=}4|G\ G\ G\ G\ I_0\ I_0\ G\ G}$, and then one of $\mathtt{I_0}$ tuples is splitted:

$\mathtt{RLS{=}4|G\ G\ G\ G\ I_0\ I_0\ G\ G\ G\ I_0}$, one more $\mathtt{I_0}$ split causes new layer:

$\mathtt{RLS{=}12|G\ G\ G\ G\ I_0\ I_0\ G\ G\ G\ I_0G\ I_0\ I_1\ I_1}$.

This structure could provide average tree fan-out as low as desired. An analysis of a similar approach is given in [10]. But this approach is too sophisticated for industrial implementation within PostgreSQL codebase with a steep learning curve. This algorithm would require *pg_upgrade* with the previous version of GiST, intervene into the physical structure of index tuple. Recovery is affected by this structure too: there are new possible valid states in case of a crash, that has to be handled during recovery from WAL. After the discussion in pgsql-hackers list and few technical seminars, we decided to give up this structure in favor of a more simple and maintainable design.

3.2 Simplified Intra-page Indexing

To simplify intra-page indexing we decided to use a two-level indexing scheme instead of multilevel. Also, one of the design goals was to exclude the necessity of *pg_upgrade* of old GiST indexes to a newer version.

To achieve these goals, we decided to introduce the concept of skip tuples. The skip tuple is the tuple, which allows skipping the next few tuples if the key of skip tuple indicates that the following group of tuples is of no interest to a given algorithm.

This approach relies on tuple ordering on the page. The GiST in PostgreSQL 9.6 tends to shuffle records for the sake of code simplicity. We have fixed this [3, 7] in PostgreSQL 10, introducing routines for tuple overwrite. This advancement allowed to gain about 15% of insertion performance for the price of more complex code. But what was really important is that now GiST could sustain tuple ordering and in future versions of GiST, we could afford to rely on this order to implement a skip tuples approach.

Regular GiST tuples always have reference to a page. While tuples on internal pages have references to other pages of the same GiST index, tuple on leaf page has reference to pages in a heap. Skip tuples do not have reference to any other page, and we are using a reference part of the tuple structure to store count of tuples, united by a key of given skip tuple. We've found a spare bit in the tuple header structure and used it to indicate that given tuple is the skip tuple. Thus, we did not change any bit of code responsible for tuple accommodation on the page (as of PostgreSQL 11 development codebase).

3.3 Usage of Skip Tuples

When do the skip tuples appear? Initially, the GiST index is placed on one leaf page which is the root page. New tuples are simply appended at the end of the page. Obviously, at this moment intra-page indexing is not necessary: there is no room for a sophisticated algorithm to gain significant performance difference on 8 Kb of the data.

When the page is overflown, it is split and a new root is formed. GiST does not always split the page into 2 halves but can have up to 75 parts (arbitrary number chosen by GiST developers as a sane upper limit). For each new page GiST forms downlinks from the root page. These downlinks are placed into one skip group of the skip tuple. This defines a moment when the first skip tuple appears: when the root is first split by the codebase with support of skip tuples.

At this exact moment, we have to decide on one more tradeoff: if we place skip tuples on a leaf page, we favor faster scans, if we do not we favor faster inserts. From our experience, most of GiST use cases encounter an insufficient performance of GiST inserts and update. We decided to work on inserts-oriented algorithms. This is a subjective decision influenced mostly by GIS users. But proposed algorithms can be adjusted for scan-oriented without compatibility issues. Also, a multilevel structure can be added over two-level too.

Each time when the page is split new downlink must be formed (see Fig. 1). This downlink is possibly added to some skip group; thus, this group may overgrow some limit, we call this limit *skip group threshold* T. In case of this overflow, skip group is split with the regular split algorithm, which is already provided by data type for page split. This algorithm outputs some new skip groups, which replace the overflown skip group.

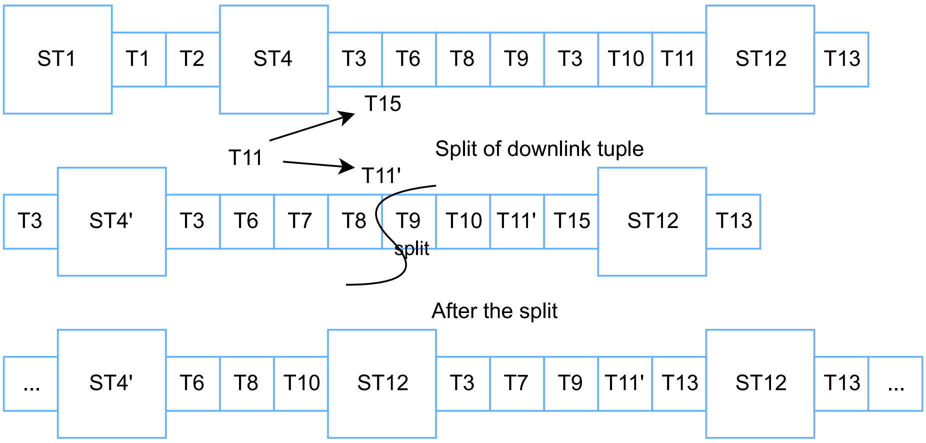


Fig. 1. Split of the skip group.

When the page with skip groups is overflown and has to be split, we pick vector of skip tuples from a page, split it, and distribute skip groups to new pages according to skip tuples split. On a rare occasion, this may produce skip group vectors, which do not fit a single page. Then we fall back to split of regular tuples from the same page.

When GiST is choosing subtree for insertion, it must pick downlink with minimal *penalty* value for inserting a given item (new index tuple). Since the penalty is the measure of “how much key space of subtree will be extended in case of insertion”, we suppose that for given item *penalty* of skip tuple is always no greater than for any tuple inside its skip group. We checked that this assumption holds true for *penalty* function bundled by all extensions in *contrib* directory of PostgreSQL source code and *penalty* functions in PostGIS. But we have no strict proof that this assumption holds everywhere since GiST does not apply enough restrictions on *penalty* function. For penalty function P, union function U, new entry e, items on page $i_1 \dots i_n$ we assume:

$$P(e, U(i_1 \dots i_n)) \leq P(e, i_x) \quad \forall x. \tag{2}$$

This inequality allows skipping the skip group during choose subtree algorithm if the penalty of its skip tuple is greater than already found. Most of the insertion performance improvement comes from this change. But that's not the only use of skip tuples.

During the execution of the scan, if the search conditions do not collide with a key of skip tuple, we can skip the whole skip group. Our theoretic analysis intentionally skips algorithms for deleting tuples from the index. PostgreSQL MVCC implementation prescribes that there are no routines to delete a single tuple from the access method, just scheduled vacuum – bulk deletion process, which is not affected by skip tuples directly.

4 Experimental Analisys

We have implemented proposed changes as a patch and published the patch on pgsql-hackers mailing list [6]. The patch is fully functional, passes all available regression and stress tests. We are going to work on inclusion it to mainstream PostgreSQL. Bug reports and experience feedback will be appreciated.

All conducted tests were single-threaded, conducted on a machine with Intel Core i7 (I7-4770HQ), 1600 MHz DDR3 SDRAM. PostgreSQL memory setting *shared_buffers* were configured to guaranty that all test data reside in RAM. The database cluster was completely wiped before each test. We had chosen a built-in data type *point* as the most basic and suitable for benchmarking. The *point* type is 16 bytes wide, represents a point on a plane. For indexing, it uses increasing of covered size by minimum bounding box as a penalty function and Korotkov split algorithm [15], which can be considered as one of the most advanced to date and is also used in PostGIS extension.

All test scripts are published on GitHub [4], along with bash scripts to run benchmarks for rapid results reproduction.

4.1 Tests with randomized data

We used following script to generate dataset and benchmark GiST insertion:

```
CREATE UNLOGGED TABLE x(c point);
CREATE INDEX ON x USING gist(c);
INSERT INTO x SELECT point(random(), random()) c
FROM generate_series(1,10000000) y;
VACUUM.
```

This script creates table textitx, which contains only one column textitc of a *point* type. Then the script creates a GiST index on this column. After that, the table is filled with ten million points, both coordinates are uniformly distributed between 0 and 1. We measured the time of insertion by *psql* metacommand \timing, which obtains time from the database server. Time of the table and index creation and VACUUM does not affect the time of insertion. It is worth noting that during data insertion the heap is also populated and this affects insertion time, to minimize this influence we used UNLOGGED table, but results for WAL-logged tables do not differ significantly.

We used following script to benchmark index scan time:

```
SET enable_bitmapscan = off;
EXPLAIN ANALYZE
WITH pts AS (SELECT random() x, random() y
FROM generate_series(1, 100000) y),
QUERIES AS
(SELECT box(point(x,y), point(x+0.01, y+0.01)) b FROM pts)
SELECT (SELECT count(*) FROM x WHERE x.c <@ q.b) FROM queries q.
```

This script generates one hundred thousand of points with coordinates uniformly distributed between 0 and 1, creates boxes from these points with edge 0.1, and for each box counts a number of data points within the box. Each box is expected to contain slightly less than one thousand data points on average. EXPLAIN ANALYZE is appended to control the execution plan during benchmarks.

We found that on default cluster configuration with source code from the master branch (git branch for bleeding edge version) insertion takes on average 123 seconds, while with intra-page indexing with T = 16 same task takes 81 seconds. This constitutes a 34.5% performance improvement. At the same time on master selection task takes 35.13 seconds, while with patch the task takes on average 32.87 seconds, 6.5% improvement.

The task of the insertion of randomized data is further referenced as RI (random inserts), the task of scanning index for counts computation is referenced as RS (random selects).

4.2 Tests with Ordered Data

To build an efficient index, GiST relies on *penalty* and *split* functions. Given that assumption about a penalty of skip group holds, it is provable that the use of penalty function is unaffected by skip group mechanics, besides the fact, that skip tuples occupy space on a page and allow to store fewer tuples. But it is a known fact that penalty function may degrade if data is provided to GiST in sorted order [7]. There are some techniques to mitigate this problem, but it allows us to construct the “worst case” for skip tuple technique.

In this case data is generated and inserted by script:

```
CREATE UNLOGGED TABLE x(c point);
CREATE INDEX ON x USING gist(c);
INSERT INTO x SELECT point(x / 1000.0, y / 10000.0) c
FROM generate_series(1, 1000) y, generate_series(1, 10000) x.
```

This script uses the same table and index as RI but inserts Cartesian product of two evenly increasing series of numbers between 0 and 1. The task of this data insertion is further references as OI (ordered inserts). To test searches, in this case, we used the same script as in RS, but on OI data, this task is called OS (ordered-data selects).

We observe that on master OI task takes 127 seconds, while with patch and T = 16 it takes 88 seconds, which is 30% improvement. In turn, OS on master takes 22 seconds, while with the intra-page index it takes 31 seconds, which is 39% degraded. Obviously, this degrade can be mitigated by *penalty* function enhancement, but this constitutes that intra-page indexing cannot rely solely on *split* algorithm.

4.3 The Case of Big Pages

We observed opinion, that GiST performance is degrading on big pages. In this case, intra-page indexing could help prevent degradation. Currently, PostgreSQL allows to specify page size before compilation, the maximum page size is 32 Kb. This size is restricted mainly by tuple placement structure *ItemIdData* which leaves 15 bits to tuple offset on the page.

We have done a series of tests with different T's on 32 Kb pages.

Table 1. RI and RS tasks time on 32Kb pages (ms).

T	RI			RS		
	T_{master}	$T_{patched}$	$T_{patched}/T_{master}$	T_{master}	$T_{patched}$	$T_{patched}/T_{master}$
8	201635	84671	0.42	36307	38453	1.06
16	201635	80375	0.40	36307	36448	1.00
24	201635	80870	0.40	36307	34573	0.95
32	201635	80135	0.40	36307	40717	1.12

From these results, we can see that while performance gain is sufficient on RI task, there is neither gain in RS task nor visible dependency from the threshold T.

4.4 Effect of Different Thresholds

Following is the test result, obtained with the most common page size of 8 Kb.

From these results, we can conclude that on given physical characteristics of the database and data type, $T = 16$ is somewhat optimal, but have no significant influence on performance if T is picked from sane numbers.

5 Current Limitations and Future Work

The implementation of the new approach brings performance benefits for insertion tasks, fixing common bottlenecks.

Table 2. RI and RS tasks time on 8 Kb pages (ms).

T	RI			RS		
	T_{master}	$T_{patched}$	$T_{patched}/T_{master}$	T_{master}	$T_{patched}$	$T_{patched}/T_{master}$
8	123260	83111	0.67	35135	33556	0.96
16	123260	80930	0.66	35135	32873	0.94
24	123260	88469	0.72	35135	34662	0.99
32	123260	87234	0.71	35135	35400	1.01

But currently, the proposed approach has several unresolved questions. We hope to address these questions first in AGS and next in mainstream GiST.

5.1 Buffered GiST Build

GiST has buffered build, which is used to build an index structure for a preexisting table. Because buffered build initializes itself with building a small tree with regular inserts, it is installing skip tuples too. That is why buffered build has performance improvement from the described technique too. But it is certain that buffered build could benefit more if it consciously used skip tuples rather than benefiting from side effects of initialization with skip tuples.

5.2 kNN Search in GiST

Skip tuples are not integrated into the kNN search pairing heap. If the kNN has a search condition operator, this search will use skip tuples to improve its performance. But granularity of pairing heap is still a page and not a skip group. Changing the granularity of kNN pairing heap will incur CPU performance cost because pairing heap will be larger. But skip tuples are itself technique to improve CPU usage. Thus, at the present state, the implementation of the proposed approach does not use skip tuples during the kNN ordering scan.

6 Related Work

Currently PostgreSQL has space-partitioned GiST SP-GiST [1]. This index is also free from the concept one node is one page, but in another manner. SP-GiST is an unbalanced tree and each page can represent the subgraph of this tree. Unfortunately, SP-GiST inherits few limitations from this its nature of dividing space into subspaces without overlap: it cannot index overlapping keys. In terms of GIS this means that SP-GiST cannot store areas, it can be used only as point access method. Usually, for scans and inserts, SP-GiST uses less CPU but more IO of page buffers. Since it's unbalanced nature, the SP-GiST scan can be unpredictably deep.

7 Conclusion

The current implementation of intra-page indexing can make GiST inserts and updates 1.5x faster. It protects GiST from performance degradation if PostgreSQL is compiled with big pages. But the most important achievement of intra-page indexing is that it opens a way to advance GiST API towards better-generalized algorithms, less code for data type developers and more performant data access methods for these data types.

Acknowledgments

The work was partially supported by Act 211 Government of the Russian Federation, contract No. 02.A03.21.0006.

References

1. Aref, W.G., Ilyas, I.F.: Sp-gist: An extensible database index for supporting space partitioning trees. Journal of Intelligent Information Systems **17**(2-3), 215–240 (2001)
2. Beckmann, N., Seeger, B.: A revised r*-tree in comparison with related index structures. In: Proceedings of the 2009 ACM SIGMOD International Conference on Management of data. pp. 799–812. ACM (2009)
3. Borodin, A.: Gist inserts optimization with pageindextupleoverwrite, `https://commitfest.postgresql.org/10/661`, [Last accessed 30-June-2019]
4. Borodin, A.: Intra-page indexing benchmark, `https://gist.github.com/x4m/56f912ba9278a97f24dfa2b6db46fa7f`, [Last accessed 30-June-2019]
5. Borodin, A.: [proposal] improvement of gist page layout, `https://www.postgresql.org/message-id/flat/CAJEAwVE0rrr\%2B0BT-P0gDCtXbVDkBBG_WcXwCBK\%3DGHo4fewu3Yg\%40mail.gmail.com`, [Last accessed 30-June-2019]
6. Borodin, A.: [wip] gist intrapage indexing, `https://www.postgresql.org/message-id/7780A07B-4D04-41E2-B228-166B41D07EEE@yandex-team.ru`, [Last accessed 30-June-2019]
7. Borodin, A., Mirvoda, S., Kulikov, I., Porshnev, S.: Optimization of memory operations in generalized search trees of postgresql. In: International Conference: Beyond Databases, Architectures and Structures. pp. 224–232. Springer (2017)
8. Borodin, A., Mirvoda, S., Porshnev, S., Bakhterev, M.: Improving penalty function of r-tree over generalized index search tree possible way to advance performance of postgresql cube extension. In: 2017 IEEE 2nd International Conference on Big Data Analysis (ICBDA)(. pp. 130–133. IEEE (2017)
9. Chilingarian, I., Bartunov, O., Richter, J., Sigaev, T.: Postgresql: The suitable dbms solution for astronomy and astrophysics. In: Astronomical Data Analysis Software and Systems (ADASS) XIII. vol. 314, p. 225 (2004)
10. Christodoulakis, S., Manolopoulos, Y., Larson, P.Å.: Analysis of overflow handling for variable length records. Information Systems **14**(2), 151–162 (1989)
11. Guttman, A.: R-trees: a dynamic index structure for spatial searching, vol. 14. ACM (1984)
12. Hellerstein, J.M., Naughton, J.F., Pfeffer, A.: Generalized search trees for database systems. September (1995)

13. Kornacker, M.: Access methods for next-generation database systems. University of California, Berkeley (2000)
14. Kornacker, M., Mohan, C., Hellerstein, J.M.: Concurrency and recovery in generalized search trees. In: ACM SIGMOD Record. vol. 26, pp. 62–72. ACM (1997)
15. Korotkov, A.: A new double sorting-based node splitting algorithm for r-tree. Programming and Computer Software **38**(3), 109–118 (2012)
16. Theodoridis, Y., Sellis, T.: A model for the prediction of r-tree performance. In: PODS. vol. 96, pp. 161–171 (1996)

Synchronization Aspects of The Optimistic Parallel Discrete Event Simulation Algorithms

Liliia Ziganurova[1,2] and Lev Shchur[1,2]

[1] Scientific Center in Chernogolovka, 142432, Chernogolovka, Moscow region,
[2] National Research University Higher School of Economics, 101000, Moscow
E-mails: ziganurova@gmail.com, levshchur@gmail.com

Abstract. We study synchronization aspects in parallel discrete event simulation (PDES) algorithms. Our analysis is based on the recently introduced model of virtual times evolution in an optimistic synchronization algorithm. This model connects synchronization aspects with the properties of the profile of the local virtual times. The main parameter of the model is a "growth rate" $q = 1/(1 + b)$, where b is a mean rollback length. We measure the average utilization of events and the desynchronization between logical processes as functions of the parameter q. We found that there is a phase transition between an "active phase", i.e. when the utilization of the average processing time is finite, and an "absorbing state" with zero utilization, vanishing at a critical point $q_c \approx 0.136$. The average desynchronization degree (i.e. the variance of local virtual times) grows with the parameter q. We also investigate the influence of the sparse distant communications between logical processes and found that they do not change drastically the synchronization properties in the optimistic synchronization algorithm, which is the sharp contrast with the conservative algorithm [1]. Finally, we compare our results with the existing case-study simulations.

Keywords: discrete event simulation, parallel discrete event simulation, PDES, optimistic algorithm, small-world

1 Introduction

Parallel discrete event simulation (PDES) [2] is a powerful tool of programming on high-performance computing systems [3]. It is widely used for modeling complex systems in computer science, engineering, physics, economics, and society [4]. The main advantage of PDES is that it is highly scalable by construction, for example, PDES simulator ROSS [5] is able to scale up to 1.9 million cores running a synthetic PHOLD model [6]. Even though the ideas of the method emerged around 40 years ago [7], the study of the PDES is still important nowadays. State-of-the-art and research challenges in the area of parallel simulation can be found in recently published papers [8, 9]. The study of PDES is going in many directions: studying different properties of PDES models [10], optimization of simulation kernels [11–14], different usage of PDES, e.g. internet of things [15], etc. In this paper we investigate properties of optimistic PDES algorithm, using a model of evolution of local virtual times.

The idea of PDES is that the physical system is simulated as a set of subsystems, which communicate with each other by time-stamped messages. The subsystems are mapped on programming objects, or logical processes (LPs). The logical process executes a sequential subprogram with its own local state variables and its own local virtual time (LVT) on some processing elements (nodes, processors, cores, or threads). During the simulation, the LPs interact by sending time-stamped event messages to each other. Each LP has an input and output queues of events. The received event messages, which are waiting for execution, are stored in the input queue, and the messages, which must be sent to other LPs, are located in the output queue. The messages in both queues are sorted by timestamp order. The simulation process goes as follows: each LP takes the first message from its input queue, executes an event, changes its local state and local virtual time, and sends messages to other LPs, if necessary. LPs works in parallel independently, without global synchronization. The simulation result will be correct (i.e. as if the simulation was sequential) if all the events have been executed by all LPs in correct non-decreasing timestamp order. In PDES the synchronization is carried out by each LP by the analysis of the values of timestamps in the queue, according to some synchronization protocol. There are three classes of the synchronization protocols: conservative, optimistic, and Freeze-and-Shift (FaS) protocol [2, 16, 17]. PDES algorithm can be classified using the mapping of the algorithm onto the partial differential equation describing the surface growth [18] and analyzing the boundary conditions [17]. In this scheme, the open boundary conditions correspond to the optimistic algorithm, the periodic boundary conditions correspond to the conservative algorithm, and the fixed boundary conditions correspond to the FaS algorithm.

In conservative synchronization, only secure events are allowed to be processed. The event is called secure, if we are sure that during the execution of this event the LP will not receive a message with a lower timestamp. This is usually implemented by using block-resume mechanisms, such that flags, semaphores, etc. The optimistic algorithm, in contrary, allows causality violations but provides a rollback mechanism for causality recovery.

All of the synchronization algorithms have their pros and cons and should be used according to the available computational facilities and the particular knowledge on the simulated system. For example, conservative synchronization is a better choice for systems with good lookahead information, i.e. the information on the minimal time between two dependent events. The conservative algorithm is easier to implement, but it generally works more slowly than the optimistic one. Realization of the optimistic algorithms are more complex, but usually, have better performance and can be used for a wider class of models [19].

Our research is focused on the study the synchronization properties of the optimistic PDES algorithm on different communication networks via the analysis of the local virtual time profile (Fig. 1). Such an approach was introduced for conservative synchronization algorithm in [20], and extended to other PDES algorithms and topologies in works [1, 21–27]. The approach provides rather a theoretical point of view on the synchronization PDES algorithms and allows to make general predictions about their behaviors. Moreover, the model can be attributed to the models of surface growth in physics, which allows using a rich instrument of statistical physics for the analysis of our model.

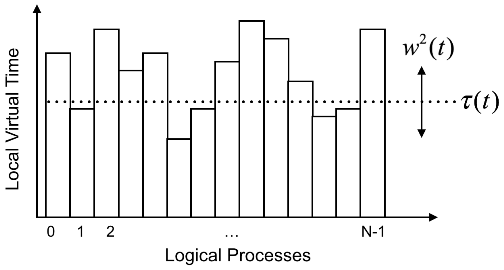


Fig. 1. A snapshot of local virtual time profile at a simulation step t. The LPs have their values of the LVT. $\tau(t)$ is a local virtual time averaged over all logical processes and $w^2(t)$ is an average squared width of the profile

The paper is organized as follows. In Section 2 we describe the model of evolution of LVTs in optimistic PDES algorithm. Section 3 provides the simulation results. In Section 4 we discuss the results and compare them with the existing case-studies of PDES models.

2 Model description

In this section, we describe a model [27] of evolution of LVT profile in optimistic PDES. We do not simulate any particular optimistic synchronization algorithm. Instead, we focused on the behavior of the local virtual times simulating the *model* of the optimistic PDES algorithm. There is a one-to-one correspondence between the LVT profile and the synchronization aspects of the optimistic algorithm. The average speed of the profile reflects the utilization of events or the effectiveness of processors load, and the profile width can be thought as a measure of desynchronization degree between LPs. The desynchronization shows the deviation of LVTs from the average between all LPs. A small deviation from the average time indicate that the LPs work at more or less equal pace, and none of the LPs are too ahead or behind from the others, while a high value of the desynchronization degree implies that some LPs are ahead and some of the LPs are behind, what increases a probability of causality violations and makes the leading processes to wait for the actual information from the lagging processes.

As a consequence, the average efficiency of the simulation slows down because of high desynchronization between LPs.

In optimistic PDES algorithms, the LPs are allowed to execute events independently without synchronization. At this stage, the LVT profile is growing freely. When the causality of computations is violated, i.e. some LP receives a message with a timestamp lower, than its LVT, the mechanism of rollback is run. This LP changes its LVT and state variables to the value when the receiving the erroneous message would be safe. After that, all sent messages must be "unsent". This is done by sending so-called anti-messages – the same messages but with the opposite sign. When a message and its anti-message occur in the same queue, they annihilate. It is clear, that one rollback can cause an avalanche of rollbacks. When the processing of rollbacks has been finished, the LVT profile will be in average lower and flatter, since some of the LPs changed their LVT to the lower values.

We simulate this process as follows. First, we set a communication topology. The communication topology determines the dependencies of LPs and can be presented as a graph, where vertices represent the LPs and edges represents the dependencies between the LPs (Fig. 2). The dependent LPs exchange by the messages and the independent ones do not communicate. Then we initiate an array of LVTs (i.e. the LVT profile) and update it according with the rules described below in this chapter. During the simulation, we calculate the observables: the average speed and the average squared width of the LVT profile. We use the following assumptions:

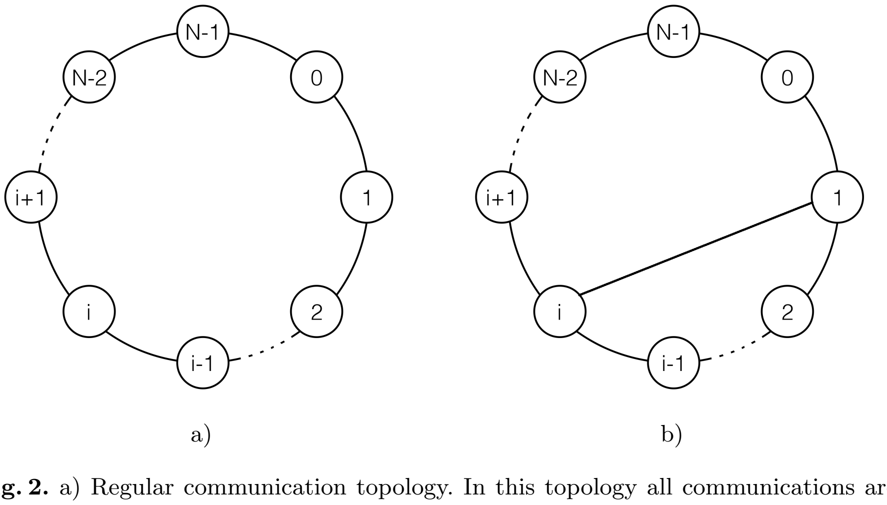


Fig. 2. a) Regular communication topology. In this topology all communications are local. b) Small-world communication topology. Besides the local communications, there are additional distant communications. The LP_1 and LP_i are distant processes, connected by a communication link

1. The communication topology is known and fixed in advance: it is known, which LPs exchange by messages and which LPs are independent.
2. Times between two events are random variables exponentially distributed with the mean 1.
3. Sending time and receiving time are equal (i.e. there is no communication overhead).
4. The causality violation may occur with equal probability at any LP.
5. If LP$_i$ depends on the information from several LPs, the causality violation on the LP$_i$ may be caused by any of those LPs with equal probability.
6. The number of rollbacks is a discrete random variable exponentially distributed with the mean b.

1. Setting the communication topology. We consider a system of N logical processes connected into a communication graph. We are interested in two communication topologies: regular and small-world because they reflect interconnections in real systems [28]. In regular topology, the LPs are arranged into a ring such that each LP depends on the two neighboring LPs: one at the left and one at the right (Fig. 2a). In small-world topology, we add a small amount of communications between distant LPs above the ring (Fig. 2b). Small-world topology is characterized by the low value of the average shortest path – it scales logarithmically with the system size, while in regular networks the average shortest path growth linearly with N. It was shown in [24] that these distant communications significantly enhance the synchronization between LPs in conservative synchronization algorithm.

The amount of distant communications is controlled by the parameter p. The total amount of distant communications is equal to pN. The case of $p = 0$ corresponds to a regular network.

2. Simulation of evolution of the LVT profile. When the communicational graph is determined, we start to simulate an evolution of LVTs. We begin with a flat LVT profile $\tau_i(t = 0) = 0$, $i = 1, 2, ., N$, where N is a number of LPs, ant t is one simulation step in our model. We assume that one simulation step consists of two stage: 1) simulation of profile growth, and 2) simulation of rollbacks. At the first stage each LP$_i$ increases its LVT by random value η_i, drawn from the Poisson distribution: $\tau_i(t+1) = \tau_i(t) + \eta_i, i = 1, \ldots, N$.

To simulate a rollback, we randomly choose a LP$_i$ and compare its LVT τ_i with the value of LVT τ_r of one of its neighbors LP$_r$, chosen with equal probability. If $\tau_i > \tau_r$, we set τ_i equal to τ_r. We repeat this action several times, assuming that the number of rollbacks is a random value drawn from the Poisson distribution with the mean b. The actions described above constitute one simulation step t. The full simulation consists of M simulation steps.

3. Calculation of the observables. After one simulation step t (increasing LVT profile + rollbacks) we calculate the observables:

1. The average height of the LVT profile $\tau(t)$ – an arithmetical mean of all LVTs at simulation step t:

$$\tau(t) = \frac{1}{N}\sum_{i=1}^{N}\tau_i(t).$$

2. The average speed of the profile $u(t)$ – an increment of the average height of the profile after one simulation step:

$$u(t) = \tau(t+1) - \tau(t).$$

3. The average squared width of the profile $w^2(t)$ – a statistical variance of LVTs from the mean value $\tau(t)$:

$$w^2(t) = \frac{1}{N}\sum_{i=1}^{N}[\tau_i(t) - \tau(t)]^2.$$

The described algorithm of evolution of LVT profile in optimistic PDES in pseudocode looks as follows:

```
Set parameters $N, M, p, b$;
Create a communication graph;
for $t := 0; t < M; t++$ do
    for $i := 0; i < N; i++$ do $\tau_i(t) += \eta_i$
    k = Poisson(b)
    for $j := 1; j < kN; j++$ do
        Choose random LP$_m$
        Choose random neighbour of LP$_m$ LP$_r$
        if $\tau_m(t) > \tau_r(t)$ then $\tau_m(t) = \tau_r(t)$
    Calculate observables.
```

3 Simulation results

We investigate the average speed and the average squared width of the LVT profile, which reflect such properties of the optimistic algorithm as the utilization of events and desynchronization between LPs, accordingly. We performed our simulation on regular and small-world topologies, varying the parameter p from 0 to 0.1. Number of LPs is fixed to $N = 10^4$, number of the simulation step M changes from 10^3 to 10^5. We also introduce a parameter $q = 1/(1+b)$, where b is a mean rollback length. The parameter q controls a growth rate of the profile and changes in our models from 0 to 1. We conduct the simulation using random number generation library RNGAVXLIB [29] and average the results over 1000 independent realizations of the models with fixed parameters.

The average speed on a regular topology. The average speed of the profile shows, how fast the LPs utilize the events. In our model the LVT profile growth with constant velocity, therefore we omit time dependence in the next formulas. We found, that the average speed u decreases with the parameter q, and when q approaches to some critical value q_c, the speed becomes equal to 0. Such behavior can be explained by a high amount of rollbacks, which do not let the profile of LVT grow (q is reversely proportional to the number of rollbacks).

We approximate the average speed u as a function of q by the following formula:

$$u(q) = u_0(q - q_c)^{\nu}. \tag{1}$$

The results of the fit of the data to the expression (1) are: $u_0 = 1.26(2)$, $q_c = 0.136(1)$, $\nu = 1.78(2)$. The behavior of the speed shows phase transition between an "active phase" (when $u > 0$), and "pinned phase" (when $u = 0$). Such behavior reminds a transition in directed percolation models [30]. It is interesting, that the critical exponent ν is also close to the critical exponent of directed percolation universality class.

The average speed on a small-world topology. When the LPs are connected into a small-world communication network, the behavior of the average speed slightly changes. The critical point q_c shifts to the right, when the parameter p increases. It happens, because the number of dependencies between LPs is increasing with p, therefore the probability of longer rollback avalanche is higher. The critical exponent ν also grows with p.

The average squared width on a regular topology. The average squared width of the LVT profile characterizes the degree of desynchronization between LPs. The width grows in a power-law manner with time t and then saturates. The saturation time and saturation value is higher for the larger parameter q. It is explained by the fact, that the number of rollbacks, in this case, is low, therefore the LVT profile grows freely.

The average squared width on a small-world topology. The behavior of LVT profile in the optimistic PDES algorithm does not exhibit qualitatively changes, when the underlying topology changes from regular to a small world, as it happens in the conservative algorithm [1]. The average squared width also grows with time and the parameter q but decreases slightly with the concentration of long-range connections p. As in the conservative algorithm, the additional communication links make the LVT profile smoother, i.e. the LPs work more synchronized. However, the difference between regular and small-world topologies in the optimistic algorithm is not so significant, because the mechanism of rollback reduces the difference between LVTs, even in the absence of additional communications between LPs.

4 Discussion

We analyzed the synchronization properties of optimistic PDES algorithm on regular and small-world communication topologies, using the model [27]. The model was introduced for the optimistic algorithm with only local interactions between logical processes. The results of our study have shown that the model is also applicable to the qualitative predictions of the synchronization properties of the optimistic algorithm with more general types of communication topology.

We found, that there is a critical point, at which the growth of the LVT profile stops, i.e. the utilization of events becomes zero. It means, that for systems with a high probability of rollbacks the optimistic algorithm would not be efficient. We also compared the results on regular and small-world topologies and found that the additional distant communications do not play such an important role as in the conservative PDES algorithm, where the synchronization was significantly better on small-world topology than on the regular topology [1].

For the application of our model to the real simulations, it is necessary to find an analogy between the parameters of the simulated systems and the parameter q of the present model. It is also possible to compare our results with the existing case-studies of various PDES models.

Paper [31] summarizes the profile data captured from 22 discrete-event simulation models from 4 simulators: NS-3 [32, 33], ROSS [5], WARPED2 [34], and Simian [35]. The research focuses on the communication properties of events exchanges between the LPs, namely, LP connectivity, betweenness centrality, and modularity. The analysis of LP connectivity has shown that in most models the LPs have either a fixed amount of connections (regular topology) or 1-8 connections in some proportion (as in small-world topology). The tendency of LPs to communicate with only a few other LPs makes the models good for parallel execution. The same LP connectivity is seen in our model as well, however, we cannot provide a detailed description of betweenness centrality and modularity of communication graphs in our model.

In [36] the performance of PDES is studied on ROSS simulator running PHOLD model on Knights Landing Processor. It was shown that the number of submitted events is decreasing with the fraction of remote events (event-messages passing between different cores). However, the simulation performance scales linearly with the number of cores, if each LP is assigned to its core, and the fraction of remote events is less than 10%. In our simulations we studied the topologies with a small fraction of remote connections (from 0.1 to 10%), and also found that they slightly slow down the performance (i.e. the average speed of the LVT profile) in both, the conservative and the optimistic algorithms. At the same time in the conservative algorithm they drastically enhance the synchronization (i.e. the average squared width of the LVT profile).

Another analogy between the observations in [36] and our results can be drawn regarding the interval of Global Virtual Time (GVT) update. The GVT is a minimum value among all LVTs. The state variables of LPs are stored only until the GVT. Smaller GVT interval requires less state information to be kept, but increase the overhead of GVT calculations. On the other hand, the rollback length is shorter, therefore the calculation of rollbacks goes faster. The interval of GVT computation has some similarity to the parameter q of our model.

The average speed of the profile in our model has values from 0 to 1. It can be compared with the average utilization of events in [10] varying from 0.47 for an epidemic model to 0.0043 for traffic model, and down to $5 \cdot 10^{-5}$ for wireless network model on running ROSS [5] and WARPED2 [34] simulators.

In the future, we plan to perform case-study simulations of the existing PDES models and establish relationships between the parameters of the real parallel discrete-event simulations and the parameters of our models.

Acknowledgments

The work has been done within the research theme 0236-2019-0001.

References

1. Ziganurova, L. and Shchur, L.N.: Synchronization of conservative parallel discrete event simulations on a small-world network. Physical Review E **98**, 022218 (2018). doi:10.1103/PhysRevE.98.022218
2. Fujimoto, R.M.: Parallel discrete event simulation. Communications of the ACM **33**, 30–53 (1990). doi: 10.1145/84537.84545
3. Bailey, D.H., David, H., Dongarra, J., Gao, G., Hoisie, A., Hollingsworth, J., Jefferson, D., Kamath, C., Malony, A., and Quinian, D.: Performance Technologies for Peta-Scale Systems: A White Paper Prepared by the Performance Evaluation Research Center and Collaborators. White paper, Lawrence Berkeley National Laboratories (2003). doi: 10.2172/15004540
4. Tropper, C.: Parallel Discrete-Event Simulation Applications. Journal of Parallel and Distributed Computing **62** (3), 327–335 (2002). doi: 10.1006/jpdc.2001.1794.
5. Carothers, C.D., Bauer, D., and Pearce, S.: ROSS: A high-performance, low-memory, modular Time Warp system. Journal of Parallel and Distributed Computing **62**, 1648–1669 (2002). doi: 10.1016/S0743-7315(02)00004-7
6. Barnes, Jr, P. D., Carothers, C.D., Jefferson, D.R., and LaPre, J.M.: Warp speed: executing time warp on 1,966,080 cores. In Proceedings of the 1st ACM SIGSIM Conference on Principles of Advanced Discrete Simulation, 327–336 (2013). doi: 10.1145/2486092.2486134
7. Jefferson, D., and Fujimoto, R.: A Brief History of Time Warp. In Advances in Modeling and Simulation, 97–134 (2017). Springer, Cham. doi: 10.1007/978-3-319-64182-9_7
8. Balci, O., Fujimoto, R.M., Goldsman, D., Nance, R.E., and Zeigler, B.P.: The state of innovation in modeling and simulation: the last 50 years. In 2017 Winter Simulation Conference (WSC), 821–836 (2017). IEEE. doi: 10.1109/WSC.2017.8247835
9. Fujimoto, R., Bock C., Chen, W., Page E., and Panchal, J.H.: Research challenges in modeling and simulation for engineering complex systems. Springer (2017). doi: 10.1007/978-3-319-58544-4
10. Wilsey, P.A.: Some Properties of Events Executed in Discrete-Event Simulation Models. In: Proceedings of the 2016 annual ACM Conference on SIGSIM Principles of Advanced Discrete Simulation, 165–176 (2016). ACM, New York. doi: 10.1145/2901378.2901400

11. Ross, C.J., Carothers, C.D., Mubarak, M., Ross, R.B., Li, J.K., and Ma, K.L.: Leveraging Shared Memory In The Ross Time Warp Simulations. In 2018 Winter Simulation Conference (WSC) 3837–3848 IEEE. (2018). doi: 10.1109/WSC.2018.8632333
12. Eker, A., Williams, B., Mishra, N., Thakur, D., Chiu, K., Ponomarev, D., and Abu-Ghazaleh, N.: Performance Implications of Global Virtual Time Algorithms on a Knights Landing Processor. In 2018 IEEE/ACM 22nd International Symposium on Distributed Simulation and Real Time Applications (DS-RT), 1–10. IEEE (2018). doi: 10.1109/DISTRA.2018.8600923
13. Masko, L. and Tudruj, M.: Application global state monitoring in optimization of parallel event-driven simulation. Concurrency and Computation: Practice and Experience **e5015** (2018). doi:10.1002/cpe.5015
14. Knopov, P. and Pardalos, P.M., eds. Simulation and optimization methods in risk and reliability theory. Nova Science Pub Incorporated, 2009.
15. D'Angelo, G., Ferretti, S., and Ghini, V.: Simulation of the Internet of Things. In 2016 International Conference on High Performance Computing and Simulation (HPCS), 1–8. IEEE (2016) doi: 10.1109/HPCSim.2016.7568309
16. Jefferson, D.R.: Virtual time. ACM Transactions on Programming Languages and Systems (TOPLAS) **7**, 404–425 (1985). doi: 10.1145/3916.3988
17. Shchur, L.N. and Novotny, M.A.: Evolution of time horizons in parallel and grid simulations. Physical Review E **70**, 026703 (2004). doi: 10.1103/PhysRevE.70.026703
18. Kardar, M., Parisi, G., and Zhang, Y.C.: Dynamic scaling of growing interfaces. Physical Review Letters **56**, 889 (1986). doi: 10.1103/PhysRevLett.56.889
19. Jefferson, D.R. and Barnes, Jr P.D.: Virtual time III: Unification of conservative and optimistic synchronization in parallel discrete event simulation. In Proceedings of the 2017 Winter Simulation Conference, 55. IEEE Press (2017). doi: 10.1109/WSC.2017.8247832
20. Korniss, G., Toroczkai, Z., Novotny, M.A., and Rikvold, P.A.: From massively parallel algorithms and fluctuating time horizons to nonequilibrium surface growth. Physical Review Letters **84**, 1351 (2000). doi: 10.1103/PhysRevLett.84.1351
21. Shchur, L.N. and Shchur, L.V.: Relation of Parallel Discrete Event Simulation algorithms with physical models. Journal of Physics: Conference Series **640**, 012065 (2015). doi: 10.1088/1742-6596/640/1/012065
22. Shchur L. and Shchur L.: Parallel Discrete Event Simulation as a Paradigm for Large Scale Modeling Experiments. In: Selected Papers of the XVII International Conference on Data Analytics and Management in Data Intensive Domains (DAMDID/RCDL 2015), Obninsk, Russia, October 13–16, 2015, pp. 107–113 (2015). http://ceur-ws.org/Vol-1536/
23. Alon, U., Evans, M.R., Hinrichsen, H., and Mukamel, D.: Roughening transition in a one-dimensional growth process. Physical Review Letters **76**, 2746 (1996). doi: 10.1103/PhysRevLett.76.2746
24. Guclu, H., Korniss, G., Novotny, M.A., Toroczkai, Z., and Racz, Z.: Synchronization landscapes in small-world-connected computer networks. Physical Review E **73**, 066115 (2006). doi: 10.1103/PhysRevE.73.066115
25. Ziganurova, L. and Shchur, L.: Properties of the Conservative Parallel Discrete Event Simulation Algorithm. Lecture Notes in Computer Science **10421**, 246–253 (2017). doi: 10.1007/978-3-319-62932-2_23
26. Shchur, L. and Ziganurova, L.: Simulation of Virtual Time Profile in Conservative Parallel Discrete Event Simulation Algorithm for Small-World

Network. Lobachevskii Journal of Mathematics **38** (5), 967–970 (2017). doi:10.1134/S1995080217050316

27. Ziganurova, L., Novotny, M.A., and Shchur, L.N.: Model for the evolution of the time profile in optimistic parallel discrete event simulations. In: Journal of Physics: Conference Series **681**, 012047 (2016). doi: 10.1088/1742-6596/681/1/012047
28. Watts, D.J. and Strogatz, S.H.: Collective dynamics of small-world networks. Nature **393** (6684), 440 (1998). doi: 10.1038/30918
29. Guskova, M.S., Barash, L.Y., and Shchur, L.N.: RNGAVXLIB: Program library for random number generation, AVX realization. Computer Physics Communications **200**, 402–405 (2016). doi: 10.1016/j.cpc.2015.11.001
30. Odor, G.: Universality classes in nonequilibrium lattice systems. Reviews of modern physics **76** (3), 663 (2004). doi:10.1103/RevModPhys.76.663
31. Crawford, P., Eidenbenz, S.J., Barnes, P.D., and Wilsey, P.A.: Some properties of communication behaviors in discrete-event simulation models. 2017 Winter Simulation Conference (WSC), Las Vegas, NV, 1025–1036 (2017). doi: 10.1109/WSC.2017.8247852
32. Henderson, T.R., Lacage, M., Riley, G.F., Dowell, C., and Kopena, J.: Network Simulations with the ns-3 Simulator. SIGCOMM demonstration **14**, 527 (2008).
33. Riley, G.F. and Henderson, T.R.: The ns-3 Network Simulator, 1534. Springer (2010).
34. Weber, D.: Time warp simulation on multi-core processors and clusters. Master's thesis, University of Cincinnati, Cincinnati, OH (2016).
35. Santhi, N., Eidenbenz, S., and Liu, J.: The Simian Concept: Parallel Discrete Event Simulation with Interpreted Languages and Just-In-Time Compilation. In Proceedings of the 2015 Winter Simulation Conference, edited by L. Yilmaz, W. K. V. Chan, I. Moon, T. M. K. Roeder, C. Macal, and M. D. Rossetti, 3013-3024. Piscataway, New Jersey, USA: Institute of Electrical and Electronics Engineers, Inc. (2015) doi:10.1109/WSC.2015.7408405
36. Williams, B., Ponomarev, D., Abu-Ghazaleh, N., and Wilsey, P.: Performance characterization of parallel discrete event simulation on knights landing processor. In Proceedings of the 2017 ACM SIGSIM Conference on Principles of Advanced Discrete Simulation, 121–132. ACM (2017). doi: 10.1145/3064911.3064929

Linguistic Big Data: Problem of Purity and Representativeness

V.D. Solovyev [1][0000-0003-4692-2564] and S.S. Akhtyamova [2][000-0001-5863-9841]

[1] Kazan (Volga Region) Federal University, Kazan, Russia
[2] Kazan National Research Technological University, Kazan, Russia
Maki.solovyev@mail.ru

Abstract. This paper deals with the quality problem of linguistic big data exemplified by the corpus of Google Books Ngram. The criticism of this corpus has been summarized and discussed. Special attention is paid to the matters of the corpus balance, spelling errors, and errors in metadata. It is also compared to the Russian National Corpus and to the General Internet-Corpus of Russian. A new concept, "diachronically balanced corpus", has been introduced. The methods are discussed for enhancing the quality of Google Books Ngram.

Keywords: Text Corpora, Data Representativeness, Time-series, Data Noisiness.

1 Introduction

Modern linguistic studies can hardly go without using large text corpora or linguistic databases of various types or without applying to them computer-based or mathematical methods to obtain valid statistics. For the Russian language, the Russian National Corpus [1] (abbreviated as RNC) is well known, which is specified in [2, 3]. RNC contains over 350 million words [1]. It is carefully checked, and its part has been disambiguated manually. All this make it exceptionally useful for research in the Russian language.

A new, interesting resource appeared recently, the General Internet-Corpus of Russian (http://www.webcorpora.ru/, abbreviated as GICR). It already contains over 20 billion words and is thought to be further expanded. GICR includes the contents from the largest resources of the Runet, such as Zhurnalny Zal (Room of Thick Literary Journals), Novosti (News), VKontakte, LiveJournal, and Blogs at Mail.ru.

Since 2009, there has been an even larger corpus, Google Books Ngram (https://books.google.com/ngrams, abbreviated as GBN). It contains data on 9 languages, including Russian. The volume of the Russian GBN sub-corpus is over 67 billion words, while it exceeds 500 billion for the English one. GBN was created by fully scanning, followed by text recognition, all books from over 40 largest libraries in the world, including those of Harvard University and Oxford University. As a result, 30 million books were digitalized, of which the 8 million best digitalized ones were selected to form the corpus, which amounted to 6 % of all the books published worldwide [4]. A detailed description of the GBN can be found in [4–6]. For so large corpora, there seems no escaping the matters of their quality and of the possibility of errors in creating them. In this paper, we are focusing on GBN as the largest corpus of all the corpora existing in the world. Some publications noted errors in the corpus [7–9]. There are three core GBN problems discussed in literature: OCR errors, balance of the corpus, and errors in metadata. Herein, we are considering the above problems and the possible ways of improving the corpus. The paper presents a review of key publications related to the GBN issues, as well as the unique results obtained by the authors in this area.

2 OCR Errors

There are recognition errors in the GBN corpus, which are primarily related to ancient books characterized by poor print quality. In the first GBN version released in 2009, things were really in a bad way. Thus, in ancient English books, letter *s* was frequently recognized as *f*. For example, the word *best* was mistaken for *beft* in up to 50 % of the 17th-century books recognized. The creator of this resource, Google, has considered the criticism and considerably improved the recognition quality. Scanning devices were upgraded every six months [6]. As a result, in the second version in 2012, *best* was incorrectly recognized as *beft* in just 10% of cases in the 17th-century books, while in contemporary books of 2000, the amount of errors only made 0.02%, that is, it was rather low and could not affect any statistics regarding the frequency of using the word *best*.

Similarly, that is the case for the Russian language. We have considered several dozens of randomly chosen words containing recognition errors. Typically, the error rate does not exceed 0.1 %. For example, letter н is sometimes recognized as и. In Fig, 1, the exemplary frequency diagrams are shown for the word “иней” (hoarfrost). The frequency of its incorrect recognition as “иией” is lower than 0.1% of the correct one.

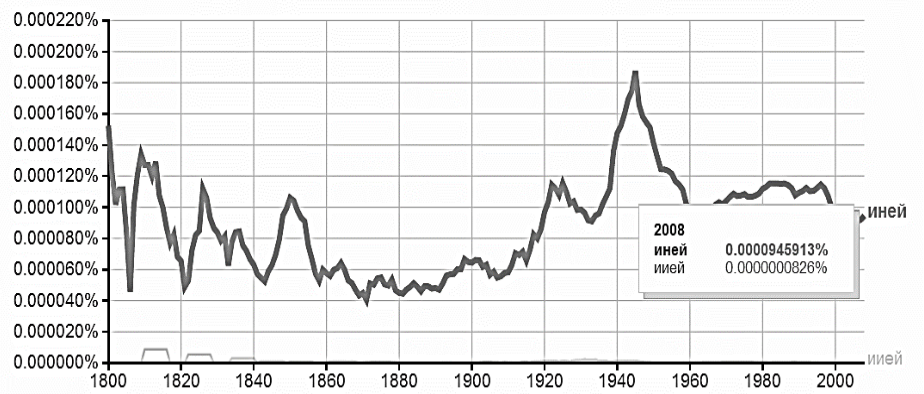


Fig. 4. Frequency of *иней* and *иией*

For Russian, certain difficulties occur with the pre-reform (before 1918) orthography. The Russian language previously used letters, such as Ѣ (yat) and Ѳ (fita), that are incorrectly recognized in GBN. For the data beyond 1918, the problem is eliminated. For many words from texts issued before 1918, the data will also be correct, since letters Ѣ and Ѳ, as well as other elements of the old orthography, are rather uncommon.

Thus, there are no apparent reasons for considering that recognition errors may essentially affect the results of counting the frequency of word usage, except for some probable rare cases with ancient books, where certain care must be exercised.

3 Balance

As a matter of principle, the problem of corpus balance is considerably more complicated. Balanced should be considered a corpus, in which all types of texts, i.e., literary, journalistic, pedagogic, scientific, business, and other texts, occur in the corpus proportionally to their shares within the texts of the chosen period [1]. It is commonly supposed that the RNC is well-balanced, which is ensured by the efforts of its developers who have “hand-picked” the texts for the corpus. GBN was created by a very different technique, its composition was not specially designed, so GBN is often faulted for being unbalanced [7, 8].

In [8], a radical opinion is expressed: “Therefore, instead of speaking about general linguistic or cultural change, it seems to be preferable to explicitly restrict the results to linguistic or cultural change ‘as it is represented in the Google Ngram data’”. In fact, the author of [7] is on the same side of the fence, suggesting that a well-balanced corpus is a utopia and that any data obtained based on a corpus reflects the content of that corpus rather than the language state.

If this were really the case, the creation of text corpora, on which many efforts have been focused, would become a little promising activity. Then corpora would be just a set of examples linguists can extract to quote in their articles, and they would not be suitable as a tool for fundamental research in the essence of a language. Fortunately, this is not the case. The best proof of the entire language representation adequacy in large corpora is the reproducibility of results demonstrated on different corpora. Let us give a simple example.

In [10], the changes were considered regarding the frequency of the members of a synonymic row that includes the words *стараться* (try) and *пытаться* (attempt). In Figures 2 and 3, the graphs are shown for the most frequently used words from the inflectional paradigm, i.e., *старался* and *пытался* (both are past-tense third-person singular masculine verbs), for GBN and RNC. The trend of the last two centuries is clearly in evidence for both corpora – *пытался* becomes more frequent than *старался*. The appearance the graphs in GBN is smoother since this corpus is larger. Even the period where the word *пытался* becomes more frequent is the same – around the year 1960. For highly-frequent words, the graphs of GBN and RNC are usually similar, as well.

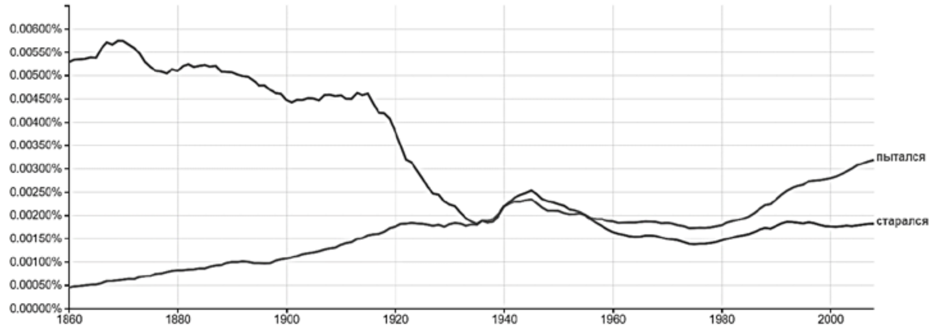


Fig. 5. Frequencies of the words *старался* and *пытался* for GBN

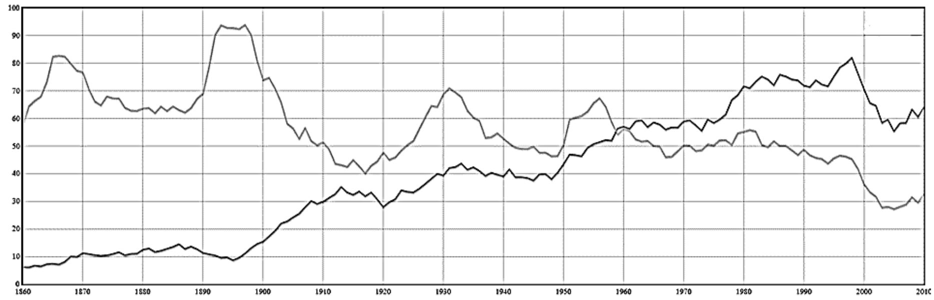


Fig. 6. Frequencies of the words *старался* and *пытался* (black line) for RNC

Such agreement of results obtained on different corpora both validates the results as such and indicates the high quality of the corpora and their consistency. Unfortunately, we cannot directly check everything on GICR within the above timeframe, since almost all GICR texts are dated the 21st century and it has just started growing deeper recently.

However, there is a curious possibility to perform an indirect comparison. Time series generated based on diachronic corpora provide a great opportunity to predict the development of the language. So far, no quantitative predictions regarding language changes have been made based on corpora. In this paper, we are probably making one of the first attempts of this kind. The scheme proposed for extrapolating time series can be useful to the research in various language-specific phenomena.

In Table 1, the GBN-based frequencies of the words *пытался* and *старался* are shown in 1978 and in 2008, as well as the ratio of the former of those values to the latter one. Using the linear regression method, we compute the expected values of the frequencies for the year 2014. That year was chosen to compare our predictions with the data of [6], in which the time interval is limited to the years 2014–2015 being available to the authors at the time of writing their work. The increased number of uses of *пытался* as compared with *старался* over a 30-year period in 1978–2008 allows expecting its further growth by 2014.

Table 1. Known and predicted frequencies of the words *пытался* and *старался* in GBN

Word	1978	2008	Prediction for 2014
пытался	0.00173	0.00318	0.00347
старался	0.00140	0.00181	0.00189
пытался/старался	1.24	1.75	1.84

Let us consider the GICR-based data from [7]. We divide all the GICR sub-corpora into three groups that differ in their genres and styles: 1) Zhurnalny Zal that contains texts from literary journals and is the closest one to GBN; 2) Novosti that contains the texts of another genre, and 3) LiveJournal and VKontakte, both containing texts that fundamentally differ from book texts. Hence, we can expect that data for Zhurnalny Zal will be similar to that of GBN, while the data for Novosti and for the sub-corpora of the third group will be different.

Indeed, the ratio of the *пытался* usage number to the *старался* usage number makes 1.94 in Zhurnalny Zal, 10.41 in Novosti, and 3.30 in the sub-corpora of the third group [7]. Thus, we can see that the text genres and styles are, of course, of great importance. At the same time, for the texts of a similar nature, such as books from GBN and articles from literary journals, the values predicted based on GBN and the real values have turned out to be very close to each other, differing by less than 6%. This also indicates the high quality of the corpora being compared and the possibility to obtain rather correct predictions based on GBN.

Unfortunately, not all the studies performed on GBN can be repeated on RNC or GICR. This is because, unlike GBN, the RNC and GICR corpora are not available to users for downloading. This limits the possibilities of processing the RNC and GICR data with simple queries and does not allow applying complex computer-aided and mathematical data-processing methods that are widely used in contemporary research. The latter ones include measuring the distances between languages at some point in time or between the states of a single language at different time instants, using Kullback-Leibler's metrics [14].

In our opinion, GBN is exactly an example of the best-balanced corpus, as well-balanced as possible. Since all the books from the largest libraries were scanned when creating it, this results in all types of texts being represented in GBN proportionally to the representativeness thereof in the libraries. So GBN is as balanced as the entire human-created library system is balanced.This result cannot be achieved by manually selecting texts.

Let us discuss the term of balance. Here, the balance shall mean the generally balanced corpus as a cohesive whole. In the paper, we introduce a new term: Diachronic balance corpus.

We will apply the concept of "diachronically balanced" to a diachronic corpus that is balanced for any given moment of time, ideally for every year or decade. That is, a corpus sample within any small timespan shall already be balanced as such.

Until now, the problem of creating diachronically balanced corpora has not even been stated. However, the giant volume of GBN, as well as the adopted ideology of total scanning, make this corpus exactly like that. For Russian, over the past decades, the volume of the corpus has made about 1 billion words per year, which is triple as much as the volume of the entire RNC. For English, the corpus volume is 10 times more. Naturally, we cannot prove the diachronic balance of GBN, since there is no operational definition of balance, which would allow us to consider corpora as balanced or unbalanced ones.

Let us consider a specific example demonstrating the degree of the GBN balance as compared with RNC. In the USSR of late 1980s, the word *ускорение* (acceleration) borrowed from physics was embedded in the political vocabulary, which word meant the accelerated development of the country's economy. This term started to be widely used in political essays after April 23, 1985 on which day M.S. Gorbachev declared at the Plenum of Central Committee of the Communist Party of the USSR (CC CPSU) a large-scaled program of reforms under the slogan of accelerating the social and economic development of the country. However, just 2 years later, in January 1987, at the Plenum of the CC CPSU the task was stated aimed at cardinally reconstructing the economy management. The new slogan of *перестройка* (reconstruction) appeared, and *ускорение* started becoming irrelevant. Let us have a look at how frequently the word *ускорение* was used in GBN and RNC.

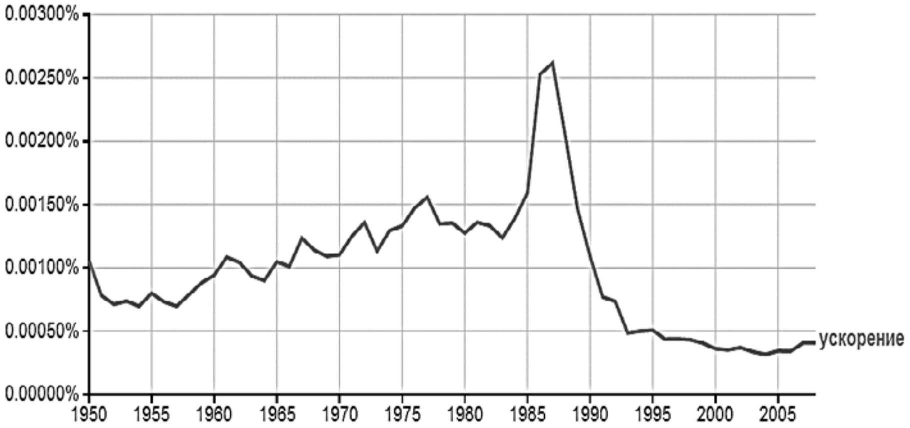


Fig. 7. Frequency of the *ускорение* uses in GBN

In Fig. 4 above, we can see that the sharp rise in the frequency of using the word *ускорение* falls exactly within the year 1985, and its frequency sharply decreases, starting from 1987. Thus, GBN reflects adequately the volume of socially- and politically-focused literature at that time and exactly reflects the processes running in the society. It is also noted in [34] that changes in languages registered in GBN correlate with social events. What about RNC?

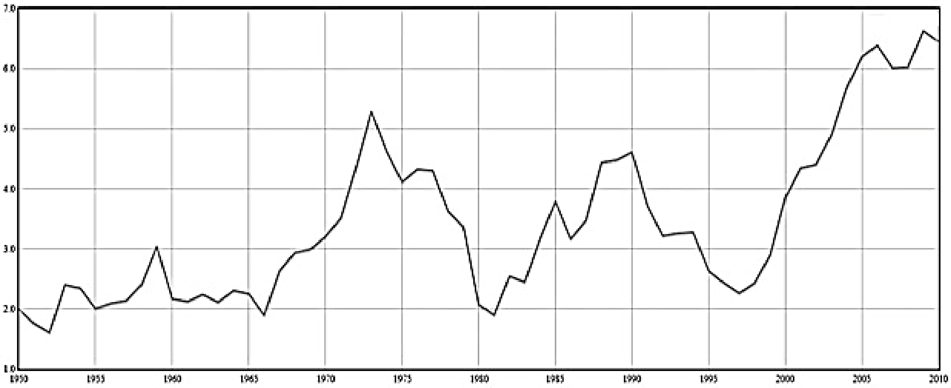


Fig. 8. Frequency of the *ускорение* uses in RNC

In Fig. 5 above, no growth in the frequency of the word *ускорение* can be seen in RNC for that period. Moreover, the frequency of *ускорение* starts falling in 1985 and growing in 1987. At the same time, no political texts are found among the specific ones containing the word ускорение in RNC in those years. This is, of course, just one example. However, it makes it clear how difficult it is to ensure the diachronic balance in manually assembling the corpus, and how naturally it occurs by itself in total digitalizing.

4 Errors in Metadata

In [9], a metadata error was found in the English sub-corpus Fiction. In the first version, many scientific books got into it. This was found based on considering the use in the Fiction corpus the word typical of scientific texts, i.e., 'Figure', compared to the word 'figure' (lowercased) that may occur in literary works, as well. In Fig. 6, you can see well the unnatural growth of the uses of 'Figure', which corresponds in time with the exponential growth of scientific publications.

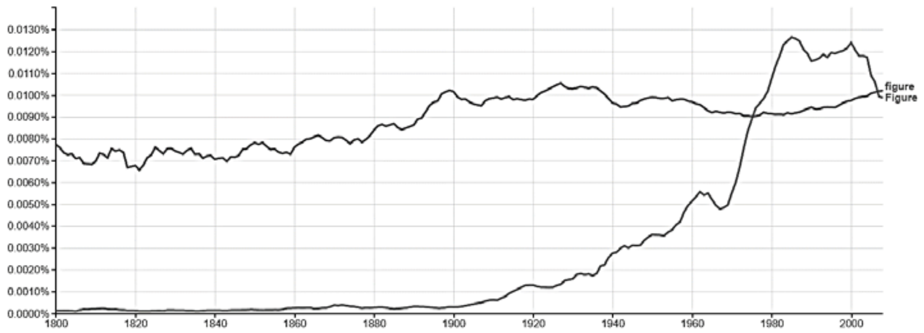


Fig. 9. Frequencies of using 'Figure' and 'figure' in the Fiction corpus, the version of 2009

This was considered in the second version of the corpus, and the books were classified correctly. Therefore, the frequency of using the word ‘Figure’ in the Fiction subcorpus fell 20 times (Fig. 7).

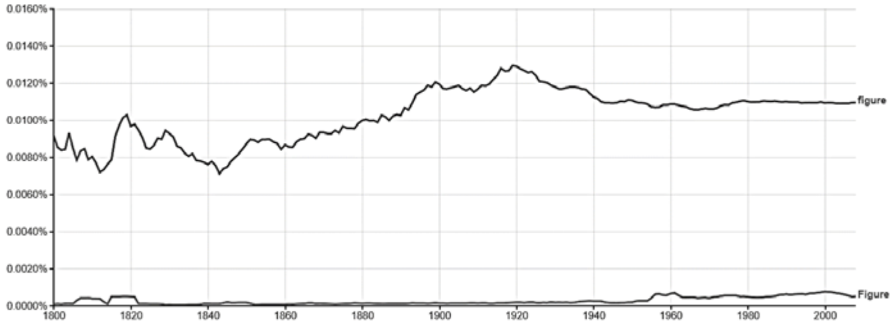


Fig. 10. Frequencies of using ‘Figure’ and ‘figure’ in the Fiction corpus, the version of 2012

Thus, in that case, again, Google rapidly responded to criticism, and the error was corrected.

Further, the authors of [8, 9] returned to using GBN in their studies of the language evolution [11, 34].

5 Using GBN

Despite the problems of the corpus mentioned above, it is widely used in various linguistic and culturological studies. There are over 6,500 articles mentioning GBN in the Google Scholar system. 187 works have already been published within the first 3.5 months of 2019. Any review of those works is far beyond the scope of this paper. However, we would like to note some of the most interesting and typical, in our opinion, trends in research, demonstrating the considerable room for using GBN in Digital Humanities.

In linguistics, the matters have been studied, such as the number and the changes in the number of words within the language vocabulary [12], the dynamics in the “births” and “deaths” of words [13], the speed of evolving the languages and their vocabularies [9, 14], the mechanisms of competing the regular and irregular forms of verbs in English [5], and comparing British English and American English [14].

In psychology, emotions [15–18] and cognitive processes [19–21] have been studied. One of the most popular matters turned out to be the changes in the psychology of collectivism/individualism. Of many works in this area, we would like to note articles [22–25], in which the growth of individualism was tracked in different countries, as exemplified by English, German, Russian, and Chinese.

In social studies, the research has been performed in gender differences and diversity [26, 27] and in global cultural trends [28].

6 Directions in Enhancing the Results

We can propose two ways of enhancing the reliability of the results obtained using GBN. The first one consists in using and comparing all types of data that can be extracted from GBN. This particularly includes considering, along with the word itself, its various inflectional forms [29] and synonyms [24]. In [29], this is exemplified with the German word *eigen* (own, peculiar), which is relatively rare to occur in this form, while it 35 times more frequently occurs in the form of *eigenen*. In [24], it is recommended to use each word with three synonyms selected in the relevant dictionaries of synonyms. Should your research be of intercultural nature, then it is natural to use corpora for several languages represented in GBN [29] in order to compare the dynamics of the frequencies of the same or close terms. For research in English, GBN provides several corpora, such as general English, American English, British English, and the Fiction corpus. They can also be used to compare and verify the results obtained. For example, in [30], the dynamics of the first-person pronoun frequencies can be tracked using both the English corpus and the fiction corpus.

The second way consists in preprocessing raw data provided by GBN. Although this way is rather labour-consuming, it can still be recommended. In [14], the corpus preprocessing is described that consists in removing all tokens (character strings) that are not words. All tokens are deleted that contain numbers or other non-alphabetic characters, except for apostrophes. (The '–' symbol is processed by the GBN system itself.) This is probably especially topical for the languages that have undergone spelling reforms, such as the Russian language. The 1918 reform removed ъ (hard sign) at the ends of masculine words ending with a consonant. To process those words correctly, it would be reasonable to delete ъ at the ends of all such words. This is just a realistic way for a researcher, which allows correctly processing an enormous number of Russian words – practically all masculine nouns.

We can find other systemic changes in spelling the words and correct the corpus accordingly in compliance with the current spelling rules. Replacing ancient orthography with the modern spelling is adopted in RNC. This is reasonable, of course, for the studies only that do not focus on researching in ancient orthography.

It is unreal, of course, to eliminate all the errors in a multibillion-word corpus. Therefore, it would be reasonable to try and apply the recently-developed methods of working with noisy language data [31, 32].

7 Conclusion

Creating very large specialized and multi-use text corpora is important for both theoretical and applied research in linguistics and allied areas of knowledge. Very large text corpora, especially diachronic ones, create fundamentally new opportunities for studies that just could not have been performed without them. GBN corpus presents very accurately both the language changes and the processes occurring in the society and reflected in the language. This allows using this corpus in various humanities research. Diachronic corpora provide a researcher with the opportunities for both describing the language properties observed and reasonably predicting about their further developments.

Creating such corpora is an extremely complicated and labour-consuming activity, and the matters regarding the quality of the corpora created emerge inevitably. If texts recently published using computer-aided techniques are quite “pure,” then the ancient books and periodicals must be scanned followed by recognizing the characters, which inherently leads to errors in the corpus. In this paper, we have considered the case of the currently largest diachronic corpus, GBN. It is shown that the main errors of the earlier version have already been eliminated in the next version of the corpus. The remaining specific minor errors are invalidated in statistical computations on a big data array. However, for Russian, some problems persist, which are related to the ancient spelling and which it would be reasonable to solve.

As to the most important issue regarding the balance/representativeness of GBN, the conceivable case for the fact has been made out that the corpus is highly balanced. GBN is specifically compared with RNC and GICR, which comparison has demonstrated their high consistency. The latest versions of spelling correction systems may be used, as well [33].

Acknowledgment

This research was financially supported by the Russian Foundation for Basic Research (Grant No. 17-29-09163), the Government Program of Competitive Development of Kazan Federal University, and through the State Assignment in the Area of Scientific Activities for Kazan Federal University, agreement No 2.8303.2017/8.9.

References

1. Russian National Corpus. http://www.ruscorpora.ru/. (2019)
2. Russian National Corpus: 2003–2005. Indrik, Moscow. (2005). (in Russian).
3. Russian National Corpus: 2006-2008. V.A. Plungyan, ed. Nestor-Istoriya. St. Petersburg. (2009) (in Russian).
4. Lin, Y., Michel, J.-B., Aiden, E., Orwant, J., Brockman, W., and Petrov, S.: Syntactic Annotations for the Google Books Ngram Corpus. Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics vol, 2: Demo Papers (ACL '12) (2012).
5. Michel, J. et al.: Quantitative Analysis of Culture Using Millions of Digitized Books. Science **331 (**6014), 176–182 (2011).
6. Aiden, E. and Michel, J.-B.: Uncharted Big Data as a Lens on Human Culture. Russian edition: Moscow. AST. 352 p. (2016) (In Russian).
7. Belikov, V.I.: What and how can a linguist get from digitized texts? In: Sibirsky philologichesky zhurnal [Siberian Journal of Philology] (3), 17–34 (2016) (In Russian).
8. Koplenig, A.: The impact of lacking metadata for the measurement of cultural and linguistic change using the Google Ngram data sets – Reconstructing the composition of the German corpus in times of WWII, *Digital Scholarship in the Humanities* **32**, 169–188 (2017). https://doi.org/10.1093/llc/fqv037
9. Pechenick, E.A., Danforth, Ch.M., Dodds, P.Sh., and Barrat, A.: Characterizing the Google Books Corpus: Strong Limits to Inferences of Socio-Cultural and Linguistic Evolution. PLOS ONE. 10 (10): e0137041 (2015).
10. Solovyev, V.D.: Possible mechanisms of change in the cognitive structure of synonym sets. In: Language and Thought: Contemporary Cognitive Linguistics. A collection of articles. Languages of Slavic Culture. Moscow, 478–487 (2015).

11. Pechenick, E.A., Danforth, Ch.M., and Dodds, P.Sh.: Is language evolution grinding to a halt? The scaling of lexical turbulence in English fiction suggests it is not. J. Comput. Science **21**, 24–37 (2017).
12. Petersen, A.M., Tenenbaum, J., Havlin, S., Stanley, H. E., and Perc, M.: Languages Cool as They Expand: Allometric Scaling and the Decreasing Need for New Words, Sci. Rep. 2, 943 p. (2012).
13. Petersen, A.M., Tenenbaum, J., Havlin, S., and Stanley, H.E.: Statistical laws governing fluctuations in word use from word birth to word death. Scientific Reports 2, 313 p. (2012).
14. Bochkarev, V., Solovyev, V., and Wichmann, S.: Universals versus historical contingencies in lexical evolution. J. R. Soc. Interface **11** (101). DOI: 10.1098/rsif.2014.0841. (2014)
15. Acerbi, A., Lampos, V., Garnett, P., and Bentley, R.A.: The expression of emotions in 20th century books. PloS One **8** (3), (2013), e59030. https://doi.org/10.1371/journal.pone.0059030 PMID: 23527080
16. Mohammad, S.M.: From once upon a time to happily ever after: Tracking emotions in mail and books. Decision Support Systems **53** (4), 730–741 (2012).
17. Morin, O. and Acerbi, A.: Birth of the cool: a two-centuries decline in emotional expression in Anglophone fiction. Cognition and Emotion **31** (8), 1663–1675 (2017). https://doi.org/10.1080/02699931. (2016). 1260528 PMID: 27910735
18. Scheff, T.: Toward defining basic emotions. Qualitative Inquiry **21** (2), 111–121 (2015).
19. Ellis, D.A., Wiseman, R., and Jenkins, R.: Mental representations of weekdays. PloS One **10** (8), e0134555 (2015). https://doi.org/10.1371/journal.pone.0134555 PMID: 26288194.
20. Hills, T.T. and Adelman, J. S..: Recent evolution of learnability in American English from 1800 to 2000. Cognition **143**, 87–92 (2015). https://doi.org/10.1016/j.cognition.2015.06.009 PMID: 26117487.
21. Virues-Ortega J. and Pear J.J.: A history of “behavior” and “mind”: Use of behavioral and cognitive terms in the 20th century. The Psychological Record **65** (1), 23–30 (2015).
22. Greenfield, P.M.: The changing psychology of culture from 1800 through 2000. Psychological Science **24** (9), 1722–1731 (2013). https://doi.org/10.1177/0956797613479387 PMID: 23925305
23. Zeng, R. and Greenfield, P.M.: Cultural evolution over the last 40 years in China: Using the Google Ngram Viewer to study implications of social and political change for cultural values. International Journal of Psychology **50** (1), 47–55 (2015). https://doi.org/10.1002/ijop.12125 PMID: 25611928
24. Younes, N. and Reips, U.-D.: The changing psychology of culture in German-speaking countries: A Google Ngram study. International Journal of Psychology **53**, 53–62 (2018). https://doi.org/10.1002/ijop. 12428 PMID: 28474338
25. Velichkovsky, B.B., Solovyev, V.D., Bochkarev, V.V., and Ishkineeva, F.F.: Transition to market economy promotes individualistic values: Analysing changes in frequencies of Russian words from 1980 to 2008. International Journal of Psychology (2017).
26. Del Giudice, M.: The twentieth century reversal of pink-blue gender coding: A scientific urban legend? Archives of Sexual Behavior **41** (6), 1321–1323 (2012). https://doi.org/10.1007/s10508-012-0002-z PMID: 22821170
27. Ye, S., Cai, S., Chen, C., Wan, Q. and Qian, X.: How have males and females been described over the past two centuries? An analysis of Big-Five personality-related adjectives in the Google English Books. Journal of Research in Personality **76**, 6–16 (2018).
28. Bochkarev, V.V., Shevlyakova, A.V., and Solovyev, V.D.: The average word length dynamics as an indicator of cultural changes in society. Social evolution & History **14** (2), 153–175 (2015).
29. Younes, N. and Reips, U.-D.: Guideline for improving the reliability of Google Ngram studies: Evidence from religious terms. PLoS ONE **14** (3), e0213554 (2019). https://doi.org/10.1371/journal.pone.0213554.

30. Twenge, J.M., Campbell, W.K., and Gentile, B.: Changes in pronoun use in American books and the rise of individualism, 1960–2008. Journal of Cross-Cultural Psychology **44** (3), 406–415 (2013).
31. Malykh, V. and Lyalin, V.: Named Entity Recognition in Noisy Domains. In: The Proceedings of the 2018 International Conference on Artificial Intelligence: Applications and Innovations. ISBN: 978-1-7281-0412-6 (2018).
32. Malykh, V. and Khakhulin, T.: Noise Robustness in Aspect Extraction Task. In: The Proceedings of the 2018 Ivannikov ISPRAS Open Conference. (2018). ISBN: 978-1-7281-1275-6.
33. Anisimov, I., Polyakov, V., Makarova, E., and Solovyev, V.: Spelling correction in English: Joint use of bi-grams and chunking. 2017 Intelligent Systems Conference, IEEE Xplore Digital library (2018). https://ieeexplore.ieee.org/document/8324234
34. Koplenig, A.: A fully data-driven method to identify (correlated) changes in diachronic corpora. https://arxiv.org/ftp/arxiv/papers/1508/1508.06374.pdf (2015).

An Approach to Data Preprocessing for the Specialists' Functional State Evaluation

Alexander Yakovlev[1], Viacheslav Matytsin [2], and Xenia Naidenova [3]

[1] Military Medical Academy, Saint Petersburg, Russian Federation
sven-7@mail.ru
[2] Military Medical Academy, Saint Petersburg, Russian Federation
matitsin@list.ru
[3] Military Medical Academy, Saint Petersburg, Russian Federation
ksennaidd@gmail.com

Abstract. The most important stage of diagnostics and forecasting of functional state and professional reliability of operators of complex man-machine systems is correctly to collect and memorize diagnostic data. The authors propose a developed instrumental measuring stand for multimodal registration of data from which the evaluation of the human functional state derived. The proposed solution uses technical and software tools to register big data sets of three types: audio, video and physiological data obtained from polygraph sensors. The algorithms are selected and tested performing primary data transformations to get a set of structured multimodal indicators, expressed in numerical form.

Keywords: human functional state, face recognition, speech recognition

1 Introduction

Functional state is a state of adaptation of a person to specific external and internal conditions. From this fact, it is easy to prove that it is impossible to say whether a person has the good functional state or not according to the measured physiological parameters alone, it is necessary to know individual physiological reserves and external conditions. The same values of physiological characteristics under the same conditions may be normal for one person and not for another. These considerations explain why there is not any conventional classification of physiological states. This also explains the established tradition of being interested in the physiological reserves of human organism and, especially, in the states or moments of transition into some critical states that can be clearly diagnosed (albeit regarding the individuality). These critical states include fatigue (muscle and mental), drowsiness, monotony, various stresses (heat and cold stresses), emotional stress, nervous stress, anxiety, etc. Usually, some dichotomies are considered: "sleepiness – wakefulness", fatigue – working condition, "monotony – lack of monotony" and so on. However, the functional state is also estimated in different multiple nominal scales. Hence we come to recognizing two or more classes of the functional states. Such diagnostic tasks require designing diagnostic rules through machine learning on a training sample.

Estimating the functional state of specialist’s organism (hereinafter – the FSO) means measuring the totality of characteristics of his physiological functions and psychophysiological reactions determining in the different environment conditions of his professional and behavioral activity [1]. The diagram (Fig. 1) presents the interrelation of the basic concepts used in the physiology of labor to solve the tasks of evaluating the specialist’s FCO and the predicting of his efficiency.

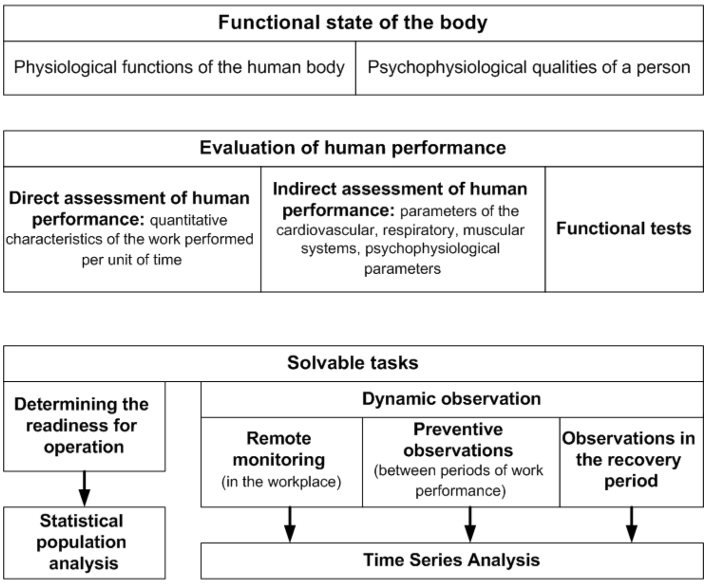

Functional state of the body
Physiological functions of the human body
Psychophysiological qualities of a person
Evaluation of human performance
Direct assessment of human performance: quantitative characteristics of the work performed per unit of time
Indirect assessment of human performance: parameters of the cardiovascular, respiratory, muscular systems, psychophysiological parameters
Functional tests
Solvable tasks
Determining the readiness for operation
Dynamic observation
Remote monitoring (in the workplace)
Preventive observations (between periods of work performance)
Observations in the recovery period
Statistical population analysis
Time Series Analysis

Fig. 1. The interrelation of the basic concepts used in the physiology of labor

The paper proposes a method of synchronous collection of data on physiological reactions of the body and psycho-emotional manifestations of a person, based on the registration of data from a variety of information channels including audio and video signals.

2 The Problem of Assessing the Specialists’ Functional Status

The information channels are combined in three modalities on the principle of registration and mathematical processing of data. First, there are the signals of physiological modality, received via multichannel polygraph registration of human physiological reactions. It is registered the following bio-signals: photoplethysmographic signal, arterial pressure, skin-galvanic reflex, frequency of thoracic and diaphragm respiration, change of posture of body (displacement of the mass center).

Secondly, there are signals of visual modality. Images of a person's face, his facial expressions, postures and gestures with their subsequent processing by means of computer vision and machine learning algorithms allow revealing and formalizing basic classes of psycho-emotional manifestations of a person during his speech activity.

Thirdly, there are signals of acoustic modality. This modality relates to the audio registration of speech signal of a person with simultaneous recording acoustic noise present in the room of study, in order to avoid systematic error. For registering speech signal, it is supposed to use the highly sensitive microphone that allowing to register voice maximally close to natural one.

The most important step in the tasks of diagnosing human condition (the FSO) is the correct collecting and memorizing diagnostic data. A hardware-software stand has been developed, providing synchronous recording physiological, video and acoustic data directly from measurement devices and registration means (hereinafter-the Stand). Its architecture is presented in Fig. 2 and repeats the typical architecture of business analytics systems [2, 3]. It contains the data collected and program components to process data.

The stand accumulates big data sets of the three types (i.e. the three modes): audio data, video data and data obtained from physiological sensors.

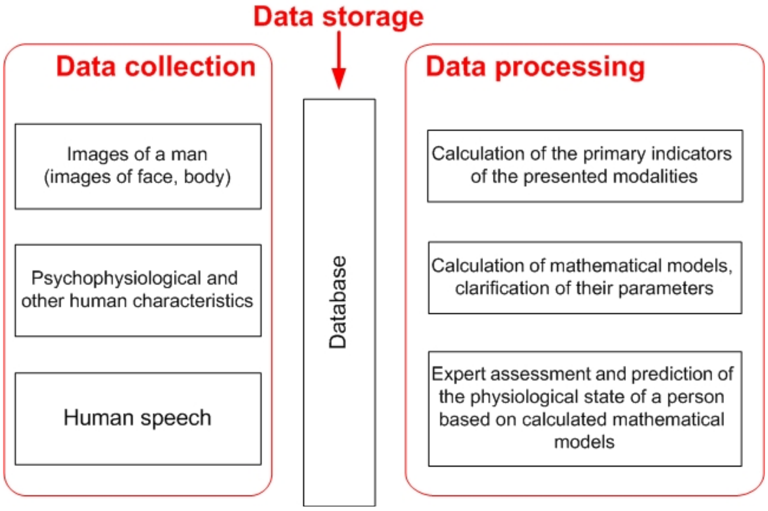

Data storage
Data collection
Images of a man (images of face, body)
Psychophysiological and other human characteristics
Human speech
Database
Data processing
Calculation of the primary indicators of the presented modalities
Calculation of mathematical models, clarification of their parameters
Expert assessment and prediction of the physiological state of a person based on calculated mathematical models

Fig. 2. General structure of the Stand [2], [3]

The database (hereinafter – the DB) stores all the information collected, including also the information of measuring instruments [4]. The data processing part of the Stand is intended for the application of modern machine learning techniques [5, 6] to extract knowledge from collected data.

3 Data Collection

The following devices of the Stand are the sources of data:

- Professional Computer Polygraph (PCP) “Diana-7M”;
- Video complex “Diana-Video”;
- Low-frequency spectrum analyzer UPV.

Professional computer Polygraph “Diana-7M” (manufacturer «Polyconius, Russia») – one of the most advanced Russian polygraphs (lie detectors), adapted to scientific research tasks. “Diana-7M” registers the human functional state data with a set of contact sensors:

- URS – Upper respiratory sensor (pneumatic), registers human chest respiration;
- LRS – Lower respiratory sensor (pneumatic), registers human diaphragm respiration;
- SGR – Sensor of skin-galvanic reaction;
- AP – Arterial pressure sensor;
- PhG – Photoplethysmogram sensor;
- MAS – Motor Activity Sensor;
- MIC – Microphone sensor in the form of a loop.

Sensors 1–6 are connected to a specialized analog-to-digital converter (hereinafter-ADC), the MIC sensor is connected to an external sound card M-Audio M-Trackmk.II. Thus, the “Diana-7M” control panel registers 7 signals of different nature.

Video complex “Diana-Video” (manufacturer “Polykonius”, Russia) provides the registration of non-verbal behavior of a person with the help of high-resolution cameras (two IP video cameras Axis, with resolution 1280 x 720), as well as high-quality synchronized registration of his speech.

The analyzer of sound oscillations R & S UPV (manufacturer Rohde & Schwartz, Germany, No. 48123-11 in the State Register of measuring instruments) with measuring Microphones 378A14 (manufacturer PCB Piezotronics, Inc., USA, No. 61720-15 in the State Register of measuring instruments) records sound vibrations in the near area of a person in the contactless way. In the digital recorder mode, the device can generate two high quality audio files. The device is controlled by the NI-VISA protocol.

It is important to note that the above listed devices register not the indicators themselves, but signals of different modalities or, as they are called, “raw” data. Such “raw” data cannot serve as the source for multimodal analysis without the primary treatment. In this connection, it became necessary to keep “raw” data in a single storage (the DB) for further processing. The task was complicated by the fact that the devices have different interfaces, different format of data presentation, limited capacity of internal memory and different ways of synchronization.

The stand is built on the “client-server” architecture, which allows to overcome the above-mentioned difficulties and ensure correct synchronous collection of “raw” data with the possibility of simultaneous presentation of stimulus material. Fig. 3 shows the scheme of switching the equipment of the Stand.

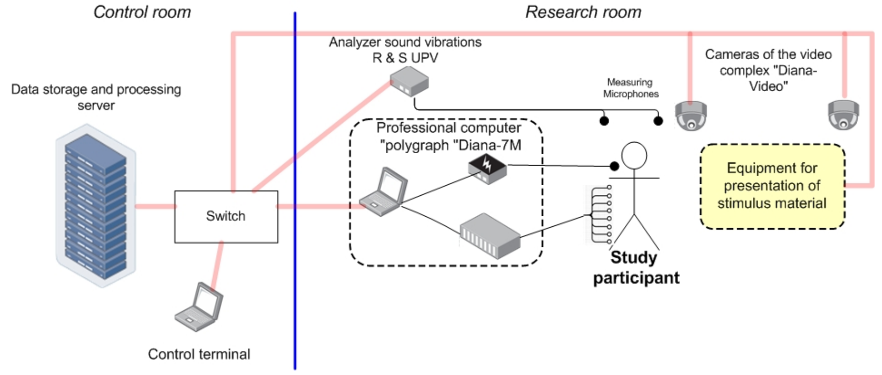

Control room
Research room
Data storage and processing server
Switch
Control terminal
Analyzer sound vibrations R & S UPV
Measuring Microphones
Cameras of the video complex "Diana-Video"
Professional computer "polygraph "Diana-7M
Equipment for presentation of stimulus material
Study participant

Fig. 3. The scheme of switching the equipment of the Stand

Fig. 4 lists the program tasks (hereinafter-PT), implemented in the Stand. They are conditionally grouped into three groups providing a) the preparation of research, b) carrying out research, and c) processing the results of research. All the PTs can be run from the Stand control terminal.

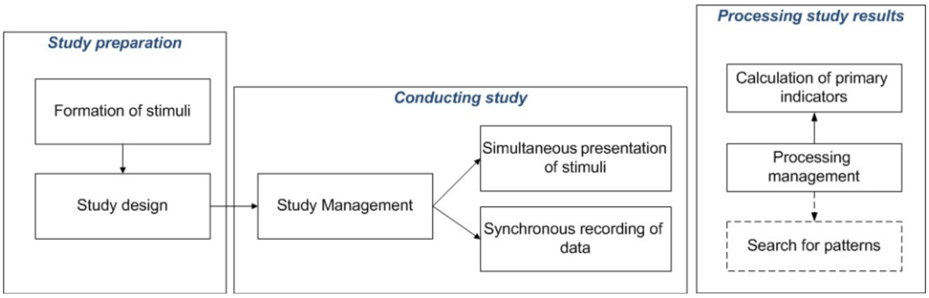


Fig. 4. Program tasks implemented in the Stand

The PT “Formation of the stimulus material” provides the user with an interface allowing to enter a stimulus material in the stand database. The stimulus material can be a video or audio sequence, questionnaires, texts, etc. The order and volume of its presentation to persons is determined during the planning of a concrete study.

The PT “Study design” provides the user with an interface for entering into the database the information about a person or a group of persons, the stimulus materials, the order of their processing, and the concrete recording equipment. The result of this PT is the research plan.

The PT “Study Management” provides the user with the functionality to manage the experiment even he is outside the study room. In case of any deviation from the research plan, the user can suspend or cancel the experiment.

The PT “Synchronous data recording” provides the initiation of devices that register raw data, the obtaining of data from them and the transmission of data over the network to the server. The PT “Synchronous presentation of the stimulus material” provides the synchronism of presenting the stimulus material to a person and the registering data.

The PT “Calculation of primary indicators” converts the “raw” data to a form convenient for further analysis. Due to the fact that the number of primary indicators can reach several thousand, the PT “Processing Management” provides the user the possibility to regulate the composition of calculated indicators and the parameters of their calculation.

The obtained primary indicators must be processed by the machine learning algorithms to search for regularities in them.

4 The Structure of the DB and an Approach to Data Processing

The software of the Stand transforms the collected data to the form convenient for further analysis.

In addition to raw data, three arrays of indicators of different modality (physiological, acoustic, visual) are stored in the DB of the Stand. To store them on the server, it is used DBMS PostgreSQL. The structure of the Stand database is presented in Fig. 5. The central part of the DB is the table “The setup data of the study”. This table associates each person with his audio and video recording, values of measured initial and secondary features (indicators).

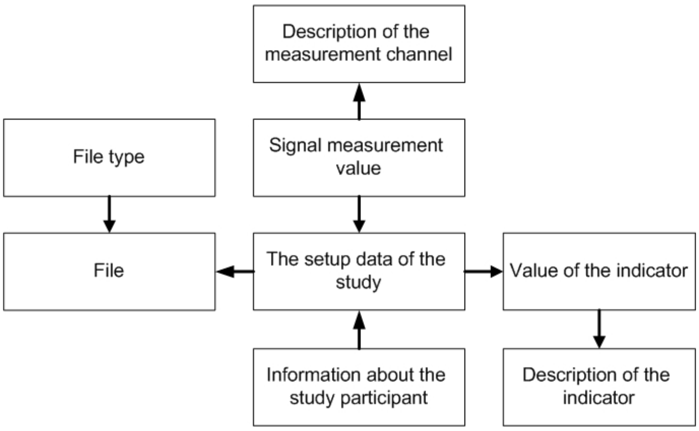

Description of the measurement channel
File type
Signal measurement value
File
The setup data of the study
Value of the indicator
Information about the study participant
Description of the indicator

Fig. 5. The structure of the DB of the Stand

The table “Signal measurement value” stores the values of all the features obtained directly during the experiment referred to the type of corresponding measurement (channel). The table “Value of the indicator” stores the values of all the features obtained as a result of initial and secondary processing. The tables “Description of the measurement channel” and “Description of the indicator” are utilized for storing the general information about each channel and feature.

5 The Approach to Data Processing

General structure of the algorithms for specialist's functional state evaluation and predicting his work capacity is depicted in Fig. 6.

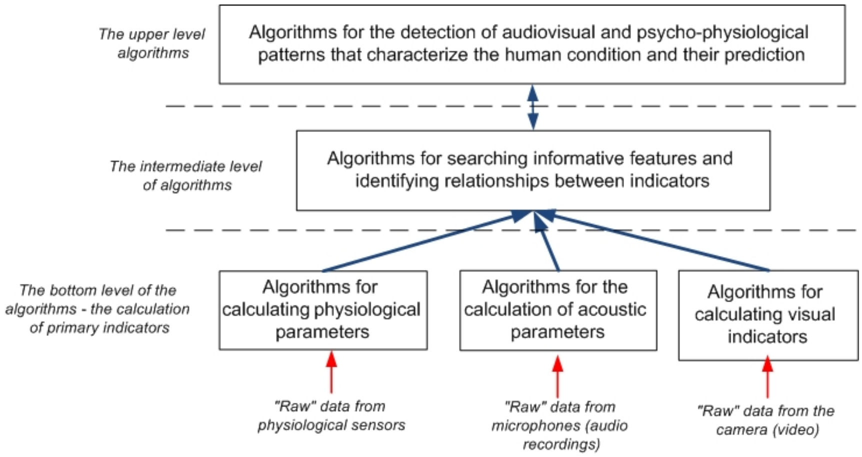

The upper level algorithms
Algorithms for the detection of audiovisual and psycho-physiological patterns that characterize the human condition and their prediction
The intermediate level of algorithms
Algorithms for searching informative features and identifying relationships between indicators
The bottom level of the algorithms - the calculation of primary indicators
Algorithms for calculating physiological parameters
Algorithms for the calculation of acoustic parameters
Algorithms for calculating visual indicators
"Raw" data from physiological sensors
"Raw" data from microphones (audio recordings)
"Raw" data from the camera (video)

Fig. 6. The structure of data processing algorithms

Diana-7M registers the signals of physiological modality from the following sensors: URS, LRS, SGR, AR, PhS, and MAS. Treating these signals are performed with the use of mathematical software GNU Octave (www.gnu.org/software/octave/) [16]. It provides an improved variant for complicated mathematical computations.

Consider the processing of photoplethysmograms for calculating several indicators to evaluate the current state of cardio-vascular system: heart rate (HR), frequency of breathing (FB), value of the interval between beating of the pulse (IBP), and some others. Let us turn to the theory. The method of photoplethysmography is based on detecting periodic changes in the volume of blood vessels and the absorption and reflection of light by hemoglobin associated with the blood flow in the microvascular bed of tissue (measurement of volume pulse of blood with the light or infrared sensor). Three kinds of waves are registered via the plethysmography [7]: waves of the first order ("pulse waves") synchronized with the vibration of heart muscle, waves of the second order ("breathing waves") synchronized with the frequency of breathing (their amplitude and period are usually smaller than amplitude and frequency of the first order waves), waves of the third order reflecting the periodicity in the activity of blood vessels motor center, but they have often aperiodic character.

Thus, in order to measure HR with the help of photoplethysmogram, it is necessary to extract from it the pulse waves and determine the intervals between their peaks. The following stages of plethysmogram's treatment are required:

Stage 1. "Extracting breathing waves". Breathing waves have smaller period than pulse waves, so the polynomial-regression model is used by us for processing the sample or only some part of it [8]. Due to the difference between the periods, the pulse waves are recognized by the regression model as noise and breathing waves – as useful (required) signal.

Stage 2. "Smoothing of the interference of respiratory waves". Since the regression model is applied in accordance to the principle of a sliding window (a certain interval), the received series at the borders of the window has undesirable disturbances and capable to distort pulse waves. These disturbances were smoothed with the help of median filtration [9].

Stage 3. "Extracting useful signal". Pulse wave was extracted by the use of subtraction of breathing wave from initial source signal. Further, the obtained signal was amplified by several decibel to recognize the pulse wave peaks more clearly.

Stage 4. "Calculating indicators". Searching for the pulse wave peaks was performed by extracting all the signal's pieces with the amplitude higher than 10 and finding in these pieces their local extremums; after that the coordinates of extremums were associated with their time markers and the intervals between the markers were calculated. This process permits determining the value of HR in an arbitrary moment of time. The values of HR are stored in the DB with their markers of time. The discrete Fourie transformation is applied to obtain the frequent spectrum of IBP, the analysis of which gives a series of its frequent characteristics.

Processing the physiological signals of chest and diaphragm respirations means evaluating the frequency of respirations and also obtaining the ratio of their amplitudes to reveal which of them predominates in current moment of time.

The evaluation of breathing frequency is reduced to discovering the peaks by analogy with the case of photoplethysmogram treatment. The calculation of the number of peaks per minute permits determining the breathing frequent of a person.

The predominance of chest or diaphragm respiration is established by subtraction the signal LRS from the signal URS. If the difference is greater than zero, then the chest respiration predominates, in the opposite case – the diaphragm respiration.

The other secondary physiological features were also calculated (Table 1).

Table 1. The secondary features of physiological modality

Source signal («raw» data)	Primary features, abbreviation
signal PhG	Heart Rate, HR
	Interval between beating of the pulse (IBP)
	Variability of HR, VHR
signal URS	Frequency of chest respiration
	Amplitude of chest respiration
signal LRS	Frequency of diaphragm respiration
	Amplitude of diaphragm respiration
signal SGR	The level of long-term reaction
	The level of short-term reaction
signal MAS	Frequency of tremor

The library OpenFace 2.0 is used for digital processing the images obtained [10]. Input video flow is partitioned into separate cadres (static images) which are analyzed by sequential transforming in accordance with the algorithm of active models of person's appearance (Active Appearance Models, AAM) [11]. As a result, we have coordinates of 68 points describing the contours of face, nose, eyes, eyebrows, and lips.

The following operations are performed: human face capture, calculating basic anatomical points on a person's face, determining the inclination and turn of face in space, recognizing face motions. As a result of processing each cadre, 714 secondary features are obtained. However only part of them is interesting with respect to their links (binding) with the characteristics of physiological modality. In particular, the indicators of action units (AU) are widely used, sometimes together with speech analysis, for diagnosing neuro-psychiatry disorders, bi-polar disorders, depression and schizophrenia [12–14]. Their ancestor is Pol Exmann, who has created the Facial Action Coding System, FACS) [15]. The group indicators calculated with the use of OpenFace 2.0 contains the values of 21 AUs. These indicators automatically evaluate the main facial emotions based on above referred the Facial Action Coding System.

Digital treatment of images allows to get some physiological indicators. Euler video intensification provides revealing temporal changes in video images that impossible to see with the naked eye. When this method is used, it is possible to visualize blood-filling of tissues, like as micro-motions of head and face of a person.

We utilize the method Viola–Joice to extracting face in the image flow. There are a lot of realizations of this method including the vision library OpenCV. It is cross-platform instrument with open code to use freely by license BSD. We have also applied the Active Appearance Models from the Dlib library completely ready to be used for finding the facial key points of a person.

For parametric representation of a speech (acoustic) signal, it is divided into short segments, and each segment is transformed into a vector of features. As a result, the input signal is represented by a sequence of feature vectors. This approach assumes that speech can be seen as stationary signal at these short intervals. For a more accurate description of the signal, the speech segments are taken with overlapping. The process of creating speech segments is performed using the window method, i.e. by multiplying the signal with some function of the window in order to weaken the breaks at the window borders. Hamming window is usually used for this purpose [16].

To obtain the indicators of the speech signal, the library openSMILE (www.audeering.com/technology/opensmile/) is used. It allows to calculate about 27 thousand indicators of speech signal. However, not all indicators are used, but only part of them, described as the configuration files for the library OpenSMILE. Fig. 7 shows the structure of the calculated indicators or, as they are called, low level descriptors (LLD) [17].

The spectral group low-level descriptors' calculation is based on Fourie transformation. In particular, the Mel Cepstral Coefficients and RASTA-shaped acoustic spectra are used for automatic speaker verification (ASV) [18]. Spectral centroid is one of more frequently used parameters of timbre [19]. The descriptors, describing prosodic, amplitude characteristics of speech are attributed to the energy-dependent group of descriptors. For example, the sums of RASTA-shaped processed acoustic spectra determine the composition of noise depending on frequency [20]. The vocalized group of descriptors describes the voice characteristics and the quality of sound. For example, jitter and shimmer are used not only for analyzing voice pathology, but for describing the general quality of sounds. The frequency of main tone determines the basic frequency characteristics of acoustic data such that human voice tone [18]. The work with the program environment OpenSMILE provides calculating above listed groups low-level acoustic descriptors.

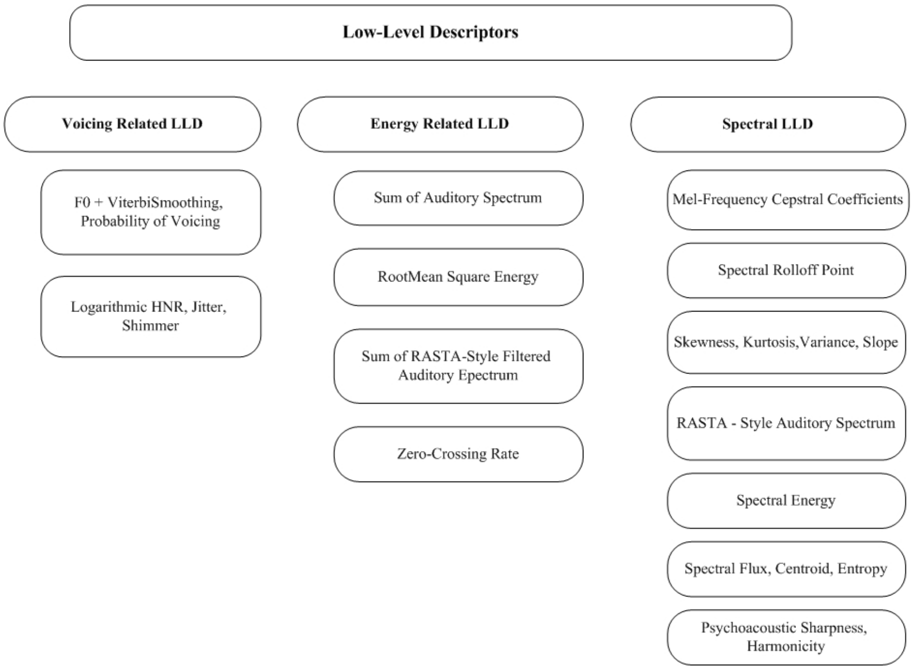

Low-Level Descriptors
Voicing Related LLD
Energy Related LLD
Spectral LLD
F0 + ViterbiSmoothing, Probability of Voicing
Logarithmic HNR, Jitter, Shimmer
Sum of Auditory Spectrum
RootMean Square Energy
Sum of RASTA-Style Filtered Auditory Epectrum
Zero-Crossing Rate
Mel-Frequency Cepstral Coefficients
Spectral Rolloff Point
Skewness, Kurtosis,Variance, Slope
RASTA - Style Auditory Spectrum
Spectral Energy
Spectral Flux, Centroid, Entropy
Psychoacoustic Sharpness, Harmonicity

Fig. 7. The low-level descriptors

6 Testing the Stand

A pilot experiment was conducted to assess the characteristics of the DB and the capabilities of the stand. Fifteen persons were presented a speech load in the form of reading specially prepared texts of the different levels of complexity. During the reading of the texts, the synchronous registration of audiovisual and psychophysiological indicators was carried out.

The resulting array of raw data amounted to 1.6 GB, of which: physiological data volume – 16 MB, acoustic data – 1.2 GB, video data – 440 MB. Such a significant amount of memory occupied by the primary indicators, is explained by the fact that for each indicator in the course of a 10-minute experiment with one tested person, about 15000 measurements are recorded. In this connection, the problem of decreasing the dimension of data, as well as searching for regular dependencies in data were originated. A good example of functional states' identification based on machine learning is the work [21]. The authors used statistical analysis of physiological parameters, including galvanic skin reaction, respiratory frequency, electrocardiography data, body temperature, electromyography data, muscle contraction data. The method of learning is divided into three stages: global learning, adjusting towards an individual, individual testing. The Gradient Boosted Trees algorithm was used as a machine learning method [22].

However, for big data, it is more convenient the approach based on incremental learning Good Maximally Redundant Diagnostic Tests [23]. This approach proposes different decompositions of classification context into sub-contexts, original methods to reduce the search for solution, and the possibility to infer diagnostic rules with the most degree of generalization.

7 Conclusion

The authors propose the instrumental measuring stand for multimodal registration of data from which the evaluation of the human functional state derived. The stand is developing in Military Medical Academy (Saint Petersburg). The proposed solution uses technical and software tools to register big data sets of three types: audio, video and physiological data obtained from polygraph sensors. The algorithms are selected and tested performing primary data transformations to get a set of structured multimodal indicators, expressed in numerical form. A pilot experiment was conducted to assess the characteristics of the DB and the capabilities of the stand. The stand can be utilized not only for diagnostic tasks but also for fundamental investigation in physiology.

References

1. Solodkov: Physiology of sports: functional states of sportsmen and the ways of their restoration. National State University of Physical Culture, Sport and Health named after P. F. Lesgasta. St. Petersburg (2015).
2. Yakovlev, A.V. and Naidenova, X.A.: The concept of using big data technology in modern medicine. News of the Russian military Medical Academy **37** (1), 17–23 (2018) (in Russian).
3. Yakovlev, A. et al.: Substantiation of the information system structure for the analysis of a person's image for medical purposes. Preventive medicine **3**, 306–312 (2017).
4. Yakovlev, A.V. and Naidenova, X.A.: Ontology as a tool of systematization of knowledge about measuring equipment for the monitoring of organism's functional state of servicemen. In: Transactions of XIX All-Russian scientific-practical conference of RAN "Actual problems of protection and safety" **7**, 400–403. Russian Academy of Rocket and Artillery sciences (2016) (in Russian).
5. Naidenova, X.A., Ivanov, V.V., and Yakovlev, A.V.: Discretization of numerical attributes with continuous scales and extraction of concept knowledge from experimental data. In: the 9th national conference on artificial intelligence with international participation **1**, 145–153. Moscow: Physmatlit (2004) (in Russian).
6. Naidenova, X.A., Ivanov, V.V., and Yakovlev, A.V.: Discretization of Numerical Attributes and Extraction of Concept Knowledge from Data. In: Advances in Data Mining and

Knowledge Discovery: Abstracts of Conference “Mathematical Methods for Learning”, 54–55. Italy: Como (2004).

7. Omelchenko, A.V. and Fedorov, A.V.: Estimation of polynomial regression coefficients on the aggregate of realizations. Radio electronics and informatics **1**, 23–28 (2009).
8. Petrovskiy, V.: Pletizmography. In: Petrov B.V. (ed.) Large Medical Encyclopedia. 3rd Edition, V. 19. Moscow: Soviet Encyclopedia (1974) (in Russian).
9. Median filtration. In: National Library of N.UH. Bauman [Electronic Resource].
10. Baltrusaitis, T., Robinson P., and Morency, L.-P.: OpenFace: An open source facial behavior analysis toolkit In: Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV), 1–10. IEEE (2016).
11. Keomany, J. and Marcel, S.: Active Shape Models Using Local Binary Patterns. IDIAP Research Institute (2006).
12. Bersani, G. et al.: Facial expression in patients with bipolar disorder and schizophrenia in response to emotional stimuli: a partially shared cognitive and social deficit of the two disorders. Neuropsychiatr Dis Treat. **9**, 1137–1144 (2013).
13. Cohn, J.F. et al.: Detecting depression from facial actions and vocal prosody In: the 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops, 1–7 (2009).
14. Hamm, J. et al.: Dimensional information-theoretic measurement of facial emotion expressions in schizophrenia. Schizophrenia research and treatment **2014**, paper 243907 (2014).
15. Ekman, P., Friesen, W.V., and Hager, J.C.: Facial action coding system. In: A Human Face. Salt Lake City, Utah (2002).
16. Kipjatkova, I.S., Ronzhin, A.L., and Karpov, A.A.: Automatic processing of colloquial Russian speech. GUAP, St. Petersburg (2013).
17. Velichko, A.N., Bukov, V. Yu., and Karpov, A.A.: Analytical review of computer paralinguistic systems for automatic detection of lies in human speech. Information and Control Systems **5** (90), 30–41 (2017).
18. Weninger, F. et al.: On the Acoustics of Emotion in Audio: What Speech, Music, and Sound have in Common. Front Psychol, 4 (2013).
19. Peeters, G. et al.: The Timbre Toolbox: Extracting audio descriptors from musical signals. The Journal of the Acoustical Society of America **130** (5), 2902–2916 (2011).
20. Rakhmanenko, I., and Meshyeryakov, R.: Analysis of identification characteristics in speech data by means of GMM-UBM System of verification of the announcer. Transaction of Spiniran **3** (52), 32–50 (2017).
21. Lobacheva, E., Galatenko, Y. et al.: Automated real-time classification of functional states based on physiological parameters. Social and Behavioral Sciences **86**, 373–378 (2013).
22. Friedman, J.H.: Greedy function approximation: a gradient boosting machine. Ann. Statist **29**, 1189–1232 (2001).
23. Naidenova, X., Parkhomenko, V., Shvetsov K., Yusupov, V., and Kusina, R.: Modification of good tests in dynamic contexts: application to modelliing intellectual development of cadets. In: M.O. Aciego, D.I. Ignatov, A. Lepsky (eds), The Soft Computing Applications and Knowledge Discovery. The proceedings of the Second International Workshop, 63–74 (2016).

DATA ANALYSIS IN ASTRONOMY

Use of Machine Learning for Anomaly Detection Problem in Large Astronomical Databases

Konstantin Malanchev[1,2,*], Alina Volnova[3], Matwey Kornilov[1,2,+], Maria Pruzhinskaya[1,++], Emille Ishida[4], Florian Mondon[4], and Vladimir Korolev[5,6]

[1] Lomonosov Moscow State University, Sternberg Astronomical Institute, Universitetsky pr. 13, Moscow, 119234, Russia
* malanchev@sai.msu.ru
+ matwey@sai.msu.ru
++ pruzhinskaya@gmail.com

[2] National Research University Higher School of Economics, 21/4 Staraya Basmannaya Ulitsa, Moscow, 105066, Russia
[3] Space Research Institute of the Russian Academy of Sciences (IKI), 84/32 Profsoyuznaya Street, Moscow, 117997, Russia
[4] Université Clermont Auvergne, CNRS/IN2P3, LPC, F-63000 Clermont-Ferrand, France
[5] Central Aerohydrodynamic Institute, 1 Zhukovsky st, Zhukovsky, Moscow Region, 140180, Russia
[6] Moscow Institute of Physics and Technology, 9 Institutskiy per., Dolgoprudny, Moscow Region, 141701, Russia

Abstract. In this work, we address the problem of anomaly detection in large astronomical databases by machine learning methods. The importance of such study is justified by the presence of a large amount of astronomical data that cannot be processed only by human resource. We focus our attention on finding anomalous light curves in the Open Supernova Catalog. Few types of anomalies are considered: the artifacts in the data, the cases of misclassification and the presence of previously unclassified objects. On a dataset of ~ 2000 supernova (SN) candidates, we found several interesting anomalies: one active galactic nucleus (SN2006kg), one binary microlensing event (Gaia16aye), representatives of rare classes of SNe such as super-luminous supernovae, and highly reddened objects.

Keywords: Machine learning; Isolation forest; Gaussian processes; Supernovae; Transients

1 Introduction

During the last couple of decades, astronomy eventually became the source of huge amounts of data produced by different dedicated surveys and experiments, which require careful processing to extract valuable information. Gigabytes of data are collected daily in every domain of electromagnetic spectrum: in high-energy range [1], optics [2, 3], and radio [4], as well as in cosmic particles window [5] and gravitational waves [6, 7]. The search for yet unknown statistically significant features of astronomical objects, as well as the distinction of real features from processing artifacts is an important problem of the automated data analysis.

Supernovae (SNe) are among the most numerous objects discovered in astronomy, and their total amount increases by several thousand per year. These objects help to solve many astronomical puzzles: they produce the majority of heavy chemical elements [8] and high energy cosmic rays [9], they trigger star formation in galaxies [10]. Moreover, the study of different types of SNe allows us to probe the composition and distance scale of the Universe imposing strong constraints on the standard cosmological model [11]. In the last 15 years several large surveys have already gathered many observational data on SNe and their candidates (Carnegie Supernova Project – CSP [12], the Panoramic Survey Telescope and Rapid Response System – Pan-STARRS [13] and the Dark Energy Survey – DES [14]). The surveys of new generation, like Large Synoptic Survey Telescope [15], will produce data of unprecedented volume and complexity.

The exponential growth of astronomical data volume makes the use of machine learning (ML) methods inevitable in this field [16]. Most of the ML efforts in astronomy are concentrated in classification [e.g., 17] and regression [e.g., 18] tasks. A large variety of ML methods were applied to supervised photometric SN classification problem [19–21] and unsupervised characterization from spectroscopic observation [e.g., 22].

Astronomical anomaly detection is the field where ML methods may be used quite effectively taking into account the enormous amount of data that has been gathered, however, they have not been fully implemented yet. Barring a few exceptions, most of the previous studies may be divided into only two different trends: clustering [23] and subspace analysis [24] methods. More recently, random forest algorithms have been used extensively by themselves [25] or in hybrid statistical analysis [26]. Although all of this has been done to periodic variables there is not much done for transients and even less for supernovae.

Supernovae surveys detect hundreds of SNe candidates per year, but the lack of spectroscopic information makes the processing algorithms to classify discovered SNe basing on secondary features (proximity to the galaxy, monotonous flux changing with time, absolute magnitude, etc.) Anomaly detection may solve two problems: (a) minimize the contamination of non-SNe in large supernova databases, and (b) find inside the SNe data rare or new classes of objects with unusual properties.

In this paper, we suggest the algorithm of anomaly detection using the isolation forest method and basing on real photometrical data from the Open Supernova Catalog (OSC) [27, 28].

2 Data Preprocessing

2.1 Open Supernova Catalog

The Open Supernova Catalog [27] is constructed by combining many publicly available data sources. It includes many catalogs and surveys, such as Pan-STARRS [29], the SDSS Supernova Survey [30], the All-Sky Automated Survey for Supernovae (ASAS-SN [31]), the intermediate Palomar Transient Factory (iPTF [32]) among others, as well as information from individual studies. It represents an open repository for supernova metadata, light curves, and spectra in an easily downloadable format. This catalog also includes some contamination from non-SN objects. It contains data for more than 55000 SNe candidates among which ~13000 objects have >10 photometric observations and for ~7500 spectra are available.

The catalog stores the light curves (LCs) data in different magnitude systems. Since we need a homogeneous data sample, we extracted only the LCs in *BRI*, *gri*, and *g'r'i'* filters. We assume, that g'r'i' filters are close enough to gri filters to consider them as the same filters. We also transform BRI magnitudes to gri using the Lupton's photometrical equations [33]. We also require a minimum of three photometric points in each filter with 3-day binning. After this first cut, our sample consists of ~3000 objects.

2.2 Light Curves Approximation

Traditionally, ML algorithms require a homogeneous input data matrix which, unfortunately, is not the case with supernovae. A commonly used technique to transform unevenly distributed data into a uniform grid is to approximate them with Gaussian processes (GP [34]). Usually, each light curve is approximated by GP independently. However, in this study we use a Multivariate Gaussian Process [35] approximation. For each object it takes into account the correlation between light curves in different bands, approximating the data by GP in all filters in a one global fit (for details see Kornilov et al. 2019, in prep.). As an approximation range we chose [–20; +100] days. We also extrapolated the GP approximation to fill this range if needed. With this technique we can reconstruct the missing parts of LC from its behavior in other filters.

Gaussian process is based on the so-called kernel, a function describing the covariance between two observations. The kernel used in our implementation of Multivariate Gaussian Process is composed of three radial-basis functions

$$k(t_1, t_2) = e^{\left(\frac{-(t_2-t_1)^2}{2l_i^2}\right)},$$

where i denotes the photometric band, and l_i are the parameters of Gaussian process to be found from the light curve approximation. In addition, Multivariate Gaussian Process kernel includes 6 constants, three of which are unit variances of basis processes and three others describe their pairwise correlations. Totally, Multivariate Gaussian Process has 9 parameters to be fitted.

Once the Multivariate Gaussian Process approximation was done, we visually inspected the resulting light curves. Those SNe with unsatisfactory approximation were removed from the sample (mainly the objects with bad photometric quality). Since each object has its own flux scale due to the different origin and different distance, we normalized the flux vector by its maximum value. Based on the results of this approximation, for each object we extracted the kernel parameters, the log-likelihood of the fit, LC maximum and normalized photometry in the range of [−20, +100] days with 1-day interval relative to the maximum. These values were used as features for the ML algorithm (Sect. 3). Our final sample consists of 1999 objects, ∼30% of which have at least one spectrum in the OSC. Less than 5% of our sample have <20 photometric points in all three filters.

2.3 Dimensionality Reduction

After the approximation procedure, each object has 374 features: 121×3 normalized fluxes, the LC flux maximum, 9 fitted parameters of the Gaussian process kernel, and the log-likelihood of the fit.

We apply the anomaly detection algorithm not only to the full data set but also to the dimensionality-reduced data. The reason for this is that the initial high dimensional feature space can be too sparse for the successful performance of the isolation forest algorithm. We applied t-SNE [36], a variation of the stochastic neighbor embedding method [37], for the dimensionality reduction of the data. As a result of the dimensionality reduction, we obtain 8 separate reduced data sets corresponding to 2 to 9 t-SNE features (dimensions).

3 Anomaly Detection

3.1 Isolation Forest

Isolation forest [38, 39] is an outlier detection technique that identifies anomalies instead of normal observations. It is built on an ensemble of random isolation trees. Each isolation tree is a space-partitioning tree similar to the widely known Kd-tree [40]. However, in contrast to the Kd-tree, a space coordinate (a feature) and a split value are selected at random for every node of the isolation tree. The tree is built until each object of a sample is isolated in a separate leaf – the shorter path corresponds to a higher anomaly score. For each object, the measure of its normality is the arithmetic average of the depths of the leaves into which it is isolated. The idea of identifying normal data vs. anomalies is presented in Fig. 1.

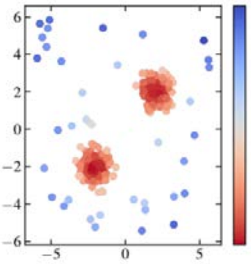


Fig. 1. This `scikit-learn` example presents the generated 2D dataset. Regions of high density are normal data while the outliers are spread around, which is also illustrated by the colour – bluer colour means more anomalous behaviour

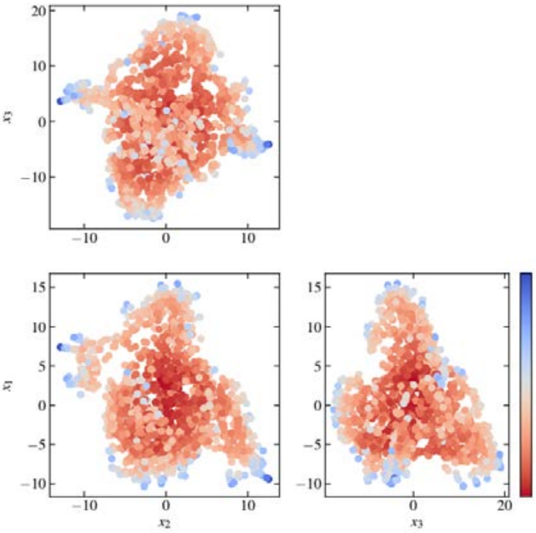


Fig. 2. Three-dimensional t-SNE reduced data after application of the isolation forest algorithm. Each point represents a supernova light curve from the data set projected into the three-dimensional space with the coordinates (x1, x2, x3). The intensity of the colour indicates the anomaly score for each object as estimated by the isolation forest algorithm – bluer colour means more anomalous light curve behaviour

We run the isolation forest algorithm on 10 data sets:

I. data set of 364 photometric characteristics (121×3 normalized fluxes, the LC flux maximum),

II. data set of 10 parameters of the Gaussian process (9 fitted parameters of the kernel, the log-likelihood of the fit),

III. 8 data sets obtained by reducing 374 features to 2–9 t-SNE dimensions (Sect. 2.3).

For each data set we obtained a list of anomalies. Contamination levels were set to 1% (20 objects with highest anomaly score) for data sets I and II. For all data sets in case III we considered 2% contamination (40 objects with highest anomaly score). This larger contamination was chosen to take into account the influence of the dimensionality reduction step in the final data configuration. Given different representations of the data and the stochastic nature of the isolation forest algorithm, the same object can be assigned a different anomaly score depending on how many t-SNE dimensions are used. Thus, only those objects which were listed within the 2% contamination in at least 2 of the data sets in case III were included in the resulting group of objects to further astrophysical analysis.

An example of the isolation forest algorithm applied to the three-dimensional reduced data set is shown in Fig. 3.

3.2 Evaluation of t-SNE Technique

Since t-SNE is a stochastic technique, we have also taken additional precautions to ensure that the resulting anomaly list does not depend on the t-SNE initial random state. For each number of dimensions, we run the t-SNE 1000 times. Then, the isolation forest algorithm is applied to the newly obtained reduced dataset and a list of anomalies is produced. Next, we counted how often each supernova is listed in the anomaly list. Fig. 3 shows the distribution of supernovae by the frequency of appearance in anomaly list for the three-dimensional t-SNE reduced dataset (filled bars). The y-axis is normalized to the total number of runs.

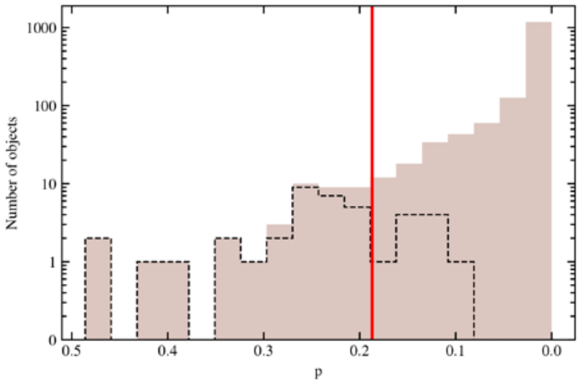


Fig. 3. The distribution of supernovae by the frequency of appearance in anomaly list for the three-dimensional t-SNE reduced dataset (filled bars). The red line denotes 2% of supernovae with the highest anomaly score contamination (40 objects). The distribution of supernovae that were subjected to the further analysis as anomalies in this work are marked by dashed line. The y-axis is normalized to the total number of runs (i.e., 1000)

4 Results

The isolation forest algorithm found ~100 anomalies among 1999 objects from all our samples. We inspected visually the LCs of selected anomalies and analyzed them using other publicly available information. Basing on this analysis, we decided whether the object is an anomaly or not. Among the detected anomalies, there are few cases of miss-classifications, representatives of rare classes of SNe and highly reddened objects. Here we list a few particular cases.

4.1 Peculiar Supernovae

By their spectral and photometric characteristics, the "normal" supernovae are historically divided into two wide types: Type I and Type II. The more recent classification distinguishes Type Ia, Ib, Ic, IIn, IIb, IIP, IIL supernovae. In terms of physics Type Ia SNe are a thermonuclear explosion of a white dwarf which mass exceeded the Chandrasekhar limit either due to accretion from a companion star or by a merging of two white dwarfs [41–43]. These SNe tend to have approximately the same luminosity in maximum and are considered as standard candles for cosmological scale estimates [44, 45]. However, the class of SNe Ia is not homogeneous – some of SNe are on average by 0.2–0.3 magnitudes brighter in maximum than others and some of them are on the contrary subluminous and fast-declining [46, 47]. The presence of non-standard SNe Ia in cosmological samples may introduce a systematic bias [48].

Other types of SNe mark the death of massive stars during the collapse of the core. The envelopes of these stars extend to hundreds of solar radii and contain large amounts of hydrogen (Type II). More massive progenitors of core-collapse supernovae can lose mass by the stellar wind and end their lives losing all (Ib) or part of the hydrogen envelope (IIb). An even more effective stellar wind can blow out not only the hydrogen but also the helium envelope (Ic).

SN2013cv is a peculiar Type Ia supernova with a large peak optical and UV luminosity and with an absence of iron absorption lines in the early spectra. It was suggested [49] to be an intermediate case between the normal and super-Chandrasekhar events.

SN2016bln/iPTF16abc belongs to a subtype of over-luminous Type Ia SNe. Its early-time observations show a peculiar rise time, non-evolving blue colour, and unusual strong C II absorption. These features can be explained by the ejecta interaction with nearby, unbound material or/and significant ^{56}Ni mixing within the SN ejecta [50].

SN2016ija was first suggested to be an early time 91T-like SN Ia with few features and red continuum. It has been also associated to the outburst in an obscured luminous blue variable, an intermediate luminosity red transient or a luminous red nova [51]. The subsequent spectroscopic follow-up revealed broad H_α and calcium features, leading to a classification as a highly extinguished Type II supernova [52].

4.2 Superluminous SNe

Superluminous SNe are supernovae with an absolute peak magnitude M<–21 mag in any band. According to [53] SLSN can be divided into three broad classes: SLSN-I without hydrogen in their spectra, hydrogen-rich SLSN-II that often show signs of interaction with CSM, and finally, SLSN-R, a rare class of hydrogen-poor events with slowly evolving LCs, powered by the radioactive decay of ^{56}Ni. SLSN-R are suspected to be pair-instability supernovae: the deaths of stars with initial masses between 140 and 260 solar masses. Our isolation forest algorithm found 4 SLSNe in our samples.

4.3 Misclassifications

SN2006kg was first classified as a possible Type II SN [54]. It is also appeared as Type II spectroscopically confirmed supernova in table 6 of [55]. However, further analysis of 3.6-m New Technology Telescope spectrum revealed that SN2006kg is an active galactic nucleus [56, 30].

Gaia16aye is an object with the most non-SN-like behavior among our set of outliers. In [57] it was reported to be a binary microlensing event – gravitational microlensing of binary systems – the first ever discovered towards the Galactic Plane.

Our analysis also revealed that 16 of detected anomalies (all from the SDSS SN candidate catalog [30]) are likely to be stars or quasars. First, we did not find any signature of supernovae on the corresponding multicolour light curves. Second, according to SDSS DR15 [58], 10 of these objects are denoted as STAR. The other 6 objects have a BOSS [59] spectrum with class "QSO" and have high redshifts.

More detailed analysis of the detected anomalies is presented in [60].

5 Conclusions

The amount of astronomical data increases dramatically with time and is already beyond human capabilities. The astronomical community already has dozens of thousands of SN candidates, and LSST survey will discover over ten million supernovae in the forthcoming decade [61]. Only a small fraction of them will receive a spectroscopic confirmation. This motivates a considerable effort in photometric classification of supernovae by types using machine learning algorithms. There is, however, another aspect of the problem: any large photometric SN database would suffer from the non-SN contamination (novae, kilonovae, GRB afterglows, AGNs, etc.). Moreover, the database will inevitably contain the astronomical objects with unusual physical properties – anomalies. In this study, we show that the isolation forest algorithm may be rather efficient in solving this problem. This method identified ~100 potentially interesting objects from 1999 supernova candidates extracted from the Open Supernova Catalog, ~30% of which were confirmed to be non-SN events or representatives of the rare SN classes. Among these objects, we report for the first time the 16 star/quasar-like objects misclassified as SNe.

It is important to note that these results are not expected to be complete. There are several known SLSNe in our sample, which were not identified as anomalies, and several objects with very distinguishing features, which do not affect the LC shape significantly, so the algorithm missed them. This may indicate some defects in Gaussian Processes approximation of initial observed LCs. Nevertheless, the above results provide clear evidence of the effectiveness of automated anomaly detection algorithms for photometric SN light curve analysis. This approach may be crucial for future surveys, like LSST, when the enormous amount of data make the search of outliers impossible for human abilities.

The code of this work and the data are available at http://xray.sai.msu.ru/snad/.

Acknowledgements

M. Pruzhinskaya and M. Kornilov are supported by RFBR grant according to the research project 18-32-00426 for anomaly analysis and LCs approximation. K. Malanchev is supported by RBFR grant 18-32-00553 for preparing the Open Supernova Catalog data. E.E.O. Ishida acknowledges support from CNRS 2017 MOMENTUM grant and Foundation for the advancement of theoretical physics and Mathematics “BASIS”. A. Volnova acknowledges support from RSF grant 18-12-00522 for analysis of interpolated LCs. We used the equipment funded by the Lomonosov Moscow State University Program of Development. The authors acknowledge the support from the Program of Development of M.V. Lomonosov Moscow State University (Leading Scientific School “Physics of stars, relativistic objects and galaxies”). This research has made use of NASA’s Astrophysics Data System Bibliographic Services and following Python software packages: NumPy [62], Matplotlib [63], SciPy [64], pandas [65], and scikit-learn [66].

References

1. GBM/Fermi Homepage, https://fermi.gsfc.nasa.gov/science/instruments/gbm.html
2. SDSS-DR12 Homepage, https://www.sdss.org/
3. Gaia project Homepage, http://sci.esa.int/gaia/
4. Event Horizon Telescope Homepage, https://eventhorizontelescope.org/
5. IceCUBE observatory Homepage, https://icecube.wisc.edu/
6. LIGO Homepage, https://www.ligo.caltech.edu/
7. Virgo observatory Homepage, www.virgo-gw.eu/
8. Nomoto, K., Kobayashi, C., and Tominaga, N.: Nucleosynthesis in Stars and the Chemical Enrichment of Galaxies. ARA&A **51**, 457 (2013).
9. Morlino, G.: High-energy cosmic rays from Supernovae. In Handbook of Supernovae, ed. Athem W. Alsabti and Paul Murdin, 1711 p. (2017).
10. Chiaki, G., Yoshida, N., and Kitayama, T.: Low-mass star formation triggered by early Supernova explosions. ApJ **762**, 50 (2013).
11. Scolnic, D.M., Jones, D.O., Rest, A., et al.: The complete light-curve sample of spectroscopically confirmed SNe Ia from Pan-STARRS1 and cosmological constraints from the combined pantheon sample. ApJ **859**, 101 (2018).
12. Carnegie Supernova Project Homepage, https://csp.obs.carnegiescience.edu/
13. Panoramic Survey Telescope and Rapid Response System Homepage, https://pan-starrs.stsci.edu/
14. Dark Energy Survey Homepage, https://www.darkenergysurvey.org/

15. Large Synoptic Survey Telescope Homepage, https://www.lsst.org/
16. Ball, N.M. and Brunner, R.J.: Data mining and machine learning in astronomy. International Journal of Modern Physics D **19**, 1049 (2010).
17. Ishida, E.E.O., Beck, R., González-Gaitán, S., et al.: Optimizing spectroscopic follow-up strategies for supernova photometric classification with active learning. MNRAS **483**, 2 (2019).
18. Beck, R., Lin, C.A., Ishida, E.E.O., et al.: On the realistic validation of photometric redshifts. MNRAS **468**, 4323 (2017).
19. Brunel, A., Pasquet, J., Pasquet, J., et al.: A CNN adapted to time series for the classification of Supernovae. arXiv e-prints, p. arXiv:1901.00461 (2019).
20. Pasquet, J., Pasquet, J., Chaumont, M., and Fouchez D.: PELICAN: deeP architecturE for the LIght Curve ANalysis. arXiv e-prints, p. arXiv:1901.01298 (2019).
21. Möller, A. and de Boissière, T.: SuperNNova: an open-source framework for Bayesian, Neural Network based supernova classification. arXiv e-prints, p. arXiv:1901.06384 (2019).
22. Muthukrishna, D., Parkinson, D., and Tucker, B.: DASH: Deep learning for the automated spectral classification of Supernovae and their Hosts. arXiv e-prints, p. arXiv:1903.02557 (2019).
23. Rebbapragada, U., Protopapas, P., Brodley, C.E., and Alcock, C.: Finding anomalous periodic time series. Machine Learning **74**, 281 (2009).
24. Henrion, M., Hand, D.J., Gandy, A., Mortlock, D.J.: CASOS: a subspace method for anomaly detection in high dimensional astronomical databases. Statistical Analysis and Data Mining: The ASA Data Science Journal **6**, 53 (2013).
25. Baron, D., Poznanski, D.: The weirdest SDSS galaxies: results from an outlier detection algorithm. MNRAS **465**, 4530 (2017).
26. Nun, I., Pichara, K., Protopapas, P., and Kim, D.-W.: Supervised detection of anomalous light curves in massive astronomical catalogs. ApJ **793**, 23 (2014).
27. Guillochon, J., Parrent, J., Kelley, L.Z., and Margutti, R.: An open catalog for Supernova Data. ApJ **835**, 64 (2017).
28. Open Supernova Catalog Homepage, https://sne.space/
29. Chambers, K.C., Magnier, E.A., Metcalfe, N., et al.: The Pan-STARRS1 Surveys. arXiv e-prints, arXiv:1612.05560 (2016).
30. Sako, M., Bassett, B., Becker, A.C., et al.: The data release of the Sloan Digital Sky Survey-II Supernova Survey. PASP **130**, 064002 (2018).
31. Holoien, T.W.-S., Brown, J.S., Vallely, P.J., et al.: The ASAS-SN bright supernova catalogue-IV. 2017. MNRAS **484**, 1899 (2019).
32. Cao, Y., Nugent, P. E., and Kasliwal, M.M.: Intermediate palomar transient factory: realtime image subtraction pipeline. PASP **128**, 114502 (2016).
33. Lupton’s transformation equations for SDSS,
http://www.sdss3.org/dr8/algorithms/sdssUBVRITransform.php
34. Rasmussen, C.E. and Williams, C.K.I.: Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning). The MIT Press (2005).
35. Multivariate Gaussian Processes code, https://github.com/matwey/gp-multistate-kernel
36. Maaten, L. v. d. and Hinton, G.: Visualizing data using t-SNE. Journal of Machine Learning Research **9**, 2579 (2008).
37. Hinton, G.E. and Roweis, S.T.: Advances in neural information processing systems, 857–864 (2003).
38. Liu, F.T., Ting, K.M., and Zhou, Z.-H.: In 2008 Eighth IEEE International Conference on Data Mining, 413–422 (2008).

39. Liu, F.T., Ting, K.M., and Zhou, Z.-H.: Isolation-based Anomaly Detection. ACM Trans. Knowl. Discov. Data **6**, 1 (2012).
40. Bentley, J.L.: Multi dimensional binary search trees used associative searching. Commun. ACM **18**, 509 (1975).
41. Whelan, J. and Iben, Jr. I.: Binaries and Supernovae of type I. ApJ **186**, 1007 (1973).
42. Iben Jr.I. and Tutukov A.V.: Supernovae of type I as end products of the evolution of binaries with components of moderate initial mass (M not greater than about 9 solar masses). ApJS **54**, 335 (1984).
43. Webbink, R.F.: Double white dwarfs as progenitors of R Coronae Borealis stars and Type I supernovae. ApJ **277**, 355 (1984).
44. Perlmutter, S., Aldering, G., Goldhaber, G., et al.: Measurements of Ω and Λ from 42 High-Redshift Supernovae. ApJ **517**, 565 (1999).
45. Riess, A.G., Filippenko, A.V., Challis, P., et al.: Observational evidence from Supernovae for an accelerating universe and a cosmological constant. AJ **116**, 1009 (1998).
46. Filippenko, A.V., Richmond, M.W., Branch, D., et al.: The subluminous, spectroscopically peculiar type IA supernova 1991bg in the elliptical galaxy NGC 4374. AJ **104**, 1543 (1992).
47. Filippenko, A.V., Richmond, M.W., Matheson, T., et al.: The peculiar Type IA SN 1991T - Detonation of a white dwarf? ApJ **384**, L15 (1992).
48. Scalzo R., Aldering, G., Antilogus, P., et al.: A Search for new candidate Super-Chandrasekhar-mass Type Ia Supernovae in the Nearby Supernova Factory Data Set. ApJ **757**, 12 (2012).
49. Cao Y., Johansson, J., Nugent, P. E., et al.: Absence of Fast-moving Iron in an Intermediate Type Ia Supernova between normal and super-chandrasekhar. The Astrophysical Journal **823**, 147 (2016).
50. Miller, A.A., Cao, Y., Piro, A.L., et al.: Early Observations of the Type Ia Supernova iPTF 16abc: A Case of Interaction with Nearby, Unbound Material and/or Strong Ejecta Mixing. ApJ **852**, 100 (2018).
51. Blagorodnova N., Neill, J.D., Kasliwal, M., et al.: Follow-up observations of DLT16am/AT2016ija with SEDM. The Astronomer’s Telegram, 9787 (2016).
52. Tartaglia, L., Sand, D., Valenti, S., et al.: The Early Detection and Follow-up of the Highly Obscured Type II Supernova 2016ija/DLT16am. ApJ **853**, 62 (2018).
53. Gal-Yam, A.: Luminous Supernovae. Science **337**, 927 (2012).
54. Bassett, B., Becker, A., Brewington, H., et al.: SUPERNOVAE 2006kg-2006lc. Central Bureau Electronic Telegrams **688**, 1 (2006).
55. Sako M., Bassett, B., Becker, A., et al.: The Sloan Digital Sky Survey-II Supernova Survey: Search Algorithm and Follow-up Observations. AJ **135**, 348 (2008).
56. Östman L., Nordin, J., Goobar, A., et al.: NTT and NOT spectroscopy of SDSS-II supernovae. A&A **526**, A28 (2011).
57. Wyrzykowski L., Leto, G., Altavilla, G., et al.: Gaia16aye is a binary microlensing event and is crossing the caustic again. The Astronomer’s Telegram, 9507 (2016).
58. SDSS-DR15 Data, http://skyserver.sdss.org/dr15/en/tools/explore/summary.aspx
59. Smee, S.A., Gunn, J.E., Uomoto, A., et al.: The Multi-object, Fiber-fed Spectrographs for the Sloan Digital Sky Survey and the Baryon Oscillation Spectroscopic Survey. AJ **146**, 32 (2013).
60. Pruzhinskaya, M.V., Malanchev, K.L., Kornilov, M.V., et al.: Anomaly Detection in the Open Supernova Catalog. MNRAS, in press (2019).
61. LSST Science Collaboration et al.: LSST Science Book, Version 2.0. arXiv e-prints, arXiv:0912.0201 (2019).

62. van der Walt, S., Colbert, S.C., and Varoquaux, G.: The NumPy Array: A Structure for Efficient Numerical Computation. Computing in Science and Engineering **13**, 22 (2011).
63. Hunter, J.D.: Matplotlib: A 2D Graphics Environment. Computing in Science and Engineering **9**, 90 (2007).
64. Jones, E., Oliphant, T., Peterson, P., et al., SciPy: Open source scientific tools for Python, http://www.scipy.org/ (2001)
65. McKinney, W.: Data Structures for Statistical Computing in Python. In: van der Walt S., Millman, J. (eds.), Proceedings of the 9th Python in Science Conference, 51–56 (2010).
66. Pedregosa, F., Varoquaux, G., Gramfort, A., et al.: Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research **12**, 2825 (2011).

Cross-matching of Objects in Large Sky Surveys

Oleg Malkov[1,2], Sergey Karpov[3,4,5], Dana Kovaleva[1], Jayant Murthy[6], Sergey Sichevsky[1], Nikolay Skvortsov[7], Sergey Stupnikov[7], Gang Zhao[2], and Aleksandr Zhukov[1,8,9]

[1] Institute of Astronomy, Moscow 119017, Russia,
malkov@inasan.ru,
WWW home page: http://www.inasan.ru/~malkov
[2] National Astronomical Observatories, Beijing 100012, China
[3] Institute of Physics, Czech Academy of Sciences, 182 21 Prague 8, Czech Republic
[4] Special Astrophysical Observatory, Nizhnij Arkhyz 36916, Russia
[5] Kazan (Volga region) Federal University, Kazan 420008, Russia
[6] Indian Institute of Astrophysics, Bengaluru 560034, India
[7] Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, Moscow 119333, Russia
[8] Sternberg Astronomical Institute, Moscow 119234, Russia
[9] Russian Technological University (MIREA), Moscow 119454, Russia

Abstract. The study of the stellar physical properties as well as the spatial distribution of interstellar extinction, is important for many investigations of galactic and extragalactic objects. We have developed a method for determination of stellar parameters and interstellar extinctions from multicolor photometry. This method was applied to objects drawn from modern large photometric surveys and, in this work, we give a review of the surveys and discuss problems of cross-identification.

Keywords: Cross-matching · Sky surveys · Photometry · Interstellar extinction

1 Introduction

An outstanding problem of astrophysics is the study of the stellar physical properties. Because the stars are observed through interstellar dust, their light is dimmed and reddened, complicating their parameterization and classification. The parameters of a given star, as well as the interstellar reddening, may be obtained from its spectrum but one must either use a large telescope or only observe bright objects in order to get spectral energy distributions with good resolution and sufficient accuracy. On the other hand, recently constructed large photometric surveys with new tools for cross-matching objects provide us with the possibility of getting multicolor photometric data for hundreds of millions of objects. From these, we may not only parameterize objects but also determine the 3-dimensional interstellar extinction in the Galaxy.

We have developed a method for the determination of stellar parameters and interstellar extinction values from multicolor photometry. The application of this method to a set of stars in a small area in the sky allows us to determine an increase of interstellar extinction with distance in that direction and, consequently, to construct a 3-d extinction map of the Milky Way Galaxy.

Published interstellar extinction maps are described in Section 2. Section 3 contains description of our procedure for parameterization of stars. In Section 4 we give a review of sky surveys, and present principles of their cross-matching. Our results are presented in Section 5. Our conclusions are given in Section 6.

2 Interstellar Extinction Maps

Three-dimensional (3D) extinction models have been constructed using spectral and photometric stellar data, open cluster data, star counts, Galactic dust distribution models.

The standard approach to construct a 3D extinction model has been to parcel out the sky in angular cells, each defined by boundaries in Galactic coordinates (l, b). The visual extinction (A_V) in each cell may then be obtained as a function of distance (d): $A_V(l, b, d)$ from the stars in the cells. The angular size of the cells has varied from study to study, although each cell was generally chosen to be large enough to contain a statistically significant number of calibration stars at different distances.

Published 3D models, using spectral and photometric data, were based on 10^4–10^5 stars, or were constructed for a very limited area in the sky (see, e.g., [30], [14], [17], the earlier studies were reviewed in [24]). Modern large surveys contain photometric (3 to 5 bands) data for 10^7–10^9 stars. However, to make those data (obtained at different wavelengths and with different observational techniques) useful for a 3D extinction model construction, one needs to run a correct cross-identification of objects between surveys. Such cross-identification was laborious and time consuming, but using Virtual Observatory (VO) data access and cross-correlation technologies, a search for counterparts in a subset of different catalogues can now be carried out in a few minutes. It is now feasible to obtain information on interstellar extinction from modern large photometric surveys.

To properly obtain astrophysical parameters from catalogued photometry one needs to study the possibility and sphere of application of the parameterization method. We indicate areas in the parameter space [effective temperature $\log T_{\mathrm{eff}}$, gravity $\log g$, metallicity $[Fe/H]$, visual extinction A_V, total-to-selective extinction ratio R_V], where observational photometry precision, achieved in modern large multi-color surveys, allows us to obtain astrophysical parameters with acceptable accuracy [32].

3 Multicolor Photometry and Parameterization of Stars

We studied a problem of classification and parameterization of stars from multicolor photometry in detail (see, e.g., [33], [34]). In particular, a problem of binary stars parameterization was studied in [26] and [25].

We have developed a method, which allows us to construct $A_V(l, b, d)$ relations from multicolor photometry. Varying (i) the spectral type of the star (SpT), (ii) its distance (d), and (iii) interstellar extinction value (A_V), we simulate the observational brightness, m, with the distance modulus equations

$$m = M_i(\mathrm{SpT}) + 5\log d - 5 + A_i(A_V) \quad (1)$$

for every photometric band, and, based on the quality of the simulation process, choose the most appropriate SpT-d-A_V set. A calibration relation M_i(SpT) and interstellar extinction law $A_i(A_V)$ should be available for each of the i photometric bands included in the original surveys.

We have to remove all non-stellar objects, unresolved photometric binaries, variable stars and other contaminating objects, based on flags included in the original surveys with flags from our simulation techniques.

This method of simulation/parameterization, as described above, allows one to plot parameterized objects in the distance-extinction (d-A_V) plane, approximate them (by the cosecant law or more complicated function) and estimate interstellar extinction parameters in a given direction on the sky.

Note that for high galactic latitude areas ($|b| > 15^o$ or so) the interstellar extinction is thought to be (roughly) uniformly distributed and to satisfy the so-called cosecant (barometric) law, suggested by Parenago in [28]. That function should be modified (complicated) for lower latitudes, as dust clouds concentrated in the Galactic plane, will have to be taken into account.

4 Sky Surveys and Cross-matching

4.1 Sky Surveys Selection

The following sky surveys are selected for our study:

- The DENIS database [10];
- 2MASS All-Sky Catalog of Point Sources [8];
- The SDSS Photometric Catalogue, Release 12 [1];
- GALEX-DR5 (GR5) sources from AIS and MIS [4], [5];
- UKIDSS-DR9 LAS, GCS and DXS Surveys [18];
- AllWISE Data Release [9];
- IPHAS DR2 Source Catalogue [3];
- The Pan-STARRS release 1 (PS1) Survey – DR1 [7];
- Gaia DR2 [13], [2].

The selected surveys satisfy the following criteria:

- the number of objects exceeds 10×10^6;
- the survey covers a large area in the sky (the only exception is IPHAS, which covers a relatively small but important area in the sky);
- the photometric accuracy is better than about 0.05 mag;
- the depth of the survey exceeds V$\sim$20 mag.

For every survey the following information should be available: absolute magnitude – spectral type (M_λ – SpT) calibration tables and $A_\lambda(A_V)$ relations for every photometric band λ. If these information is not available in literature, we construct it using response curves of photometric bands and spectral energy distribution (SED) for every spectral type, as well as the interstellar extinction law ([12], [6], [11]). Besides, relations between spectral type and atmospheric parameters (effective temperature $\log T_{\rm eff}$ and surface gravity $\log g$) for stars of different luminosity classes should be available.

To model observational photometry one needs to know spectral energy distribution, and a number of spectrophotometric atlases are designed to meet that requirement (e.g, [29], [36]). We have made a comparative analysis of the most known semi-empirical and empirical spectral atlases. The results show that the standard error of synthesized stellar magnitudes calculated with SEDs from best spectral atlases reaches 0.02 mag. It has been also found that some modern spectral atlases are burdened with significant systematic errors [16].

A preliminary analysis of applicability of SDSS and 2MASS photometry for determining the properties of stars and interstellar extinction was made by in [31].

4.2 Cross-matching of Surveys

The number of surveys available at any wavelength is large enough to construct detailed Spectral Energy Distributions (SEDs) for any kind of astrophysical object. However, different surveys/instruments have different positional accuracy and resolution. In addition, the depth of each survey is different and, depending on sources brightness and their SED, a given source might or might not be detected at a certain wavelength. All this makes the pairing of sources among catalogues not trivial, especially in crowded fields.

We have implemented an algorithm of fast positional matching of large astronomical catalogs in small (up to one degree) areas with filtering of false identification [21]. In particular, for each area and each pair we estimated the matching radius. As a result, we drew in a number 0.1-degree radius areas samples of point-like objects counterparts from the DENIS, 2MASS, SDSS, GALEX, and UKIDSS surveys, and performed a cross-identification within these surveys [15], [20]. We have compiled the corresponding subcatalogues in the VOTable [27] format. The tool developed as a result of this work can be used to cross-identify objects in arbitrary sky areas for the further classification and determination of stellar parameters, including the measurement of the amount of interstellar extinction.

In some surveys (e.g., GALEX, SDSS, UKIDSS) more than one observation per object was made and, consequently, more than one entry per object is present in the catalogue. In such cases we use weighted average values for the photometry.

In the cross-identification process (and later for the parameterization) we use all positional information and all photometry available in surveys. To select objects for further study we also pay attention to various flags, presented in the surveys. The flags can indicate quality of observations and provide information on a nature of object (duplicity, variability, extended shape). As it was mentioned above, on this stage we do not use trigonometric parallax as an input parameter.

5 Results and Discussion

In our pilot study [22] we applied this method to construct $A_V(l, b, d)$ relations for selected areas at high galactic latitudes. We have cross-matched objects in 2MASS, SDSS, GALEX and UKIDSS surveys in selected areas in the sky, using Virtual Observatory facilities. As a result of the cross-matching, we find multi-wavelength ($i = 9$ to 13 bands) photometric data for each object.

We have compared our results with LAMOST [19] data and extinction values to distant SNs (based on IRAS and DIRBE microwave data), available in the literature. The comparison exhibits a good agreement (see [22] for details). A comparison of our results with recently released Gaia DR2 data also demonstrates a good agreement for stars as faint as $19^m.6$ $\mathrm{g_{SDSS}}$, and shows that our method allows us to determine spectral type, distance and interstellar extinction of objects out to 4.5 kpc [23]. It indicates that the proposed algorithm (after some modifications, required for low galactic latitudes) can be used for construction of a 3D map of interstellar extinction in the Milky Way Galaxy.

6 Conclusion

The parameterization of stars is a well known problem and used for various purposes in astronomy (e.g., while solving the problem of searching for well defined stars to be used for secondary photometric standards [35]). We have shown that multicolor photometric data from large modern surveys can be used for parameterization of stars. A comparison of our results with independent data shows a good agreement. We prove that with sufficiently good quality photometry, one may compute a 3D extinction map by comparing catalogued multicolor photometry with photometry derived from the secondary estimators such as the distance modulus and the interstellar extinction law with suitable calibration tables for absolute magnitudes with reasonable spectral types, extinctions and distances.

With the advent of large, existing and coming, photometric surveys and the evolution of computing power and data analysis techniques (in particular, Virtual Observatory tools for cross-matching), interstellar extinction can now be computed for hundreds of millions of stars in a reasonable amount of time, and a 3D interstellar extinction map can be constructed.

Acknowledgments

We thank our reviewers whose comments greatly helped us to improve the paper. OM thanks the CAS President's International Fellowship Initiative (PIFI). The work was partly supported by the Russian Foundation for Basic Researches (project 17-52-45076) and by DST grant INT/RUS/RFBR/P-265 to JM. This research has made use of NASA's Astrophysics Data System, and use of the VizieR catalogue access tool, CDS, Strasbourg, France.

References

1. Alam, S., Albareti, F.D., Allende Prieto, C., Anders, F., Anderson, S.F., Anderton, T., Andrews, B.H., Armengaud, E., Aubourg, É., Bailey, S., et al.: The Eleventh and Twelfth Data Releases of the Sloan Digital Sky Survey: Final Data from SDSS-III. Astrophys. J. Suppl. Ser. **219**, 12 (2015). https://doi.org/10.1088/0067-0049/219/1/12
2. Bailer-Jones, C.A.L., Rybizki, J., Fouesneau, M., Mantelet, G., Andrae, R.: Estimating Distance from Parallaxes. IV. Distances to 1.33 Billion Stars in Gaia Data Release 2. Astron. J. **156**, 58 (2018). https://doi.org/10.3847/1538-3881/aacb21
3. Barentsen, G., et al: VizieR Online Data Catalog: IPHAS DR2 Source Catalogue (Barentsen+, 2014). VizieR Online Data Catalog **2321** (2014).
4. Bianchi, L., Herald, J., Efremova, B., Girardi, L., Zabot, A., Marigo, P., Conti, A., and Shiao, B.: GALEX catalogs of UV sources: statistical properties and sample science applications: hot white dwarfs in the Milky Way. Astrophys. Space. Sci. **335**, 161–169 (2011). https://doi.org/10.1007/s10509-010-0581-x
5. Bianchi, L., Shiao, B., and Thilker, D.: Revised Catalog of GALEX Ultraviolet Sources. I. The All-Sky Survey: GUVcat_AIS. Astrophys. J. Suppl. Ser. **230**, 24 (2017). https://doi.org/10.3847/1538-4365/aa7053
6. Cardelli, J.A., Clayton, G.C., and Mathis, J.S.: The relationship between infrared, optical, and ultraviolet extinction. Astrophys. J. **345**, 245–256 (1989). https://doi.org/10.1086/167900
7. Chambers, K.C., et al: The Pan-STARRS1 Surveys. arXiv e-prints (2016)
8. Cutri, R.M., et al: VizieR Online Data Catalog: 2MASS All-Sky Catalog of Point Sources (Cutri+ 2003). VizieR Online Data Catalog **2246** (2003)
9. Cutri, R.M., et al: VizieR Online Data Catalog: AllWISE Data Release (Cutri+ 2013). VizieR Online Data Catalog **2328** (2014).
10. DENIS Consortium: VizieR Online Data Catalog: The DENIS database (DENIS Consortium, 2005). VizieR Online Data Catalog **1** (2005).
11. Fitzpatrick, E.L. and Massa, D.: An Analysis of the Shapes of Interstellar Extinction Curves. V. The IR-through-UV Curve Morphology. Astrophys. J. **663**, 320–341 (2007). https://doi.org/10.1086/518158
12. Fluks, M.A., Plez, B., The, P.S., de Winter, D., Westerlund, B.E., and Steenman, H.C.: On the spectra and photometry of M-giant stars. Astron. and Astrophys. Suppl. Ser. **105**, 311–336 (1994).
13. Gaia Collaboration, Brown, A.G.A., Vallenari, A., Prusti, T., de Bruijne, J.H.J., Babusiaux, C., Bailer-Jones, C.A.L., Biermann, M., Evans, D.W., Eyer, L., et al.: Gaia Data Release 2. Summary of the contents and survey properties. Astron. Astrophys. **616**, A1 (2018). https://doi.org/10.1051/0004-6361/201833051
14. Green, G.M., et al: A Three-dimensional Map of Milky Way Dust. Astrophys. J. **810**, 25 (2015). https://doi.org/10.1088/0004-637X/810/1/25
15. Karpov, S.V., Malkov, O.Y., and Mironov, A.V.: Cross-identification of large surveys for finding interstellar extinction. Astrophysical Bulletin **67**, 82–89 (2012). https://doi.org/10.1134/S1990341312010087
16. Kilpio, E.Y., Malkov, O.Y., and Mironov, A.V.: Comparative analysis of modern empirical spectro-photometric atlases with multicolor photometric catalogues. In: P. Prugniel, H.P. Singh (eds.) Astronomical Society of India Conference Series, *Astronomical Society of India Conference Series*, vol. 6, p. 31 (2012).
17. Lallement, R., Capitanio, L., Ruiz-Dern, L., Danielski, C., Babusiaux, C., Vergely, L., Elyajouri, M., Arenou, F., and Leclerc, N.: Three-dimensional maps of

interstellar dust in the Local Arm: using Gaia, 2MASS, and APOGEE-DR14. Astron. Astrophys. **616**, A132 (2018). https://doi.org/10.1051/0004-6361/201832832

18. Lawrence, A., et al: The UKIRT Infrared Deep Sky Survey (UKIDSS). Mon. Not. R. Astron. Soc. **379**, 1599–1617 (2007). https://doi.org/10.1111/j.1365-2966.2007.12040.x
19. Luo, A.L., et al: The first data release (DR1) of the LAMOST regular survey. Research in Astronomy and Astrophysics **15**, 1095 (2015). https://doi.org/10.1088/1674-4527/15/8/002
20. Malkov, O., Dluzhnevskaya, O., Karpov, S., Kilpio, E., Kniazev, A., Mironov, A., and Sichevskij, S.: Cross Catalogue Matching with Virtual Observatory and Parametrization of Stars. Baltic Astronomy **21**, 319–330 (2012). https://doi.org/10.1515/astro-2017-0390
21. Malkov, O. and Karpov, S.: Cross-Matching Large Photometric Catalogs for Parameterization of Single and Binary Stars. In: I.N. Evans, A. Accomazzi, D.J. Mink, A.H. Rots (eds.) Astronomical Data Analysis Software and Systems XX, *Astronomical Society of the Pacific Conference Series*, vol. 442, p. 583 (2011).
22. Malkov, O., Karpov, S., Kilpio, E., Sichevsky, S., Chulkov, D., Dluzhnevskaya, O., Kovaleva, D., Kniazev, A., Mickaelian, A., Mironov, A., Murthy, J., Sytov, A., Zhao, G., and Zhukov, A.: Interstellar extinction from photometric surveys: application to four high-latitude areas. Open Astronomy **27**, 62–69 (2018). https://doi.org/10.1515/astro-2018-0002
23. Malkov, O., Karpov, S., Kovaleva, D., Sichevsky, S., Chulkov, D., Dluzhnevskaya, O., Kniazev, A., Mickaelian, A., Mironov, A., Murthy, J., Sytov, A., Zhao, G., and Zhukov, A.: Verification of Photometric Parallaxes with Gaia DR2 Data. Galaxies **7**, 7 (2018). https://doi.org/10.3390/galaxies7010007
24. Malkov, O. and Kilpio, E.: A Synthetic Map of the Galactic Interstellar Extinction. Astrophys. Space. Sci. **280**, 115–118 (2002). https://doi.org/10.1023/A:1015526811574
25. Malkov, O., Mironov, A., and Sichevskij, S.: Single-binary star separation by ultraviolet color index diagrams. Astrophys. Space. Sci. **335**, 105–111 (2011). https://doi.org/10.1007/s10509-011-0613-1
26. Malkov, O.Y., Sichevskij, S.G., and Kovaleva, D.A.: Parametrization of single and binary stars. Mon. Not. R. Astron. Soc. **401**, 695–704 (2010). https://doi.org/10.1111/j.1365-2966.2009.15696.x
27. Ochsenbein, F., Williams, R., Davenhall, C., Durand, D., Fernique, P., Hanisch, R., Giaretta, D., McGlynn, T., Szalay, A., and Wicenec, A.: VOTable: Tabular Data for the Virtual Observatory. In: P.J. Quinn, K.M. Górski (eds.) Toward an International Virtual Observatory, p. 118 (2004). https://doi.org/10.1007/10857598_18
28. Parenago, P.P.: On interstellar extinction of light. Astron. Zh. **13**, 3 (1940).
29. Pickles, A.J.: A Stellar Spectral Flux Library: 1150-25000 Å. Publ. Astron. Soc. Pac. **110**, 863–878 (1998). https://doi.org/10.1086/316197
30. Sale, S.E., Drew, J.E., Barentsen, G., Farnhill, H.J., Raddi, R., Barlow, M.J., Eislöffel, J., Vink, J.S., Rodríguez-Gil, P., and Wright, N.J.: A 3D extinction map of the northern Galactic plane based on IPHAS photometry. Mon. Not. R. Astron. Soc. **443**, 2907–2922 (2014). https://doi.org/10.1093/mnras/stu1090
31. Sichevskij, S.G.: Applicability of Broad-Band Photometry for Determining the Properties of Stars and Interstellar Extinction. Astrophysical Bulletin **73**, 98–107 (2018). https://doi.org/10.1134/S199034131801008X
32. Sichevskij, S.G., Mironov, A.V., and Malkov, O.Y.: Accuracy of stellar parameters determined from multicolor photometry. Astrophysical Bulletin **69**, 160–168 (2014). https://doi.org/10.1134/S1990341314020035

33. Sichevskiy, S.G., Mironov, A.V., and Malkov, O.Y.: Classification of stars with WBVR photometry. Astronomische Nachrichten **334**, 832 (2013). https://doi.org/10.1002/asna.201311932
34. Sichevsky, S. and Malkov, O.: Estimating stellar parameters and interstellar extinction from evolutionary tracks. Baltic Astronomy **25**, 67–74 (2016). https://doi.org/10.1515/astro-2017-0112
35. Skvortsov, N.A., Avvakumova, E.A., Bryukhov, D.O., Vovchenko, A.E., Vol'nova, A.A., Dluzhnevskaya, O.B., Kaigorodov, P.V., Kalinichenko, L.A., Kniazev, A.Y., Kovaleva, D.A., Malkov, O.Y., Pozanenko, A.S., and Stupnikov, S.A.: Conceptual approach to astronomical problems. Astrophysical Bulletin **71**(1), 114–124 (2016). https://doi.org/10.1134/S1990341316010120
36. Wu, Y., Singh, H.P., Prugniel, P., Gupta, R., and Koleva, M.: Coudé-feed stellar spectral library – atmospheric parameters. Astron. Astrophys. **525**, A71 (2011). https://doi.org/10.1051/0004-6361/201015014

Search and Observations of Optical Counterparts for Events Registered by LIGO/Virgo Gravitational Wave Detectors

Elena Mazaeva[1], Alexei Pozanenko[1], Alina Volnova[1], Pavel Minaev[1], Sergey Belkin[1], Raguli Inasaridze[2], Evgeny Klunko[3], Anatoly Kusakin[4], Inna Reva[4], Vasily Rumyantsev[5], Artem Novichonok[6,7], Alexander Moskvitin[7], Gurgen Paronyan[8], Sergey Schmalz[6], and Namkhai Tungalag[10]

[1]Space Research Institute (IKI), 84/32 Profsoyuznaya Str, Moscow, Russia, 117997
[2]KharadzeAbastumani Astrophysical Observatory, Ilia State University, Tbilisi, 0162, Georgia
[3]Institute of Solar Terrestrial Physics, Irkutsk, Russia, 664033
[4]Fesenkov Astrophysical Institute, Almaty, 050020, Kazakhstan
[5] Crimean Astrophysical Observatory, Nauchny, Crimea 298409
[6]Keldysh Institute of Applied Mathematics, Russian Academy of Sciences, Miusskaya 4, Moscow, Russia, 125047
[7]Petrozavodsk State University, Leninast., 33, Petrozavodsk, Russia, 185910
[8] Special Astrophysical Observatory of Russian Academy of Sciences, Nizhniy Arkhyz, Russia, 369167
[9]Byurakan Astrophysical Observatory, 0213, Byurakan, Aragatzotn Province, Republic of Armenia
[10]Institute of Astronomy and Geophysics, Mongolian Academy of Sciences, 13343, Ulaanbaatar, Mongolia
elena.mazaeva@phystech.edu

Abstract. The problem of search for optical counterpart of LIGO/Virgo events are discussing. Multi-messenger astronomy boosts the use a huge amount of astronomical data obtained by virtually all observatories around the world. We are discussing different methods used for observations, problem of search for transients in the extremely large localization error-box of LIGO/Virgo events, and lessons obtained during second observational run of LIGO/Virgo in 2017. In particular we present our experience and results of follow up observations of LIGO/Virgo optical counterpart candidates.

Keywords: Multi-messenger astronomy, gravitational waves, LIGO/Virgo, gamma-ray bursts, afterglow, kilonova, photometry.

1 Introduction

The problem of search and observations of new transient objects is one of the main problems in modern astrophysics. It requires wide-field observations with some initial all-sky catalogue of stationary sources for comparison. Dedicated surveys and experiments produce huge amounts of data daily in every domain of electromagnetic spectrum: in high energy range [1], optics [2, 3], and radio [4], as well as in cosmic particles window [5] and gravitational waves [6]. The surveys of new generation, like Large Synoptic Survey Telescope [7], will produce data of unprecedented volume and complexity. Reduction and analysis of these enormous data sets is already out of human's capacity and is similar to a search of a needle in a haystack. This problem is also connected to the search of transients related to the gravitational waves detections in very large localization areas, provided by LIGO and Virgo observations during theirs third scientific observational run in 2019.

The Laser Interferometer Gravitational-Wave Observatory (LIGO) is designed to open the field of gravitational-wave astrophysics through the direct detection of gravitational waves predicted by Einstein's General Theory of Relativity [8]. LIGO's multi-kilometer-scale gravitational wave detectors use laser interferometry to measure the minute ripples in space-time caused by passing gravitational waves from cataclysmic cosmic events such as merging neutron stars (NSs) or black holes (BHs), or by supernovae. LIGO consists of two widely separated interferometers within the United States – one in Hanford, Washington and the other in Livingston, Louisiana – operated in unison to detect gravitational waves.

The first success of LIGO observations came in 2015 with the first direct observations of gravitational waves from the binary black hole merging GW150914 [9]. In 2017, when the sensitivity of LIGO detectors increased, and Virgo detector in Italy started its first observational cycle [10], the merging of binary neutron star was detected for the first time [11].

In the context of gravitational waves detection, the most important problem for astrophysics is the search, identification, and observations of the possible electromagnetic (EM) counterpart of the event. The General Relativity predicts no EM radiation from the binary BH coalescence since, in theory, there is no enough matter outer the source that can produce it. In practice, there may be some radiation caused by accretion of a circumstellar matter on the resulting black hole, but its predicted flux is extremely low (e.g. [12]). Quite different situation is the binary NS merging (BNS). In this case, the merging objects consist of an ordinary matter that may produce high-energy EM radiation process (short gamma-ray burst) after the merging BNS an afterglow of wide energy range, and most interesting BSN counterpart which is called 'kilonova'.

The association between BNS merging, short gamma-ray bursts and kilonovae was first predicted theoretically [13], and then was confirmed observationally with the detection of GW170817/ GRB 170817A / AT2017gfo [14]. Besides the fact that GW170817 was the first case of the registration of gravitational waves from a BNS merging, it was also the first detection of gravitational waves and EM radiation from the same source [15].

A signal from the binary system merging is modeled numerically based on the Einstein's General Theory of Relativity and represents a package of oscillations increasing in amplitude with a decreasing period. The processing algorithm of LIGO and Virgo detectors searches for modeled templates in the received data using wavelet analysis. Localization on the sky is performed by triangulation method, measuring the time lag between the detection time for spatially distributed detectors, which determines the sky area of the most probable localization of the source. The time of the signal registration is measured with a high accuracy; however, the localization area may be very large, tens to hundreds of square degrees (see Table 1). GW170817 [16] has a localization region of ~30 square degrees, and there were reported ~190 galaxies in the volume limited by the sky area and distance estimates [17]. The kilonova AT2017gfo was discovered independently by 6 survey projects and was observed during several dozens of days in wide energy range from X-rays to radio [18]. The co-authors of the paper were used the mosaic method to search for optical counterpart, observed of kilonova and proposed the model of prompt emission of GW170817/GRB 170817A [19].

In this paper, we discuss the problem of the search of a new transient optical source in large areas provided by detections of gravitational wave sources. We describe two basic methods of the search: mosaic observations of localization area and pre-determined goals observations, i.e. search for transients in galaxies inside the detection volume. We also provide several examples of such semi-manual searches using available ground-based optical telescopes performed during the LIGO/Virgo observational run O2. Multi-messenger Astronomy is becoming a commonplace [20].

Table 1. Selected source parameters of the eleven confident GW detections [21]

Event	Type [a]	d_L/Mpc[b]	$\Delta\Omega$=deg^2[c]
GW150914	BBH	430 (+150/-170)	180
GW151012	BBH	1060 (+540/-480)	1555
GW151226	BBH	440 (+180/-190)	1033
GW170104	BBH	960 (+430/-410)	924
GW170608	BBH	320 (+120/-110)	396
GW170729	BBH	2750 (+1350/-1320)	1033
GW170809	BBH	990 (+320/-380)	340
GW170814	BBH	580 (+160/-210)	87
GW170817	BNS	40 (+10/-10)	16
GW170818	BBH	1020 (+430/-360)	39
GW170823	BBH	1850 (+840/-840)	1651

[a] BBH – binary black holes. BNS – binary neutron stars.

[b] Luminosity distance.

[c] Error box of sky localization.

2 The Optical Transient Search Procedure

After receiving the alert signal from LIGO/Virgo, our observations are carried out on ground-based optical telescopes to search for counterpart.

One can observe the whole range of localization with wide-field telescopes. This observation tactic is suitable if the localization area is not very large (up to about one hundred square degrees), or it is possible to observe on a large number of telescopes.

Since we know not only the localization region in the celestial sphere of the gravitational-wave event, but also the distance to the source, we can only observe galaxies from the localization region that are located at a given distance. For this purpose, there is a value-added full-sky catalogue of galaxies, named as Galaxy List for the Advanced Detector Era, or GLADE [22]. GLADE was constructed by cross-matching and combining data from five separate (but not independent) astronomical catalogues: GWGC, 2MPZ, 2MASS XSC, HyperLEDA, and SDSS-DR12Q. But GLADE is complete up only to d_L=37(+3/-4) Mpc in terms of the cumulative B-band luminosity of galaxies within luminosity distance d_L, and contains all of the brightest galaxies giving half of the total B-band luminosity up to d_L=91 Mpc. While the distance to the registered source can be several thousand Mpc (see Table 1).

But whatever method we use, we need to find a transient on the obtained optical images.

We use the method of comparison with all-*sky catalogs* using the generated catalog of sources selected from the image. A block diagram of the method is presented in Fig. 2.

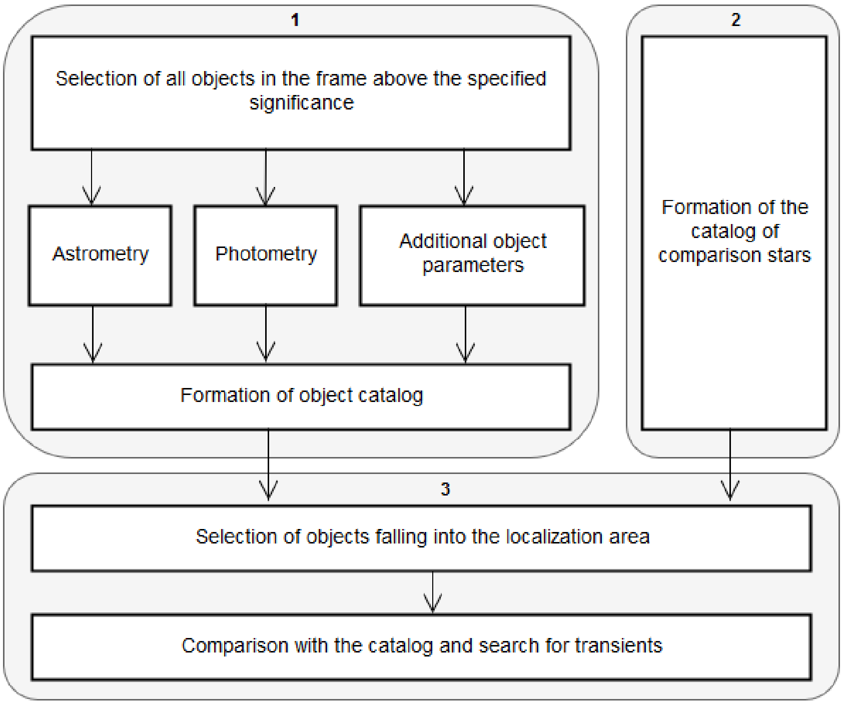

1
Selection of all objects in the frame above the specified significance
Astrometry
Photometry
Additional object parameters
Formation of object catalog
2
Formation of the catalog of comparison stars
3
Selection of objects falling into the localization area
Comparison with the catalog and search for transients

Fig. 1. A block diagram of the match algorithm

Block 1 – detection, measure and classification of sources from astronomical images, the formation of an object catalog. In this case, we used SExtractor – software for source extraction [23]. Comparison of object catalogs is best done using equatorial coordinates of objects; first of all, astrometry is necessary (for example, using Apex [24] or other software). To avoid incorrect comparison, it is useful to reject objects at the border of a frame that do not fully fit into the frame (either at a distance of < 4 FWHM from the border or to use the value of the SExtractor flags).

Block 2 – formation of the catalog of the comparison stars. For comparison of objects it is better to use photometric catalogs (e.g. SDSS, Pan-STARRS, APASS, 2MASS, it depends on the filter of the original image, the image upper limit and the region of the celestial sphere).

Block 3 – search for transient sources.Comparison of objects is performed simultaneously by equatorial coordinates and magnitude within the measurement error.

3 Results Obtained for our Procedures of Search and Identification

Our collaboration is based at the Space Research Institute and provided follow-up gravitational wave observations in the optical range during Second Observing Run of LIGO/Virgo.

The optical data were obtained by IKI GRB Follow-up Network which is collaborating with Crimean Astrophysical Observatory (CrAO), Sayan Solar Observatory (Mondy), Tian Shan Astrophysical Observatory (TShAO), Abastumani Astrophysical Observatory (AbAO), Special Astrophysical Observatory (SAO), ISON-Khureltogoot, Koshka observatory of INASAN and Byurakan Astrophysical Observatory (BAO).

3.1 LIGO/Virgo G299232: Compact Binary Coalescence Candidate

GW170825 G299232is a low-significance compact binary coalescence candidate identified from LIGO Hanford Observatory (H1) and LIGO Livingston Observatory (L1) at 2017-08-25 13:13:31 UTC. If the candidate is astrophysical in origin, it appears consistent with the merger of a black hole and a neutron star [25]. Subsequently, the event was not confirmed.

Localization generated by the BAYESTAR pipeline [26] including information from H1, L1, and V1 is presented in Fig. 1. The 90% credible region spans about 2040 deg^2. The a posteriori luminosity distance estimate is 339 +/-110 Mpc [25].

The IceCube Neutrino Observatory (a cubic-kilometer neutrino detector operating at the geographic South Pole, Antarctica) searched IceCube online track-like neutrino candidates (GFU) detected in a [-500,500] second interval about the LIGO/Virgo trigger G299232 [27]. Comparison of the candidate source directions of 7 temporally-coincident neutrinos to the BAYESTAR skymap is presented in Fig. 2.

One of the neutrino candidate (marked as X1) was within the LIGO/Virgo localization area and detected 233.82 seconds before LIGO/Virgo trigger G299232.X1 sky location is R.A.=28.2, Dec.=44.8 with 3.8 degrees uncertainty of direction reconstruction [28].

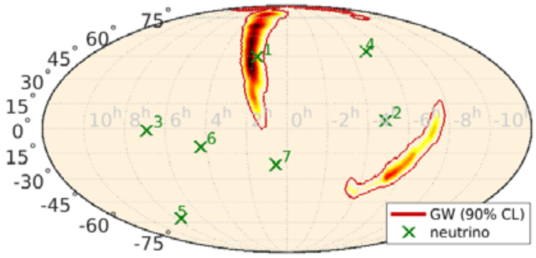


Fig. 2. The localization with distance information generated by the BAYESTAR pipeline [26] including information from H1, L1, and V1.X1 - X7 are neutrino candidates (GFU) detected in a [–500,500] second interval about the LIGO-Virgo trigger G299232

We observed the field of LIGO/Virgo trigger G299232 [25] and error circle of IceCube candidate X1 [27, 28] with wide field of view VT-78a telescope of ISON-Khureltogoot observatory. We obtained several unfiltered images with the two time series starting on 2017-08-25 (UT) 15:24:13 and 16:32:52 (time since LVC trigger are 0.11289 and 0.16054 days), each centered to the position of localization reported in [27] and [28], respectively. Total coverage of the error region of IceCube candidate X1 [28] is 85.7 %. The map of the coverage can be found in Fig. 3.

Using the algorithm described in Chapter 3 we have distinguished 94.7 thousand objects from the images (field of view is 7 x 7 degrees). After comparing these 94.7 thousand objects with the USNO-B.1 catalog we have 834 candidates left, of which 818 are processing artifacts. Finally, we found one cataloged asteroid (895) Helio and 24 objects, the magnitude of which was brighter than R2 of USNO-B1.0, but weaker than R1 (see Table 1). There is no presented R-magnitude for the object 1352-0033439 in USNO-B1.0 catalog, but magnitudes B1=18.27, B2=15.32 and I=13.87 for the object 1352-0033439 are presented in catalog and correspond to our photometric magnitude (column name is “Mag SExtractor” in Table 2).

We found no significant variability of the sources between the two epochs. We found no significant brighter sources, which could be galaxies, than their R-magnitudes presented in the USNO-B.1 catalog. Upper limit on the stellar magnitude of possible optical candidate is 19.2.

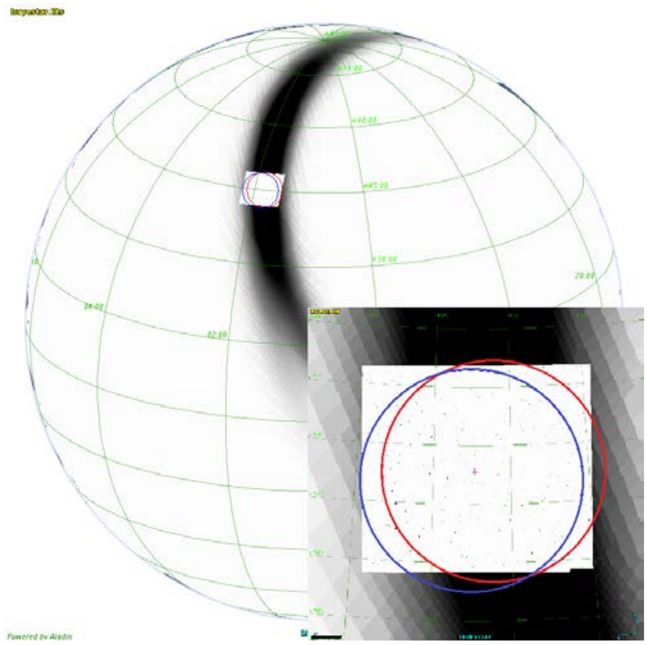


Fig. 3. The map of the coverage IceCube candidate X1 localization by VT-78a telescope of Khureltogoot observatory. Red circle is preliminary IceCube X1 error box [27], blue circle is final error box [28]

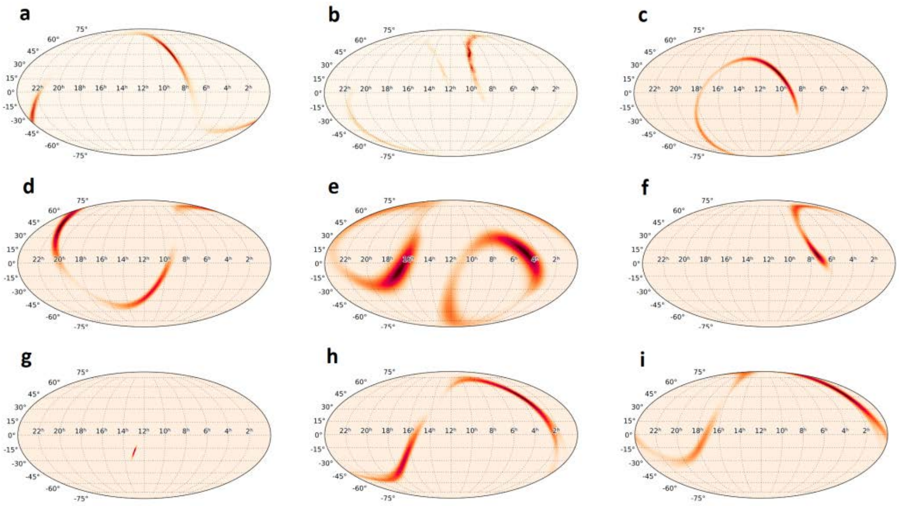


Fig. 4. Sky localization of LIGO/Virgo events. a – GW170104_G268556, b – GW170120_G270580, c – GW170217_G274296, d – GW170227_G275697, e – GW170313_G277583, f – GW170608_G288732, g – GW170817_G298048, h – GW170823_G298936, i – GW170825_G299232

Table 2. List of object which magnitudes are brighter than R2 of USNO-B1.0

USNO-B1.0 id	R1	R2	Mag SExtractor
1377-0046508	13.51	16.26	14.22
1377-0046571	15.49	18.48	16.41
1377-0046983	14.23	20.16	14.80
1378-0048129	11.44	14.42	12.40
1378-0048633	15.51	18.38	16.20
1378-0048730	13.89	16.87	14.49
1377-0048562	15.79	19.76	16.48
1377-0048684	13.69	16.37	14.34
1376-0047873	14.53	18.27	15.06
1372-0048226	15.56	19.83	16.30
1355-0039997	14.13	17.78	14.92
1352-0033439	-	-	16.46
1327-0048518	13.94	18.19	14.11
1328-0048333	14.47	17.22	14.87
1327-0048405	15.19	20.70	15.79
1328-0048274	13.14	15.59	13.31
1324-0046318	19.27	-	15.76
1327-0038107	12.93	19.03	13.68
1327-0038122	15.42	20.47	16.24
1328-0038377	11.94	15.03	12.66
1328-0038222	12.87	18.22	13.66
1328-0038148	12.68	19.20	13.61
1328-0038142	14.50	20.47	15.29
1328-0038084	14.11	18.98	14.92

3.2 Observations of LIGO/Virgo Optical Candidates

In addition to searching the object in the localization area, we also observed objects in the localization area of GW events that were found by other research groups.

The objects that we have observed are listed in the Table 3, the areas of localization of each gravitational-wave event can be seen in the Fig. 4. Some gravitational-wave events, the areas of which we observed, later were not officially confirmed and continue remained candidates.

Table 3. Observations of optical candidates of LIGO/Virgo events

Event	Type [a]	Optical candidates	Type[b]
GW170104_G268556	CBC (+) [29]	PS17fn	n/c
		PS17fl	n/c
		PS17dp	n/c
		PS17gl	n/c
GW170120_G270580	n/c (-) [30]	PS17yt	SN Ia
		MASTER OT J090737.22+611200.5	n/c
		PS17lk	n/c
		PS17nv	n/c
		PS17pv	n/c
		PS17qk	n/c
		PS17rc	n/c
GW170217_G274296	n/c (-) [31]	PS17bek	SLSN
GW170227_G275697	CBC (-) [32]	iPTF17bue	SN Ia
		XRT23	n/c
GW170313_G277583	n/c (-) [33]	ATLAS17cgg	SN IIn
GW170608_G288732	BBH (+) [21]	GW170608X2	n/c
GW170817_G298048	BNS (+) [21]	GW 170817	GRB, KN
GW170823_G298936	BBH (+)[21]	GWFUNC-17ure	SN Ia
GW170825_G299232	NS+BH (-) [34]	SwiftJ014008.5+343403.6	n/c
		MASTER OT	SN IIb

[a] BBH – binary black holes merging, BNS – binary neutron stars merging, NS+BH – neutron star and black hole merging, CBC – compact binary coalescence, n/c – this event candidate does *not* have a chirp signature, and thus does not suggest a compact binary merger or the morphology of the event candidate is unclear. (+) – event, (-) – candidate.

[b] SN – supernova, KN – kilonova, SLSN – super-luminous supernova, GRB – gamma-ray burst, n/c – non classified.

GW170120_G270580. The Pan-STARRS covered northern area of the GW170120_G270580 localization and detected 124 transients including rapidly rising transient – PS17yt (R.A. 10:03:57.96 Dec. +49:02:28.3) [35,36]. Our collaboration observed PS17yt source in BVR filters and a light curve of PS17yt were constructed (see Fig. 5a). It was subsequently shown that PS17yt is Ia type supernova at a redshift z ~ 0.026 [37].

Furthermore, we observed orphan sources PS17lk (R.A. 09:29:58.27 Dec. +15:11:58.5), PS17nv (R.A. 09:57:41.01 Dec. +17:49:33.4), PS17qk (R.A. 09:29:12.15 Dec. +25:49:06.4), PS17pv (R.A. 09:25:07.35 Dec. +50:12:28.9), PS17rc (R.A. 09:32:19.16 Dec. +47:03:38.3) and MASTER J090737.22+611200.5 (R.A. 09:07:37.22 Dec. +61:12:00.5) in the field of the LIGO G270580 localizations. Results of observations see in Table 4.

GW170217_G274296. Pan-STARRS covered 501 square degrees on the first night following the release of the G274296 alert. They have located and vetted 10 transients with host spectroscopic redshifts and 60 unknown transients with no host spectroscopic redshifts. [38] We observed one of the transients with no host spectroscopic redshifts (PS17bek) and the light curves in BR-filters are presented in Fig. 5b.

Afterwards a good correlation betweenPS17bek spectrum and the spectra of superluminous supernovae (SLSNe type I) was found. In particular, a good match with the spectra of SN 2010gx at -5 days before peak if PS17bek is at a redshift of z~0.31 was found. The weak emission line at 6559.4 is consistent with [O III] 5007 at z=0.31, and we also detect [O III] 4959 at a consistent redshift but lower significance [39].

GW170825_G299232. Global MASTER robotic net discovered optical transient source – MASTER OT J033744.97+723159.0(R.A. 03:37:44.97 Dec. +72:31:59.0). [40]. Analysis of the MASTER spectrum suggests that it is a supernova Type IIb [41] MASTER OT observation with the RoboPolpolarimeter shown that the R-band fractional polarization of the source is 1.8+/–0.47% [42].

Our observations of the MASTER OT are shown in the Fig. 5c.

4 Summary

In 2017, coordinated hardworking of thousands of astronomers and other scientists around the world allowed to find and successfully observe the electromagnetic counterpart of the gravitational wave event GW170817 of binary neutron star merging. The associated GRB 170817A and kilonova AT2017gfo were observed by hundreds of space and ground-based experiments in all ranges of electromagnetic spectrum. The unprecedented collaboration allowed to obtain detailed properties of kilonova and to verify existing physical models of this phenomenon, which is not fully studied yet. At the same time, there was no any reliable EM counterpart candidate detected for 10 binary black holes coalescences discovered during O1/O2 scientific runs of LIGO and Virgo detectors. However, a huge amount of observational data, which covered vast localization area of the events, led to the discovery of many other new transient sources unrelated to the GW. The problem of search of a new optical transient with specific properties in large localization areas arose here with the great actuality.

Table 4. The photometric observation results of orphan sources

Orphan	Date	Filter	MJD	Magnitude
PS17lk	2017-01-25	R	57778.72969	20.92 +/– 0.16
	2017-01-29	R	57782.74483	> 22.5
	2017-01-30	R	57783.70939	21.10 +/– 0.11
	2017-01-31	R	57784.69874	21.22 +/– 0.12
	2017-02-07	R	57791.93095	21.60 +/– 0.40
	2017-02-18	R	57802.73167	21.89 +/– 0.24
PS17pv	2017-01-25	R	57778.82847	>20.5
	2017-01-28	R	57781.86958	>22.4
	2017-01-30	R	57783.80134	20.74 +/- 0.11
PS17nv	2017-01-25	R	57778.74557	>22.2
	2017-01-27	R	57780.82339	>22.2
	2017-01-31	R	57784.03700	>23.4
PS17qk	2017-01-25	R	57778.77275	21.01 +/– 0.12
	2017-01-29	R	57782.78859	20.61 +/– 0.11
	2017-01-30	R	57783.74301	20.46 +/– 0.06
	2017-01-31	R	57784.73083	20.36 +/– 0.05
	2017-01-31	R	57784.84479	20.50 +/– 0.03
	2017-01-31	B	57784.85668	20.87 +/– 0.04
	2017-02-01	R	57785.77138	20.17 +/– 0.05
	2017-02-18	R	57802.75716	20.55 +/– 0.07
	2017-03-06	CR	57818.86623	20.53 +/– 0.09
PS17rc	2017-01-25	R	57778.81470	20.96 +/– 0.15
	2017-01-30	R	57783.77229	21.04 +/– 0.10
	2017-01-31	R	57784.76358	21.13 +/– 0.11
	2017-01-31	R	57784.88885	21.28 +/– 0.05
	2017-01-31	R	57784.87345	23.57 +/– 0.24
	2017-02-01	R	57785.80742	21.28 +/– 0.09
MASTER OT	2017-01-21	R	57774.62327	19.07 +/– 0.01
J090737.22+611200.5	2017-01-22	R	57775.64491	19.10 +/– 0.02
	2017-01-23	R	57776.63227	19.18 +/– 0.03

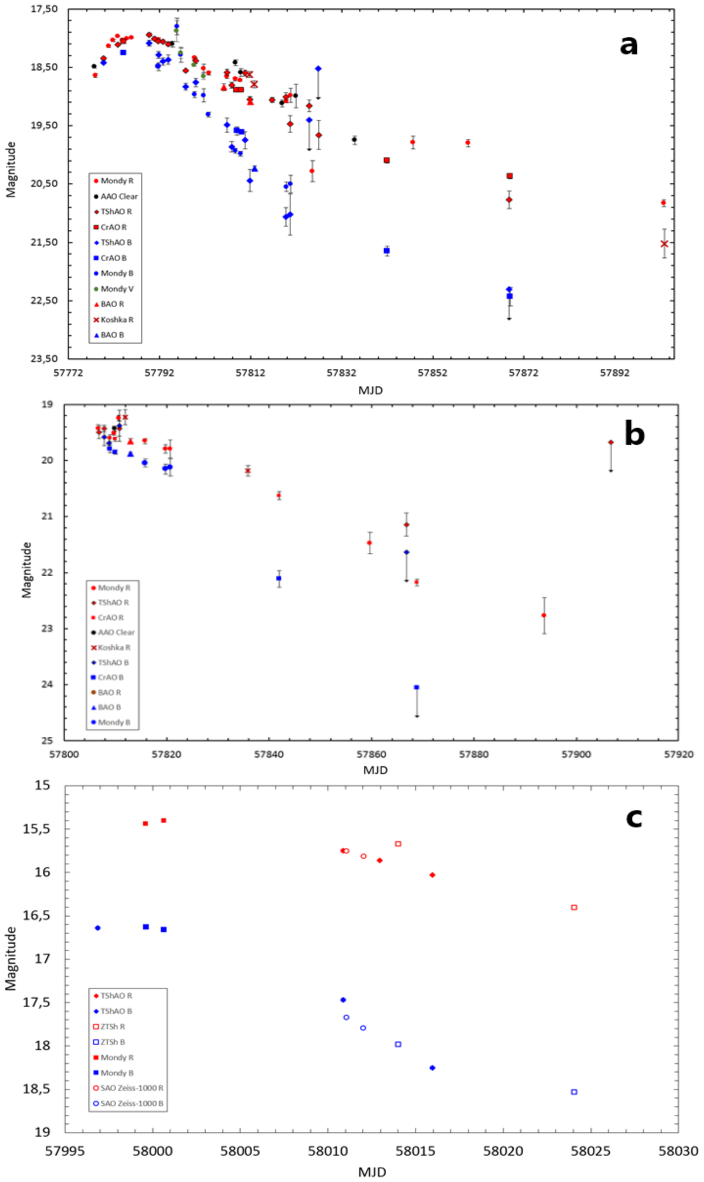


Fig. 5. a – light curves of PS17yt (GW170120_G270580 optical candidate), b – light curves of PS17bek (GW170217_G274296 optical candidate), c – light curves of MASTER-OT (GW170825_G299232 optical candidate). Red points are R-band, blue points are B-band and green points are V-band. Observations were obtained by TShAO (Zeiss-1000), CrAO (ZTSh – 2.6m), Mondy (AZT-33IK), AAO (AS-32), BAO (ZTA 2.6-m), Simeiz/Koshka (Zeiss-1000). Host galaxy is not subtracted

We discussed the two main methods of the search for optical transients in the areas of tens and hundreds of square degrees: mosaic surveys and observations of pre-defined targets (potential host galaxies). The case of mosaic surveys is suitable for small-aperture telescopes with wide fields of view, with rather low optical upper limit, though. The search of the transient inside pre-defined target galaxies requires deeper limits and thus require observations with large-aperture telescopes with >1 meter diameter. The second case involves compiled catalogues of galaxies with known distance like Galaxy List for the Advanced Detector Era (GLADE) [22]. This fact increases the actuality of deep surveys of galaxies with measured distances. These methods are suitable not only for the search of the EM counterpart of gravitational waves events detected by LIGO/Virgo, but also for the search of optical counterparts of ordinary GRBs with large localization region (e.g., from GBM/Fermi experiment).

We also provided results of the observations of localization regions of candidates for real GW events detected with LIGO/Virgo during their second scientific run O2. We did not find any optical transients with our facilities; however, we conducted a follow-up of transients discovered by other teams worldwide. This valuable experience is now being adapted for the third scientific run O3 of LIGO/Virgo, which started on April 1, 2019 and would continue for 1 year. Nevertheless, the problem of automatization of the data processing algorithms remains unsolved for all cases and requires the development of new conceptual approach, and generalized pipelines for data reduction are required.

Almost all space and ground-based astronomical facilities are now involved in the follow-up of GW events. This makes multi-messenger astronomy a commonplace nowadays. Quick availability of new obtained data and vast collaboration of observatories and observers may guarantee further success.

Acknowledgments

Authors are grateful to the grants RFBR 19-42-910014, 17-51-44018 and 17-02-01388 for a partial support.

References

1. GBM/Fermi, https://fermi.gsfc.nasa.gov/science/instruments/gbm.html, last accessed 2019/04/21.
2. SDSS, https://www.sdss.org, last accessed 2019/04/21
3. Gaia, http://sci.esa.int/gaia, last accessed 2019/04/21
4. Event Horizon Telescope, https://eventhorizontelescope.org, last accessed 2019/04/21
5. IceCUBE, https://icecube.wisc.edu, last accessed 2019/04/21
6. LIGO,https://www.ligo.caltech.edu, last accessed 2019/04/21
7. Virgo, www.virgo-gw.eu, last accessed 2019/04/21
8. Aasi, J., Abbott, B.P., Abbott, R. et al.: Advanced LIGO. Classical and Quantum Gravity, **32** (7), article id. 074001 (2017).
9. Abbott, B.P., Abbott, R., Abbott, T.D. et al.: Observation of Gravitational Waves from a Binary Black Hole Merger. Physical Review Letters **116** (6), id.061102 (2016).

10. Acernese, F., Agathos, M., Agatsuma, K. et al.: Advanced Virgo: a second-generation interferometric gravitational wave detector. Classical and Quantum Gravity **32 (2)**, article id. 024001 (2015).
11. Abbott, B.P., Abbott, R., Abbott, T.D. et al. GW170817: Observation of Gravitational Waves from a Binary Neutron Star Inspiral. Phys. Rev. Lett. **119**, 161101 (2017).
12. Bisikalo, D.V., Zhilkin, A.G., Kurbatov, E.P.: Possible: Electromagnetic Manifestations of Merging Black Holes. Astronomy Reports **63** (1), 1–14 (2019).
13. Li, L. and Paczynski, B.: Transient Events from Neutron Star Mergers. Astrophysical Journal **507** (1), L59–L62 (1998).
14. Abbott, B.P., Abbott, R., Abbott, T.D. et al.: Multi-messenger Observations of a Binary Neutron Star Merger. Astrophysical Journal Letters **848**, L12 (2017).
15. Tanvir, N.R., Levan, A.J., González-Fernández, C. et al.: The Emergence of a Lanthanide-rich Kilonova Following the Merger of Two Neutron Stars. Astrophysical Journal Letters **848** (2), article id. L27, 9 pp. (2017).
16. The LIGO Scientific Collaboration and the Virgo Collaboration. LIGO/Virgo Identification of a binary neutron star candidate coincident with Fermi GBM trigger 524666471/170817529. GCN Circ. 21509 (2017).
17. Cook, D.O., Van Sistine, A., Singer, L., and Kasliwal, M.M.: LIGO/Virgo G298048: Nearby Galaxies in the Localization Volume. GCN Circ. 21519 (2017).
18. Valenti, S., Sand, D.J., Yang, S., Cappellaro, E., Tartaglia, L., Corsi, A., Jha, S.W., Reichart, D.E., Haislip, J., and Kouprianov, V.: The Discovery of the Electromagnetic Counterpart of GW170817: Kilonova AT 2017gfo/DLT17ck. Astrophysical Journal Letters **848**, L24 (2017).
19. Pozanenko, A.S., Barkov, M.V., Minaev, P.Yu., Volnova, A.A., Mazaeva, E.D., Moskvitin, A.S., Krugov, M.A., Samodurov, V.A., Loznikov, V.M., and Lyutikov, M.: GRB 170817A Associated with GW170817: Multi-frequency Observations and Modeling of Prompt Gamma-Ray Emission. Astrophysical Journal Letters **852** (2), article id. L30, 18 pp. (2018).
20. Позаненко, А., Вольнова, А., Минаев, П., Самодуров, В.: Поиск компонентов источников гравитационных волн в электромагнитном диапазоне и с помощью методов астрономии космических лучей. Аналитика и управление данными в областях с интенсивным использованием данных: XVIII Международная конференция DAMDID / RCDL’2016.
21. Abbott et al.:GWTC-1: A Gravitational-Wave Transient Catalog of Compact Binary Mergers Observed by LIGO and Virgo during the First and Second Observing Runs. Eprint arXiv:1811.12907 (2018).
22. Dálya, G., Galgóczi, G., Dobos, L., Frei, Z., Heng, I.S., Macas, R., Messenger, C., Raffai, P., and de Souza, R.S.: GLADE: A galaxy catalogue for multimessenger searches in the advanced gravitational-wave detector era. Monthly Notices of the Royal Astronomical Society **479** (2), 2374–2381
23. Bertin, E. and Arnouts, S.: SExtractor: Software for source extraction. Astronomy and Astrophysics Supplement **117**, 393–404 (1996).
24. Devyatkin, A.V., Gorshanov, D.L., Kouprianov, V.V., and Verestchagina, I.A.: Apex I and Apex II software packages for the reduction of astronomical CCD observations. Solar System Research **44** (1), 68–80 (2010).
25. The LIGO Scientific Collaboration and Virgo report: LIGO/Virgo G299232: Identification of a GW Compact Binary Coalescence Candidate. GCN Circ. 21693 (2017).

26. SingerLeo, P., Chen, Hsin-Yu, Holz, Daniel E. et al.: Going the distance: mapping host galaxies of LIGO and Virgo sources in three dimensions using local cosmography and targeted follow-up. Astrophysical Journal Letters **829** (1) (2016).
27. Bartos, I., Countryman, S., Finley, C. et al.: LIGO/Virgo G299232: FOUND COINCIDENT IceCube neutrino observation. GCN Circ. 21694 (2017).
28. Bartos, I., Countryman, S., Finley, C. et al.: LIGO/Virgo G299232: COINCIDENT IceCube neutrino observation UPDATE. GCN Circ. 21698 (2017).
29. The LIGO Scientific Collaboration and Virgo report: LIGO/Virgo G268556: Updated sky map from gravitational-wave data. GCN Circ. 20385 (2017).
30. The LIGO Scientific Collaboration and Virgo report: LIGO/Virgo G270580: Identification of a GW Burst Candidate. GCN Circ. 20486 (2017).
31. The LIGO Scientific Collaboration and Virgo report: LIGO/Virgo G274296: Identification of a GW Burst Candidate. GCN Circ. 20689 (2017).
32. The LIGO Scientific Collaboration and Virgo report: LIGO/Virgo G275697: Updated localization from LIGO data. GCN Circ. 20833 (2017).
33. The LIGO Scientific Collaboration and Virgo report: LIGO/Virgo G277583: Identification of a GW Burst Candidate. GCN Circ. 20860 (2017).
34. The LIGO Scientific Collaboration and Virgo report: LIGO/Virgo G299232: Identification of a GW Compact Binary Coalescence Candidate. GCN Circ. 21693 (2017).
35. Chambers, K.C., Smith, K.W., Young, D.R. et al.: LIGO/Virgo G270580: Pan-STARRS coverage and bright, rising transient PS17yt. GCN Circ. 20512 (2017).
36. Huber, M.E., Chambers, K.C., Smith, K.W. et al.: LIGO/Virgo G270580: Pan-STARRS coverage and 124 optical transients. GCN Circ. 20518 (2017).
37. Castro-Tirado, A.J., Casanova, V., Zhang, B.-B. et al.: LIGO/Virgo G270580: 10.4m GTC spectroscopic observations of PS17yt. GCN Circ. 20521 (2017).
38. Chambers, K.C., Smith, K.W., Huber, M.E. et al.: LIGO/Virgo G274296: Pan-STARRS imaging and discovery of 70 transients. GCN Circ. 20699 (2017).
39. Gal-Yam, A., Leloudas, G., Vreeswijk, P. et al.: LIGO/Virgo G274296: PS17bek is a superluminous supernova at z=0.31. GCN Circ. 20721 (2017).
40. Lipunov, V.M., Gorbovskoy, E., Kornilov, V.G. et al.: LIGO/Virgo G299232/PGWB170825.55: MASTER Global-Net OT inside NGC1343 discovery. GCN Circ. 20719 (2017).
41. Jonker, P.G., Fraser, M., Nissanke, S. et al.: LIGO/Virgo G299232: WHT spectrum of MASTER OT J033744.97+723159.0. GCN Circ. 21737 (2017).
42. Reig, P. and Panopoulou, G.V.: LIGO/Virgo G299232: RoboPol observations of MASTER OT J033744.97+723159.0. GCN Circ. 21802 (2017).

Multi-frequency Observations and Discovery of a Supernova Associated with the GRB 181201A

S. Belkin[1,2], A. Pozanenko[1,2], E. Mazaeva[2], A. Volnova[2], P. Minaev[2], N. Tominaga[3,4], S. Blinnikov[2,4,5], D. Chestnov[6], E. Klunko[7], I. Reva[8], V. Rumyantsev[9], D. Buckley[10], R.Ya. Inasaridze[11]

[1] Moscow Institute of Physics and Technology (MIPT), Institutskiy Pereulok, 9, Dolgoprudny, 141701, Russia
[2] Space Research Institute (IKI), 84/32 Profsoyuznaya, Moscow, 117997, Russia
[3] Department of Physics, Faculty of Science and Engineering, Konan University, 8-9-1 Okamoto, Kobe, Hyogo 658-8501, Japan
[4] Kavli Institute for the Physics and Mathematics of the Universe (WPI), The University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa, Chiba 277-8583, Japan
[5] Institute of Theoretical and Experimental Physics (ITEP), Moscow, 117218, Russia
[5] Keldysh Institute of Applied Mathematics (KIAM), Russian Academy of Sciences, Miusskaya 4, Moscow, 125047, Russia
[7] Institute of Solar Terrestrial Physics (ISTP), Irkutsk, 664033 Russia
[8] Fesenkov Astrophysical Institute (FAPHI), Almaty, 050020, Kazakhstan
[9] Crimean Astrophysical Observatory (CrAO), Nauchny, Crimea, 298409
[10] South African Astronomical Observatory (SAAO), Cape Town, 7925, South Africa
[11] Kharadze Abastumani Astrophysical Observatory (AbAO), Ilia State University, Tbilisi, 0162, Georgia
astroboy96@mail.ru

Abstract. Despite the explosive growth in the amount of data in astronomy, one of the main cases is the search for new objects (and identification of its parameters) with a limited number of observations. We discuss the possibilities of making a decision in the conditions of limited amount of information (data set) obtained as a result of a large number of observations and identifying the optimal number of independent parameters that allow exploring and describing the phenomenon. These cases well known in astrophysics, e.g. when searching for electromagnetic counterparts of gravitational-wave events detected by LIGO/Virgo detectors and searching and classifying supernova associated with Gamma-Ray Bursts (GRB). We describe observations and discovery of the supernova (SN) associated with gamma-ray burst GRB 181201A and present preliminary parameters of the SN. The positive decision about SN was generated based only on five high-precision observations. This is one more discovery of SN among only about thirty cases of photometric confirmation of the SN associated with GRBs. The discovery is made possible due to networked telescopes in both Southern and Northern hemisphere.

Keywords: Gamma-ray burst, Afterglow, Supernova, Photometry, Multi-frequency observations

1 Introduction

Nowadays, the problem of recovering the parameters of a phenomenon on a limited amount of data is important in cases, when we cannot obtain more input data for any reasons. Nevertheless, this problem is also a crucial point in a situation of determining the optimal amount of experimental data necessary to describe the phenomenon. Both of these items could be solved when we have a prior information about a number of independent parameters of the phenomenon under study. Such a need arises, for example, in cases of a transient phenomenon, when we cannot repeat the observation. It is obviously that the solution of the problem of optimizing the number of observations is necessary and in demand when planning the search for transient sources in future projects ground based optical telescope LSST [1] and space born X-ray observatory SRG [2].

Two last decades of observations and investigations of gamma-ray bursts (GRBs) and their optical counterparts led to the unambiguous association between at least some long GRBs and the death of massive stars. The observational connection of long GRBs with type Ic supernovae (SNe) supports this evidence. The first reliable association between GRB 980425 and type Ic SN 1998bw with broad spectral lines was both positional and temporal, and spectral data of the two events showed the same redshift of 0.0085 (~ 40 Mpc) [3–5]. The next confirmation of GRB-SNe associations occurred in 2003, with the discovery of very bright GRB 030329 associated with type Ic SN 2003dh [6–8]. The kinetic energy of both these SNe exceeded 10^{52} erg, so they were hypernova (the name of unusual SN suggested by B. Paczynski [9]). The launch of the Swift space observatory [10] changed the way of GRBs investigation dramatically. An early discovery of GRBs optical counterparts and their fast follow-up with ground-based telescopes allowed to build detailed multicolor light curves and to obtain valuable spectroscopic data.

Generally, the optical light curve (LC) of a long GRB may be described by four prominent phases. The first phase is related to the prompt phase when the central engine is still producing energy. This phase is very hard to observe because of relatively slow reaction of optical instruments: usually, when optical telescopes begin to observe the localization region of the burst, the prompt phase is already finished, that’s why there are rather few cases of the prompt phase observations in optical domain. The second phase is usually the longest and is related to the afterglow. A simple power law or a broken power law with two different decay indices can describe it as a good model, and a break is a geometric effect related to the collimation of the GRB jet. This phase may also demonstrate some flares or wiggles [11]. On the 7–20th day the SN feature may appear. It may look like a bump or a slight re-brightening on the light curve, which deviates significantly from the afterglow power law. Spectra obtained during this phase usually show broad lines common for Ic type SNe. After the end of all activities, the source fades away, and the host galaxy may be observed at its location.

Today there are only few dozens of GRB-SNe discovered, and 23 of them are confirmed with spectroscopic observations and 28 are detected only by photometric evidence. The observed flux from the GRB-SN is composed of the afterglow flux, the SN itself and the constant flux of the host galaxy. A careful decomposition of the three components is necessary to obtain the LC of the SN for further determination of its bolometric properties. The decomposition should also take into account the line-of-sight extinction in the Milky Way (e.g., by using the extinction maps by Schlafly and Finkbeiner [12]) and in the host galaxy (e.g., by modelling its spectral energy distribution and comparing it with models of well-studied galaxies). Every listed component may be included in the fitting procedure as an additional parameter or a set of parameters [13–15]. This phenomenological approach is based on the standard GRB theory, that states that the light powering the AG is synchrotron in origin, and therefore follows a power-law behavior in both time and frequency [16].

2 Observations

Optical observations of the source

After registration of GRB181201A with INTEGRAL [17], LAT/Fermi [18], Konus-WIND [19], XRT/Swift [20], Insight–HXMT [21], AstroSAT CZTI [22], we observed this source over the next month. Optical data were made by observatories located in Chile, South Africa, Crimea and Tien Shan, which are part of our IKI GRB Follow-up Network.

The astronomical observatory Gemini [23], which is not a part of our network and located in Hawaii, also made a significant contribution to the construction of the light curve. Because the observations on Gemini were made on the Gemini-North telescope, which is 8.1 meter in aperture. That is much bigger than telescopes in other observatories (D=0.7–2.6m). This fact made observations possible even when the culmination of the object was almost gone for the daytime.

As a result of the monthly observation of the source, the data shown in Table 1 were obtained.

XRT observations of the source

When gamma-ray burst triggers our space vehicles, we must make as many observations in different ranges of energy as possible. In a prevailing number of cases, the space X-ray telescope discovers the x-ray afterglow and provides an accuracy localization of the source within an error circle about several arcseconds. This allows optical telescopes be more productive in optical component search.

The same thing happened in a case of GRB181201A, whose position was observed by Swift/XRT 4 hour after INTEGRAL trigger time [20]. The results of observations of that source are shown in Fig. 3.

Optical observation of the host galaxy

The observation of the source’s host galaxy, which was made six months after the GRB happened on ZTSh telescope in Crimea. This observation was carried out in order to further taking into account the contribution of the host galaxy to the flux from our source. This is necessary for accurate parameter evaluation of the studied phenomenon.

Comparison the host photometry and latest photometry of the afterglow we found that the host influence of the light curve is no more than 15% of any part of the light curve in r'- filter. This small influence alone cannot explain the flattening of the light curve proposed by Laskar et al. [24] instead of the discovered supernova in our study.

3 Data Processing

Before we began to analyze the observations, which had been obtained, we had made a preliminary reduction (dark subtraction and flat-field correction) of all the images from all observatories we collaborated with. This had been made by using task “ccdproc” of NOAO’ IRAF software package which is stands for Image Reduction and Analysis Facility. This is a general purpose software system for the reduction and analysis of astronomical data. IRAF is written and supported by the National Optical Astronomy Observatories (NOAO) in Tucson, Arizona. NOAO is operated by the Association of Universities for Research in Astronomy (AURA), Inc. under cooperative agreement with National Science Foundation [25]. Images from each epoch of observations in corresponding filters had been combined by “imcombine” task with a purpose of providing better signal-to-noise ratio. All magnitudes had been obtained using aperture photometry by APPHOT package within IRAF.

All our instrumental magnitudes had been calibrated according to the SDSS-DR12 photometric catalogue. The reference start was chosen so that between observations in two epochs there was no significant change in the magnitude of the filter we need. If this condition is met, then we can say that this star is not a variable and can be used as a reference. The second criterion for choosing a reference star is that it should not be overexposed on our images. It provides us an accurate determination of the magnitude of our reference star.

The reference star we had used has coordinates 319.29508 and -12.618443, which stands for RA(J2000) and DEC(J2000) respectively.

4 Observation Results

As a result of three weeks observations [26] and data processing, we discovered the supernova [27] and constructed a multicolor light curve, which is shown in Fig. 1.

We know the redshift of this source and it is z=0.45 [33]. We also know that most of the gamma-ray bursts with red-shift less than 0.4 are characterized by presence of a supernovae feature. Considering these facts, we want to check if there is a signature of supernovae in case of GRB181201A.

To do this, we build the light curve in fluxes and fit it with a power function with a slope of –1.2, which well describes the afterglow stage in all filters (see Fig. 2).

Assuming that the optical afterglow evolves achromatically and using the information about the slope of the afterglow stage in r’-filter, we can also describe this stage in z’ filter where we have only one point on a period of afterglow stage.

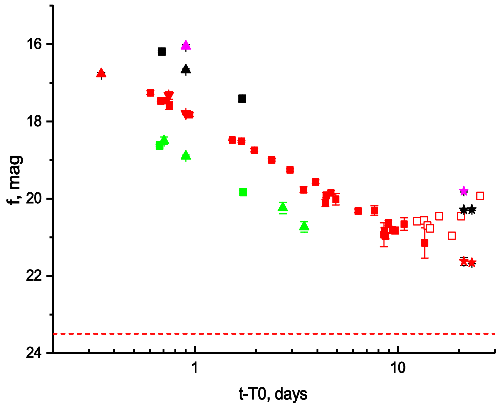


Fig. 2. Multicolor light curve in magnitude units obtained by different scopes. Where pink, black, red and green colors indicate observations on z', i', r' and g' filters respectively. Punctured points stand for the upper limits of observations made in the r' filter. Dotted line shows us the magnitude of the host galaxy in R filter. The points in filters i' and z' were raised up and in filter g' lowered by 1 magnitude for ease of viewing the graph. Here, squares represent the data obtained from the observatories of our network, empty squares stand for upper limits in r'-filter, triangles show values taken from circulars and stars show values from the Gemini Observatory. We also used data on this source, obtained at other observatories by other groups and published on GCN circulars [28–32]

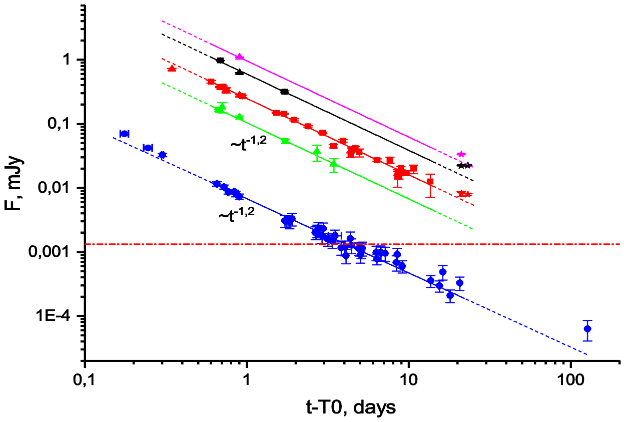


Fig. 3. Multicolor light curve in mJy fitted by power law. Here, squares represent data obtained from observatories of our network, triangles denote data taken from circulars, stars show data taken from Gemini Observatory and circles denote XRT observations. Dashed red line shows the level of the flux from host galaxy in filter r'. The power law index of both optical afterglow and XRT afterglow fitted in the same time interval (0.7–15 days) is equal to –1.2

Fig. 2 gives us the values of the fluxes in different filters only from our source, because after the source had faded we observed and calculated the magnitude and flux of the host galaxy, which is equal to 23.55 in magnitudes in filter R (equivalent of 0.001131 mJy), which was subtracted from data in corresponding filter.

Further, to determine if there are any deviations of our dots from the afterglow, we will construct the graph of residuals. It means that we need to subtract from out light curve the model curve of afterglow stage described by a power law. Before that, contribution to the flux from the host galaxy must be subtracted. It is shown in Fig. 3.

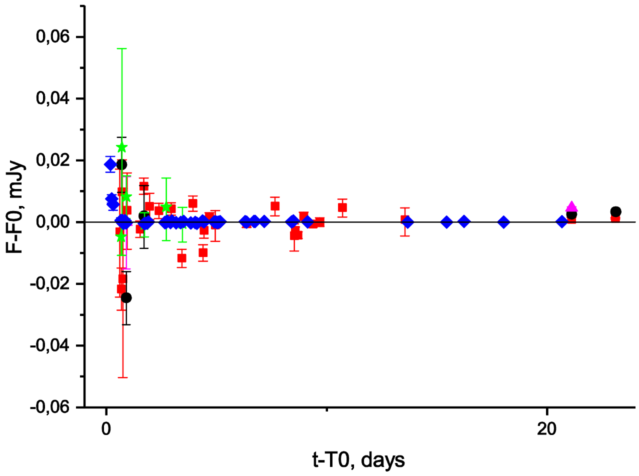


Fig. 4. Residuals of the multicolor light curve. Squares mean deviations our observations from the power law, which describe afterglow stage on the whole range of time since trigger in r' filter, circles in the i' filter, triangles in the z' filter, stars in the g' filter and rhombus in the x-ray range of wavelength

It is noticeable that there is a deviation of our points from the power law extrapolation on 22–24th days which can be better seen if we enlarge this part of a graph (see Fig. 4).

Fig. 4 shows that there is a deviation of our observations from the extrapolation of the afterglow stage and what is more important is that deviation is manifested not only in one point, but in several, in both epochs and even in three filters. This fact gives us a hint that there is some phenomenon other than just only power decay after the gamma-ray burst happened. However, before we talk about the presence of physics in this phenomenon, we must make sure that it is not associated with any kind of error. To do this we construct a graph of the deviation of our observations from the power law, expressed in units of standard deviation, from the time since trigger of the gamma-ray burst.

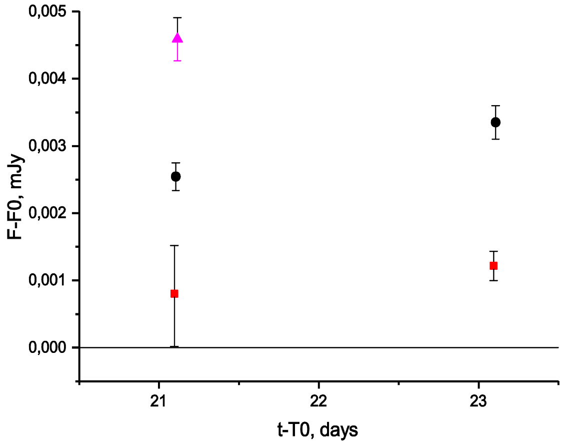


Fig. 5. Residuals of the multicolor light curve (enlarged)

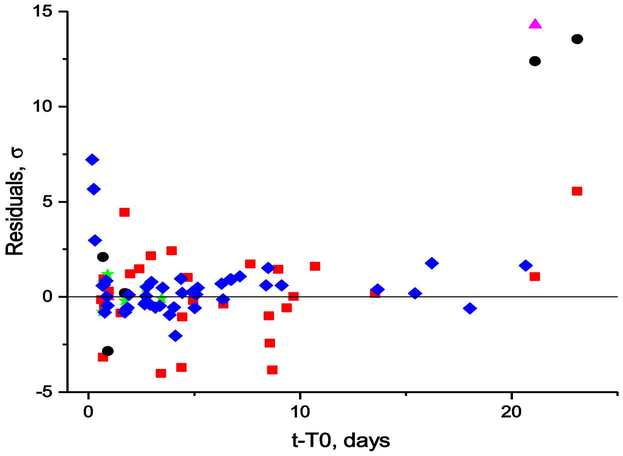


Fig. 6. Significance of deviation of our data from light curve described by the power law, expressed in standard deviation. Here squares denote observations in r’ filter, circles in i’ filter, triangles in z’ filter, stars in g’ filter and rhombus in x-ray range of wavelength

Fig. 5 makes us understand that our data after 21st day after the gamma-ray burst happened deviates from the light curve described by power law and its deviations are significant. That in turn allows us to explore that phenomenon further, rather than link it with the preset error and put it on the back burner.

Now we can go back to Fig. 4 and try to fit any curve to estimate the parameters of the proposed supernova. One of the variant of fitting curve is lognormal distribution, which with such small input data (just 2 points in the filter i' and r') allows us not precisely calculate, but only estimate some of the parameters of that phenomenon. The result of fitting lognormal distribution in our data is in the Fig. 6.

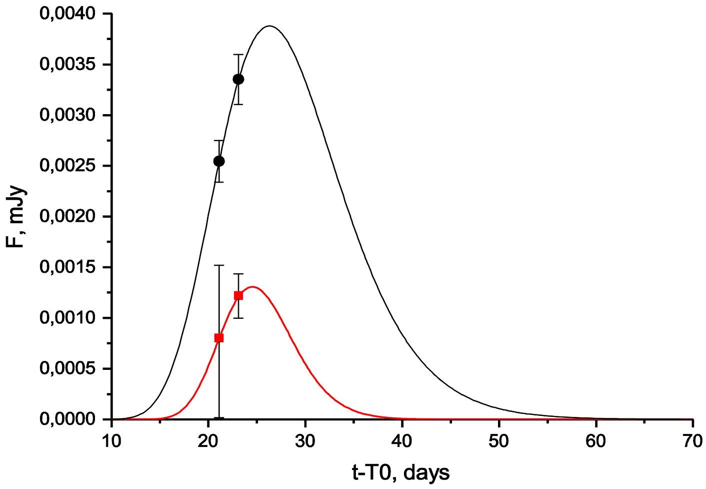


Fig. 7. Supernova light curve in i' (circles) and r' (squares) filters

It is known that there are 4 parameters in lognormal distribution and it is impossible to fit this function in only two dots without any operations with free parameters of the function like fixing one of the parameters or limiting in some range of values. All this manipulations led us to evaluation of such parameters of that phenomenon like:

- Absolute magnitude in filter V: M_v=–19.6;
- Time from the beginning of the burst to the maximum of the supernova on observer's reference frame: t–T0=26.3 days;
- Time from the beginning of the burst to the maximum of the supernova on the rest frame: t–T0=18.138 days.

Now we are able to compare our estimation of parameters of the supernova with those mentioned in Cano's paper [34] on supernova associated with gamma-ray bursts. Based on this work of Cano, it is possible to plot the dependence of the absolute magnitude in the filter V on the time from beginning of the gamma-ray burst to the maximum of the supernova in the rest frame (see Fig. 7).

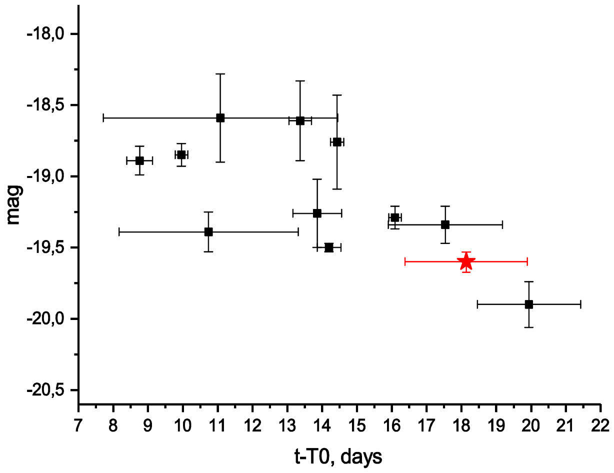


Fig. 8. Comparison of the absolute magnitude in filter V and time of maximum brightness of the SN associated with GRB181201A with the same parameters of previously studied corresponding SNs. Here squares are represent SNs' parameters from paper [34] and star shows where among the parameters of other supernova are the parameters of our

It is noticeable from Fig. 8 that the parameters of supernova, which were discovered by our group, are well placed in a row for already known supernova associated with gamma-ray bursts. This is another additional, among of 27, photometric confirmation of the presence of supernova in gamma-ray bursts.

5 Discussion

We reported preliminary analysis of an observational campaign of the GRB 181201A. Multicolor afterglow observations and a targeted search for a Supernova for this gamma-ray burst were conducted. For these observations, we used IKI GRB Follow-up Network and observations were completed with 7 observatories from all hemispheres, i.e. North, South, East and West ones. Using non-homogeneous data obtained by different observatories we build uniform multicolor light curves in g', r', i' and z' Sloan photometric filters. We also used long-term XRT/Swift X-ray observations.

Based on a few (5) high-precision optical observations on the Gemini-North telescope, a systematic significant excess above afterglow light curve model was found. We suggest that these excesses are due to emerging supernova. Found properties, namely the absolute magnitude and the time of the supernovae maximum in the rest frame, are within the known values [34], which also confirm the discovery of the SN.

Because of that, we are able to summarize that with mentioned earlier amount of observable data we can make a qualitative conclusion about the presence of a photometric signature of a supernova. Moreover, by observing the afterglow and at least two photometric points, one can determine the position of the maximum and the amplitude of the supernova. Minimal necessary conditions for all of this are the availability of at least two observations in each photometric filters at the assumed supernova appearance time interval, as well as reasonably large number of observations at the afterglow stage and host galaxy observations. Provided these conditions are fulfilled one can not only qualitatively identify the supernova, but also find bolometric luminosity and time of maximum which is necessary for minimal quantitative description of the supernova. Of course for identification of the type SN, modeling of physical parameters of SN (mass of progenitor and remnant, abundances of Ni and other chemical elements; see e.g. [35]) one need much more photometric observation for detailed light curve building and spectroscopic observation for measurement of photosphere velocity. Despite the long history of GRB observations since first SN/GRB discovery in 1998, this is one more supernovae associated with GRB among only few dozen (28+23) previously known cases. Moreover, the Supernova is one of the most distant Supernovae (z=0.45) associated with GRB 181201A.

Most photometric discoveries and confirmations of the SNs are based on a limited data set. This dictates the need to develop robust decision criteria for photometric SN confirmation using small amount of useful data based on existing data. More data cannot be obtained for a various reasons: faint source, low flux of the SN over bright afterglow, inability to conduct long-term observations due to transition of the culmination of the source in the daytime and weather conditions.

In our case of supernova search, there are such criteria as simultaneous excess of the flux over the model in different photometric filters, chromatic behavior of the light curves, i.e. color change at the time of SN rise, approximate coincidence of the time of our supernovae’s maximum with the same time of known supernovae associated with GRBs. Of course, the quality and significance of the parameters of the SN obtained will be dependent on the density of the light curve and precision of photometry. The discovery of the SN and obtaining a dense light curve can be provided only by networked telescopes to ensure a long-term observation of the source.

Acknowledgements

Authors are grateful to the grants RFBR Nos 17-52-80139 and 17-02-01388 for a partial support.

References

1. Large Synoptic Survey Telescope homepage: http://www.lsst.org, last accessed 2019/07/10.
2. Spectrum-Roentgen-Gamma homepage: http://hea.iki.rssi.ru/SRG/en/index.php, last accessed 2019/07/10.
3. Galama, T.J., Vreeswijk, P.M., van Paradijs J., et al.: Nature, 395, 670 (1998).
4. Iwamoto, K., Mazzali, P.A., Nomoto, K., et al.: Nature, 395, 672 (1998).
5. Kulkarni, S.R., Frail, D.A., Wieringa, M.H., et al.: Nature, 395, 663 (1998).
6. Hjorth, J., Sollerman, J., Møller, P. et al.: Nature, 423, 847 (2003).

7. Stanek, K.Z., Matheson, T., Garnavich, P.M. et al.: Astroph. J. Lett., 591, L17 (2003).
8. Matheson, T., Garnavich, P.M., Stanek, K.Z. et al.: Astroph. J., 599, 394 (2003).
9. Paczynski, B.: Astroph. J., 494, L45 (1998).
10. Gehrels, N., Chincarini, G., Giommi, P. et al.: Astroph. J., 611, 1005 (2004).
11. Mazaeva, E., Pozanenko, A., and Minaev, P.: International Journal of Modern Physics D **27** (10), 1844012 (2018).
12. Schlafly, E.F. and Finkbeiner, D.P.: Astroph. J., 737, id. 103 (2011).
13. Thöne, C.C., de Ugarte Postigo, A., Fryer, C.L. et al.: Nature, 480, 72 (2011).
14. Sollerman, J., Fynbo, J.P.U., Gorosabel, J. et al.: Astron. Astroph., 466, 839 (2007).
15. Greiner, J., Mazzali, P.A., Kann, D.A. et al.: Nature, 523, 189 (2015).
16. Sari, R., Piran, T., Narayan, R.: Astroph. J., 497, L17 (1998).
17. Mereghetti, S., Gotz, D., Ferrigno, C. et al.: GCN Circ. 23469 (2018).
18. Arimoto, M., Axelsson, M., and Ohno, M.: GCN Circ. 23480 (2018).
19. Svinkin, D., Golenetskii, S., Aptekar, R. et al.: GCN Circ. 23495 (2018).
20. Mereghetti, S., Pintore, F., Gotz, D. et al.: GCN Circ. 23471 (2018).
21. Cai, C., Li, C.K., Li, X.B. et al.: GCN Circ. 23491 (2018).
22. Khanam, T., Sharma, V., Vibhute, A. et al.: GCN Circ. 23501 (2018).
23. Gemini Observatory homepage: http://www.gemini.edu/, last accessed 2019/07/10.
24. Laskar, T., van Eerten, H., Schady, P. et al.: arXiv:1907.13128 (2019).
25. National Optical Astronomy Observatory homepage: http://iraf.noao.edu/.
26. Mazaeva, E., Klunko, E., Belkin, S. et al.: GCN Circ. 23522 (2019).
27. Belkin, S., Mazaeva, E., Pozanenko, A. et al.: GCN Circ. 23488 (2019)
28. Kong, A.K.H.: GCN Circ. 23475 (2018).
29. Heintz, E. Kasper, Malesani, B. Daniele, and Moran-Kelly, Shane: GCN Circ. 23478 (2018).
30. Bolmer J. and Schady P.: GCN Circ. 23486 (2018).
31. Ramsay, G., Lyman, J., Ulaczyk, K. et al.: GCN Circ. 23503 (2018).
32. Srivastava, S., Kumar, H., Otzer, S. et al.: GCN Circ. 23510 (2018).
33. Izzo, L., de Ugarte Postigo, A., Kann, D.A. et al.: GCN Circ. 23488 (2018).
34. Cano, Z., Wang, S., Dai, Z. et al.: Advances in Astronomy **2017**, id.8929054 (2017).
35. Volnova, A., Pruzhinskaya, M., Pozananko, A. et al.: Monthly Notices of the Royal Astronomical Society **467** (3), 3500–3512 (2017).

Astronomical Images in the Light of Big Data

Ekaterina S. Postnikova[1], Natalia V. Chupina[1], Andrei P. Demidov[2],
and Sergei V.Vereshchagin[1]

[1]Institute of Astronomy, Russian Academy of Sciences, Pyatnitskaya str., 48, 119017 Moscow, Russia
[2]Central Aerological Observatory, Pervomayskaya str., 3, Dolgoprudny, Moscow region, Russia
svvs@ya.ru chupina@inasan.ru es_p@list.ru the-admax@ya.ru

Abstract. The role of Big Data images in astronomy is considered, taking into account the usefulness of information in different periods of time. Estimates of the volume of accumulated data in the form of digitized photographic and CCD images are made. The rate of accumulation of information was analyzed. It is shown that only a small percentage of the obtained images at the present stage is already effectively used. We need the development of tools for intensive use and processing to obtain new knowledge. A convenient center for storing general information and simple search capabilities about the entire set of image archives, built on the principles of FAIR is important as part of the Virtual Observatory. The result of the initial development and creation of an image archive of the Zvenigorod Observatory is presented.

Keywords: Big Data, Images, Intensive processing of image archives

1 Introduction

Astronomy accumulated a huge amount of information in publications URL [1], whereas image archives are usually scattered at observatories and institutes. The images contain unique information about the Universe. This information is unrepeatable and can be used to solve scientific and other problems. The lack of a single center leads to the fact that even a professional is hard to find interesting and often necessary data.

With the help of astronomical photographic plates at the present time possible to open a new objects and phenomena. For examples, processing of photographic plates can be used to calculate the orbits of moving objects (asteroids, planets, satellites), as well as to study and search for variable stars, nova and supernova stars.

The aim of this work is determined from the problems of working with image archives carried out by us at the Zvenigorod Observatory. These are:

1) Improving the structure of the available archive of accumulated images obtained at the telescopes of the Institute of Astronomy RAS. It is important to testing the structure. The increasing volume of information and ability to combine photos and CCD images demand the development of using metadata. We need the change the scheme of his work in the framework of the principles of FAIR (Findable – Accessible – Interoperable – Reusable).

2) Estimate and use of the current world practice of storing observation archives in the field of images of objects of the starry sky. Also, develop and use of options for the specific storage of archives of various volumes both at small observatories and at the largest observatories.

3) Development of a way to include archives into service of the open access and data retrieval in the totality of astronomical information that is part of the World data.

2 Photo Images

For more than 100 years, being the pre-CCD era, a large amount of images obtained by the photo, what was the main method for that period have accumulated. Although many observatories did not carried out large-scale observation in the modern sense, the accumulated material yet has large volume, including by the large number of observatories and telescopes. We will begin consideration with the well-known observatory of our institute and proceed to the archives of other observatories and, then, to modern electronic data. More than 100 astronomical observatories operating around the world have similar archives of images taken on photographic plates. With the help of photographic plates images have discovered the satellite of Pluto - Charon Christy and Harrington [2]. Similarly, now the well-known planets outside the Solar System (exoplanets) were first found in the photo image by Farihi and Stoop [3].

3 From Photo Observations to Electronic Images at the Zvenigorod Observatory

Photographic images are used along with digital ones. For this it is necessary to convert them to digital format. For deriving astrometric solutions, object identification and photometry from digitized photo plates the systems like URL [4] – astrometry.net are used, other one – SExtractor [5]. They include photo-observation, conversion photo into electronic form and observation with CCD.

We have 4500 photo plates. Information from each plate is allocated in the amount of 1 GB. In total, it turns 0.0045PB. The observatory has two large telescopes. ASTROGRAPH ZEISS-400 with lens diameter D=40 cm and VAU CAMERA with mirror diameter D=107 cm. More information and details can be found in URL [6].

Photographic survey of the Sky (FON [7]) and FOCAT [8] programs) was carried out from the 1980 to 1992 ys. The FON and FOCAT programs also were used to determine the coordinates of 113 Galactic radio sources [9].

The most interesting images are in our archive of comets. Its distinguishing features are the long series of observations of the brightest comets of the late 20th century (Halley, Hale-Bopp, Hyakutake, etc.). The Zvenigorod archive has 220 images for 11 comets URL [10]. In Fig. 1, we see an example of an image of Hyakutake comet.

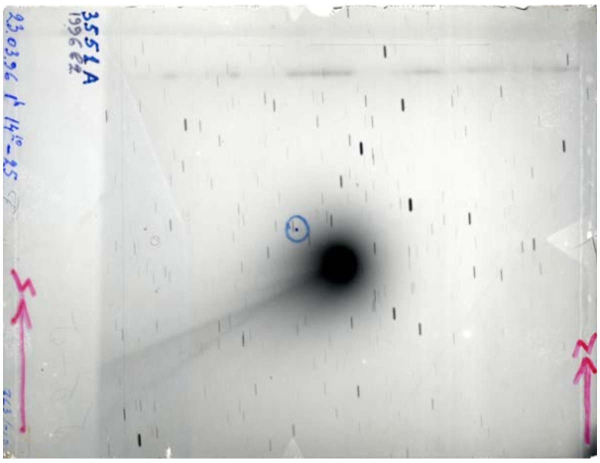


Fig. 1. Comet Hyakutake. Observation at Zvenigorod Observatory. Date of observation: 03/23/1996. Plate Number: 3551A. Plate size: 13 × 18 cm. Obtained by V.P. Osipenko

We have 4500 photo plates, or 4.5 TB. Zvenigorod Observatory can be considered to be similar in terms of information to about a hundred more observatories worldwide. This value will increase two orders and will be approximately 0.5 PB. This, it is easy to understand, is approximately the total amount of photographic images accumulated on the small observatories of the World.

The results of modern observation with Zvenigorod robot-telescope are in digital format, Fig. 2. The stream from the robot-telescope of the one image takes 32 MB. Photometric observations are carried out, on average, with an exposure of 60 – 120 seconds. From 30 to 60 images are recorded in one hour. Given that the night lasts an average of 8 hours, we get 23 GB for one night. In our area in the year 27% of nights are observant, that is, about a hundred nights per year. During this time, you can gain 2.3 TB of information.

Robotic wide-angle (field of view is 10°) system is used for the sky monitoring at the Zvenigorod Observatory INASAN. It makes possible the automatic observations according to a predetermined plan. The telescope robot is based on a wide angle telescope and photo detector with U, B, V, R, I photometric system, Terebizh [11]. The angular diameter of the field of view is 10° for the matrix 50 × 50 mm, the focal length is 395 mm. It has direct drive, [12]. Extending the knowledge of the Zvenigorod Observatory to other ones, we can say that total number of CCD images at all small observatories of the World will be approximately 0.2 PB.

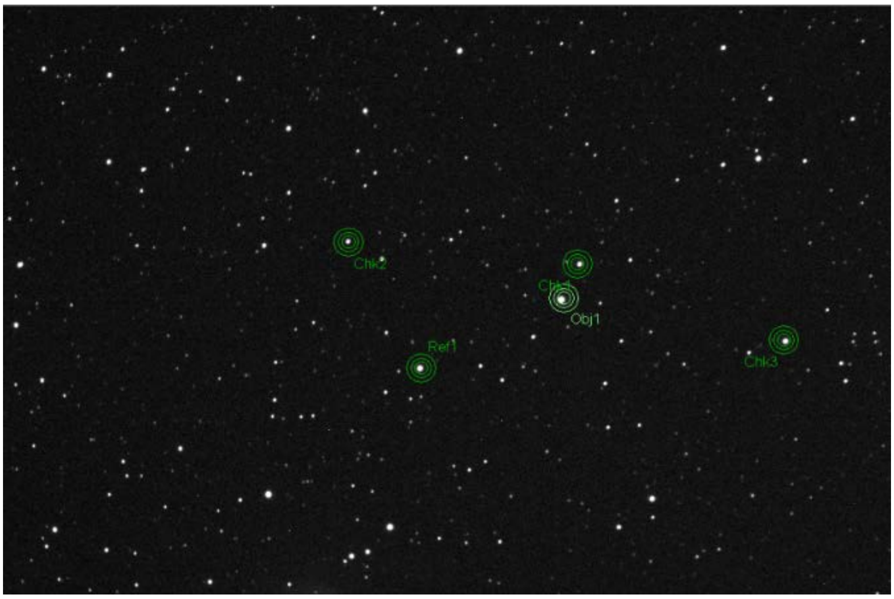


Fig. 2. The LO Peg is a young star of the K3 spectral class and is one of the most studied fast-rotating stars of the late spectral classes. The star is a member of the AB Dor [13] group of stars with common spatial motion. The star gave its name to the moving group AB of Golden Fish, a stellar association consisting of approximately 30 stars that move in the same direction and are approximately the same age. The image was obtained by S.A. Naroenkov and M.A. Nalivkin

4 Large Surveys, Observatories and Telescopes

The **Pan-STARRS** (Panoramic Survey Telescope and Rapid Response System) URL [14] automatic system consist of four Richie-Chretien telescopes with mirrors of 1.8 meters each and 1.4-gigapixel CCD cameras located on the top of Mauna Kea volcano on the island of Hawaii. Its archive size is 1.6 PB, which makes it the largest astronomical data base ever released.

Gaia data. A large space project currently underway is a Global Astrometric Interferometer for Astrophysics (Gaia) URL [15]. Now the second Gaia data, Gaia DR2, have released for approximately 1.7 billion sources brighter than magnitude 21. First Gaia realize (DR1) consist 1 billion row or 351GB of data and the second Gaia DR2 consist 1.7 billion rows or 1.2TB of data.

BTA-6. The Special Astrophysical Observatory is located at a height of 2070 meters in Karachay-Cherkessia. Mirror-reflector 605 cm, azimuthal mount, [16]. The SAO has a “General Observational Data Archive” URL [17]. Roughly it exceed 1 TB. The Harvard College Observatory Astronomical Plate Stacks URL [18]. There are over 500,000 glass photographic plates (memory volume will be a 500 GB).

Very Large Telescope (VLT) is located in the Atacama Desert, mountain Paranal, 2635 m (Chilean Andes). And it belongs to the European Southern Observatory (ESO), which includes 9 European countries. A system consists of four telescopes 8.2 meters and four auxiliary 1.8 meters. According to luminosity, is equivalent to one device with a mirror diameter of 16.4 meters This telescope generates an average of 20–25 gigabytes of information per night URL [19]. The all archive value equal no more than 25 TB.

Large Binocular Telescope (LBT) is located on Mount Graham, at an altitude of 3321m, in Safford, Arizona (USA). Two mirrors of 8.4 meters with a distance of 14.4 meters, which is in terms of sensitivity is equivalent to one mirror with a diameter of 11.8 m. According to estimates, two telescopes of 8.4 meters both generate 92 GB / night of data URL [20]. The all archive value equal no more than 90 TB.

The Gran Telescopio CANARIAS (GTC) stands on top of the extinct volcano Muchachos on one of the Canary Islands, at an altitude of 2396 m, [21]. The diameter of the main mirror is 10.4 m. It generate up to 20 MB of data per second. In the course of a typical night, therefore, it is possible to accumulate up to 720 GB.

Hobby-Eberle telescope (HET) has segmented 11x10 m mirror (effective area 9.2 meters), located in the USA (Davis Mountains – 2026 m, Texas). It is equipped with active optics [11]. The data volume was obtained on this instrument at 120 GB / night and 20 TB in a three year survey [22].

Hubble Space Telescope (HST) was launched in 1990 and should work until 2030. The diameter of its mirror is only 2.4 m, URL [23]. For 29 years, HST has generated 153 Tb of data.

As a result, a separate position is occupied by the PAN-STAR archive with a volume of 1.6 PB. The rest of the total does not exceed 0.35 PB.

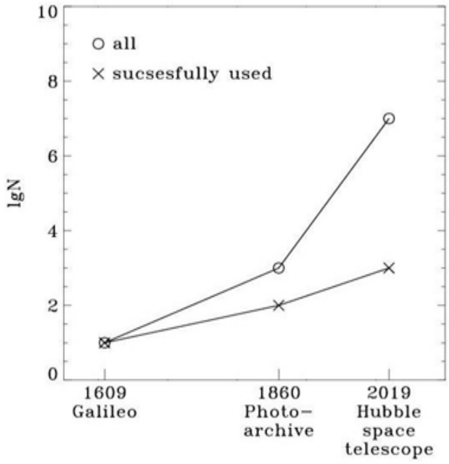


Fig. 3. The increase of information over time. From the Galileo telescope to the Hubble Space Telescope. Small circle shows the logarithm of a number of images (N) representing a full amount of data, and crosses – the volume of data that gave new results. We see that the effectiveness of the “archives” of images is now far from the same as it was before. If Galileo made a discovery from almost every image, then a modern telescope gives only one useful image out of 100,000. The data presented in this figure are estimates only

5 The Increase of Requests

On Fig. 3 the growth of information with time and the change in the efficiency of its use are shown. How many discoveries does astronomical PB contain? Is it true to say that the percentage of images giving new knowledge fell from 100% Galileo to 10 percent in modern observations? If in the search engine of the Digital Library for Physics and Astronomy ADS URL [1] we make a request to get the number of publications with the words “image” and “image&archive”, then we get a number, the dependence of which on the dates is shown in Fig. 4.

As we can see, in Fig. 4 there are no publications about the archives of images before 1980. This is due to the history of digital technology. The CCD was invented in 1969 by Willard Boyle and George Smith at Bell Laboratories (AT & T Bell Labs) URL [24]. In 1970, Bell Labs researchers learned how to shoot images using simple linear devices. Soon these devices appeared as light receivers on telescopes. And since the end of 80-s, publications about archives began to appear.

Just about the images of stars, it was almost always, starting with the book of Ptolemy [25]. Although the rapid increase in the number of publications began precisely on photographic images.

In 1839, on January 7, physicist Francois Arago, at a meeting of the Paris Academy of Sciences, first reported on the invention of daguerotype [26] by Louis Daguerre and Nicephorus Niepce. By the decision of the IX International Congress of Scientific and Applied Photography, this date is considered the day of the invention of photography. It is after this that we see a rapid increase in the number of publications on the processing of astronomical photographic images in Fig. 4. As we have said, the number of publications on archives has dramatically increased since the late 80s. Looking at Fig. 4, we will see that the graph by the number of publications with the words “image” in the headline also greatly increased its inclination from the same time.

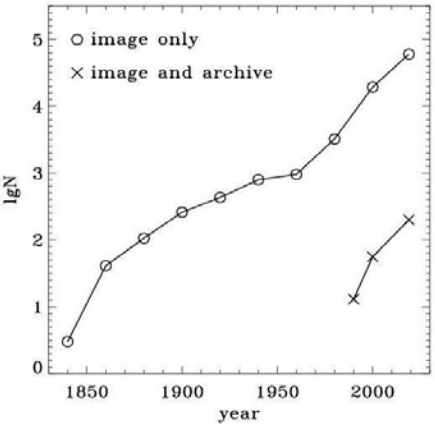


Fig. 4. The cumulative distribution of the number of publications with “image” in the title (small circles) and “image&archive” (slanting crosses)

We propose to consider the efficiency of the telescope (the coefficient of efficiency QE). Obviously, it is determined by the diameter of the mirror / lens -D (the image quality depends primarily on D) and the speed of information accumulation (delta t, time spent on accumulation). We get QE=D VTB / delta t. In Fig. 5 shows the dependence of QE on D.

Table 1. The QE values for different telescope

Telescope	D, m	VTB, in TB	Δt, yr	QE /(TB/yr)	lg QE
ZEISS-400, Zvenigorod	0.40	4.5	30	0.060	–1.20
Zvenigorod robot-telescope	0.25	2.3	1	0.575	–0.24
Gaia	4.80 eff	1.2	5	1.150	0.06
Hobby-Eberle telescope (HET)	10.00	20	3	66.7	1.82
Greater Canary Telescope, GTC	10.40	240	1	2496	3.40
Hubble	2.40	153	28	13.1	1.12

The effectiveness of small telescopes, as we see in Fig. 5, differs up to a thousand times. But even among large telescopes, the QE scatter is large. Table 1 shows data for various telescopes: mirror on lens diameter, volume of accumulated information (VTB), the time interval of data accumulation, QE value and log QE. In order to try to estimate the amount of information extracted from the universe by any telescope, we came up with an “image production coefficient,” or QE. Let us make the following assessment for ZEISS-400. About thirty years of operation were received 4.5 TB information on 40 cm lens. We get a one year by one centimeter of telescope mirror (objective lens for a refractor, as in this case) image output in bytes QE, or QE for Astrograph will equal 0.06 TB/(yr cm).

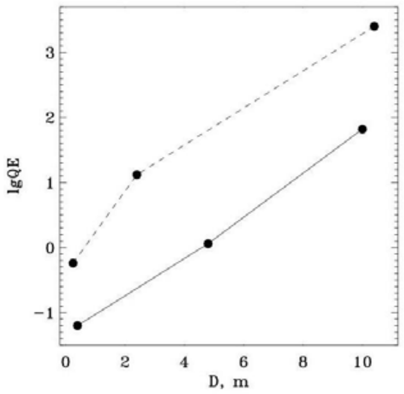


Fig. 5. The logarithm of the efficiency coefficient of the telescope depending on the diameter of the mirror (or lens)

It would seem that the accumulation time is an important parameter to get the value efficiency, but, as determined empirically, it is mainly determined by the diameter of the telescope, Fig. 5.

6 Importance of Format, Problem of Identifying

The problem of identifying stars and other objects cannot always be successfully solved by using automatic identification programs [27]. In our archive the most interesting images obtained by long series of observations of comets, Fig. 6. In such cases the images of stars are obtained extended, which greatly complicates their identification.

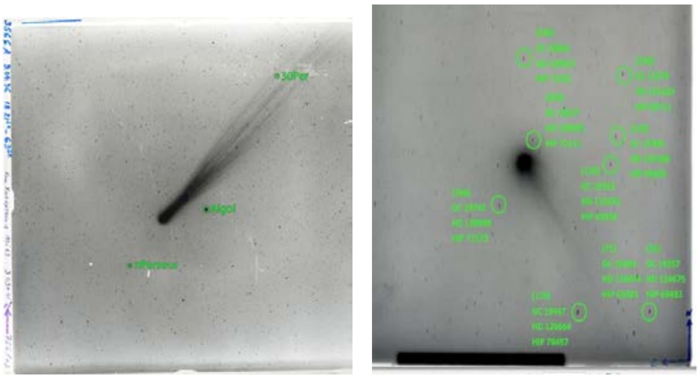


Fig. 6. Examples of the identification of objects on photo plates. On the left we see an extended object – Hyakutake comet. As the telescope was guided on the comet, the images of the stars were stretched. Stretching images of stars will create difficulties for automatic identification. Identification by several catalogs is shown on the right plate – it is Hipparcos, HD, GC, and Star Atlas. The plates were obtained by V.P. Osipenko, identified by V.P. Osipenko and M.D. Sizova

7 Inclusion of Astronomical Image Archives in the World Data Center

There are principles for working with data – FAIR Data Principles URL [28]. These data principles (stands for findability, accessibility, interoperability, and reusability) are a set of guidelines to make data searchable, compatible and reusable [29].

It is necessary to create reliable repositories of data, related audit and certification schemes (for example, Core Trust Seal URL [30] gives both repository requirements and a list of the most reliable ones). These demands are from resulting from the accession to the Data Certificate (DSA) and the certification scheme under the auspices of the Research Data Alliance (RDA, established to: ensure data sharing, overcoming technological, national and disciplinary barriers).

8 Conclusions

The total memory capacity of computer-based storage media required for astronomical archives – 1.6 PB from Pan-STARRS and 0.35 PB for other large telescopes and observatories. The share of small observatories accounts for approximately 0.5 PB photo archives and 0.2 PB for the CCD. In sum, how easy it is to see, 2.65 PB.

It is important to consider the possibilities of connecting to the World Data System. While it is not traditional for astronomical data. WDS URL [31] is building worldwide 'communities of excellence' for scientific data services by certifying Member Organizations – holders and providers of data or data products – from wide-ranging fields by using internationally recognized standards. WDS Members are the building blocks of a searchable common infrastructure, from which a data system that is both interoperable and distributed can be formed. This path, if it is actively supported financially by the government, puts the work with astronomical data before a real choice towards overall improvement.

Results. 1) The result of the development and creation of an image archive of the Zvenigorod Observatory is presented.

2) Astronomical archives of various levels from the small observatories to the largest telescopes were studied. It is shown that today astronomical images are in archives by different structure and access, both at many small observatories and at the largest specialized network resources. A significant part of the images is included in publications, which causes copyright problems. The possibilities of solving the copyright problem developed by the scientific community are shown.

3) A convenient center for storing general information and simple search capabilities about the entire set of image archives, built on the principles of FAIR, is important as part of the Virtual Observatory.

References

1. Astrophysics Data System (ADS), http://adsabs.harvard.edu/abstract_service.html.
2. Christy, J. and Harrington, R.S.: The satellite of Pluto. Astronomical Journal **83**, 1005, 1007, 1008 (1978). http://articles.adsabs.harvard.edu/pdf/1978AJ.....83.1005C
3. Farihi, J. and Stoop, J.: Extrasolar planetary systems were first observed a century ago, evidence suggests Are we celebrating the 20th or 100th anniversary of exoplanet discoveries? https://www.elsevier.com/connect/extrasolar-planetary-systems-were-first-observed-a-century-ago-evidence-suggests
4. Astrometry.net software, http://astrometry.net.
5. Bertin, E. and Arnouts, S.: SExtractor: Software for source extraction. Astronomy and Astrophysics Supplement **117**, 393–404 (1996).
6. Equipment of Zvenigorodskaya observatory, http://www.inasan.ru/en/divisions/zvenigorod/instr/
7. Kolchinsky, I.G. and Onegina, A.B.: On the Programme of Sky Photographing with Wide-Angle Astrographs Astrometriia i Astrofizika **39**, 57 (1979).
8. Bystrov, N.F., Polojentsev, D.D., Potter, H.I., Yagudin, L.I., Zallez, R.F., and Zelaya, J.A. Bulletin d'Information du Centre de Donnees Stellaires **44**, 3 (1994).
9. Rizvanov, N., Dautov, I., and Shaimukhametov, R.: The comparative accuracy of photographic observations of radio stars observed at the Engelhardt Astronomical Observatory. A&A 375, 670–672 (2001).
10. Archive of photoplates scan of the Zvenigorod Observatory INASAN with images of comet Hyakutakehttp://www.inasan.ru/divisions/zvenigorod/scan/scan_hyakutake_comet.
11. Terebizh, V.Y.: On the Capabilities of Survey Telescopes of Moderate Size, Astron. J. 152, 121 (2016).
12. Savanov, I.S., Naroenkov, S.A., Nalivkin, M.A., Puzin, V.B., and Dmitrienko, E.S.: Photometric Observations of LO Peg in 2017, Astrophysical Bulletin **73** (3), 344–350 (2018).
13. Zuckerman, B. and Inseok Song: The AB Dorados moving group, The Astrophysical Journal, (613), L65–L68 (2004).
14. The Pan-STARRS1 data archive home page, https://panstarrs.stsci.edu/
15. Gaia Archive, http://gea.esac.esa.int/archive/
16. BTA-6 telescope hamepage, http://w0.sao.ru/
17. General Observational Data Archive, https://www.sao.ru/oasis/cgi-bin/fetch?lang=ru
18. The Harvard College Observatory Astronomical Plate Stacks, http://tdc-www.harvard.edu/plates/
19. Very Large Telescope (VLT) homepage, http://www.eso.org/sci/facilities/paranal/telescopes/vlti.html
20. Large Binocular Telescope (LBT) homepage, http://oldweb.lbto.org/
21. The Gran Telescopio CANARIAS (GTC) homepage, http://www.gtc.iac.es/
22. Hobby-Eberle telescope (HET) homepage, http://www.as.utexas.edu/mcdonald/het/het_gen_01.html
23. Hubble Space Telescope (HST) homepage, http://hubble.nasa.gov
24. Boyle, W.S. and Smith, G.E.: Charge Coupled Semiconductor Devices. Bell Syst. Tech. J. 49 (4), 587–593 (1970).
25. Ptolemaeus, Claudius Astronomia, teutsch Astronomei: von art, eygenschafften, und himelischen Bildern und iren Sternen wirckung der XII Zeychen des Himels,der VII Planeten, der XXXVI himelischen Bildern und iren Sternen, by Ptolemaeus, Claudius, 1545. (1545). DOI: 10.3931/e-rara-1983.

26. Arago, François: Le Daguerréotype. In: Comptes rendus IX (July–Dec. 1839): 250-67. 4to, 903. Paris: Bachelier, 1839.
27. Automatic stars identification on astronomical images, http://nova.astronet.com
28. FAIR Data Principles, https://libereurope.eu/wp-content/uploads/2017/12/LIBER-FAIR-Data.pdf
29. Wilkinson, M.D., et al.: The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, 3, 160018. doi:10.1038/sdata.2016.18.
30. Core Trust Seal, https://www.coretrustseal.org/.
31. The World Data System (WDS), Interdisciplinary Body of the International Science Council (ISC; formerly ICSU), URL: https://www.icsu-wds.org/organization

PhD WORKSHOP

Benchmarking of Different Approaches for Objects Matching

Dmitry V. Frolov

Master student, Bauman Moscow State Technical University Moscow, Russia
frolov.dmtrii@gmail.com
This work is supervised by Roman S. Samarev, Associate Professor, Ph.D. Computer Systems and Networks Department, Bauman Moscow State Technical University Moscow, Russia
samarev@acm.org

Abstract. Data matching is widely used in different applications. However, very often we do not know data schema completely of the matching objects, because it might be changed during application use. As a result, objects require different processing methods, depending on their structure. To provide this, a special way of storing data should be used. For building an effective application, we need to choose DBMS suitable for selected storage data structure by measuring time spent on different operations with a high load database. The purpose of this article is to demonstrate a benchmark development which is performing these measurements and estimations.

Keywords: Benchmark, Friend of a friend, object matching.

1 Introduction

Nowadays, due to a huge amount of information, describing real-life objects, searching, comparison and choosing the most suitable object requests are becoming more and more actual. Data matching is widely used in different applications. It might be processing of social data and real-time similarity calculation between persons, recommending systems with computation of scores for different goods for a specific person and many other domain areas. The data matching issue can be easily solved when the data structure is fixed and simple, or data can be stored in memory only, but there are many issues when an object has complex, dynamically changed structure and it becomes impossible. In that case, we have to speak about using some storages and databases. Nevertheless, matching of data is a different task comparing with data fetching. Moreover, very often we do not know data schema completely of the matching objects because it may be changed during application use. As a result, the answer to which data structure should be used for matching of some objects is not so obvious. Besides, those objects require different processing methods for different substructures.

The main purpose of this work is to do assessment of different implementations of data storing for solving the task of data matching, on big data databases with complex objects.

To achieve this, we need to conduct a performance analysis of the number of database systems and provide systematic comparison of them. Among analytical modeling methods for are applicable for DBMS's are:

1. Queuing Models: Queuing models are effective to study the dynamics of a database system when it is modeled as a multi-component system with resource allocation constraints and jobs moving around from one component to another. Examples of such dynamic studies are concurrent transaction control algorithms, data allocation and management in distributed database systems etc;
2. Cost Models: Cost Models are useful in studying the cost in terms of Physical storage and query processing time. The cost model gives some real insight into the actual physical structure and performance of a database system;
3. Simulation Modeling: A simulation Modeling is more effective for obtaining better estimates since it not only analyses the database system in isolation but also can effectively analyze the database system with the application program running on top of it and the database system itself operating within the constrained environment of an operating system on real physical hardware;
4. Benchmarking: Benchmarking is the best method when multiple database systems need to be evaluated against each other but suffer from the inherent setback that it assumes all systems to be fully installed and operational. Benchmarking relies on the effectiveness of the synthetic workloads. Real workloads are non-repeatable and hence not good for effective benchmarking.

The most suitable is benchmarking method, because we need to measure a similar operation on a set of DBMS.

2 Review of Benchmarks

There are standardized benchmarks called as TPC or "Transaction Processing Performance Council" [1]. Which are widely used as benchmarks for transaction processing and database performance analysis. These benchmarks do not solely evaluate the database component of the system but rather evaluates the whole system of which the Database system is one of the key differentiating factors. The suite contains a mix of benchmarks for evaluating different performance aspects of such systems.

- TPC-C Benchmark – contains a mix of different types of concurrent transactions, a complex database, nine types of tables with different record and population sizes. It simulates the process of a multi-user environment making concurrent queries to a central database. The performance evaluation involves a thorough monitoring of all system state parameters for the correctness of update as well as performance parameters such as service time etc. This benchmark is most suitable for businesses that need a database to support online handling of orders, sell product and manage inventory.
- TPC-E benchmark – designed for evaluating database systems needed to be installed at brokerage firms. It is quite similar to the TPC-C benchmark in terms of setup and components differing only in the design of the transactions that are more relevant in

a brokerage firm environment such as account inquiries, online trading and market research, etc.

- TPC-H benchmark – is fine-tuned for decision support systems. The transactions in such environments are characterized by business intelligence intensive complex data mining queries and concurrent data modifications. The performance metric used to evaluate such systems is generally TPC-H composite query per hour.

TPC is the largest and most popular benchmarking authority, but, still there are some other benchmarks among them [2]:

- Bristlecone [3] – is a Java-based database performance testing benchmarking utility. It provides knobs to vary the system parameters for a single scenario across different sets of rules or environments. The standard run parameters for synthetic replication include a number of users, number of tables, number of rows per table, number of rows returned by queries or size and complexity of queries, etc.
- CIS benchmark [4] – is a set of security benchmark for the MS SQL Server. These benchmarks provide a testing tool to evaluate these database systems against common security vulnerabilities. Generally, while installing databases most administrators focus on key operating performance issues such as scalability, load balancing, failovers, availability, etc. and let security settings to be default factory settings.
- Yahoo! Cloud Serving Benchmark [5] – is a program suite that is used to compare the relative performance of NoSQL database management systems. The main goal of benchmark is to facilitate performance comparison of transaction-processing workloads which differed from ones measured by benchmarks designed for more traditional DBMSs. YCSB was contrasted with the TPC-H as YCSB is used for big data benchmark while TPC-H is a decision support system benchmark.

Nevertheless, previously mentioned benchmarks are not suitable for data matching measurement. The main problem is that we need to measure big data and decision operations on both SQL and NoSQL DBMS’s it one program. Another problem is that we do not know data schemas of the matching objects completely because it may change during application using. In addition, those objects require different processing methods for different substructures. These facts make the development of your own benchmark relevant. The definition of data storage structures is needed for this purpose.

3 Benchmark Building

While implementing a benchmark, we consider Friend of a friend (FOAF) models as a base for our complex object representation. FOAF describes the world using simple ideas inspired by the Web [6]. In FOAF descriptions, there are only various kinds of things and links, which we call *properties*. The types of things we talk about in FOAF are called *classes*. FOAF is therefore defined as a dictionary of terms, each of which is either a *class* or a *property*. Other projects alongside FOAF provide other sets of classes and properties, many of which are linked with those defined in FOAF.

Definition of operations with complex objects

Complex object operations and requests are ones that can be made to specific objects with high probability. For the benchmark, we define the following list of operations based on FOAF object.

- Find object by a set of FOAF properties;
- Find objects by FOAF property with restrictions;
- Find an object by special property and value;
- Find similar objects.

Three groups can generalize these use cases:

1. Finding an object by certain property or group of properties;
2. Finding an object by word or phrase occurrence in property;
3. Finding a similar objects, according to all properties.

For the first group it is advisable to use simple property search with property value selection by conditions, united by logical “and”, which is integrated into DBSM toolkit.

For the second group fuzzy search is the most suitable. This search type intended to preserve mistakes that are made because of the user’s misprint. The search is proceeded by the entered request first, and then by several similarly written requests.

For the third group comparison on calculating generalized estimates is the most suitable. Working principle of this method is in calculating general estimate for the whole object based on local properties’ estimates. Then the most appropriate objects are selected by comparison of objects’ general estimates.

To make a proper program that is worked with complex objects it is needed to analyze the requests execution speed on different database sizes and models. Benchmark, measuring time spent on a single operation should be programmed to get this information.

Data models definition

As we have different DBMSs’ with different logical and physical data models we need to determine most general data models but with their specifics. As FOAF models are the base for our complex objects, it is suitable to use data models containing full information about storing them. The developed data model should describe the set of object fields for each object in the database.

Three main data models have to be implemented in each test in order to compare DBMS’s performance in similar situations. Nevertheless, there can be additional structures to show DBMSs’ unique features, according to their specific.

1. Field data model – each object property has its own database object, containing the property value and root object identifier (see Fig. 1).

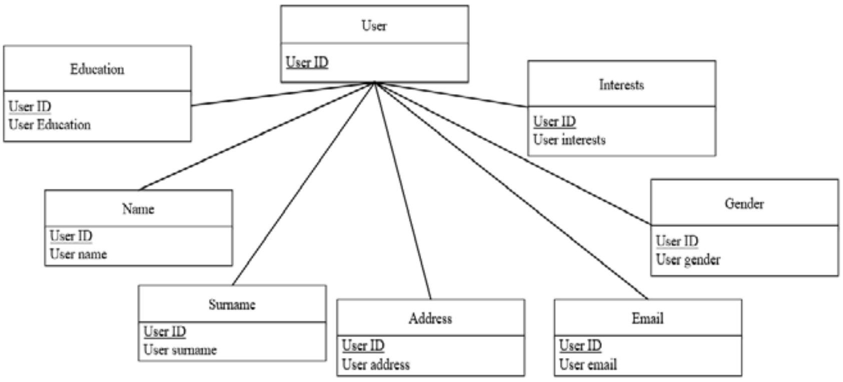


Fig. 1. Field data storage structure diagram

2. Key-value data model – each record in database is key-value pair, where the key is the name of object property and value is the value of this property for a particular object. Also, the identifier of the root object is added to this record to provide connection of the key-value pair with the root object. The example of this structure can be seen in Fig. 2.

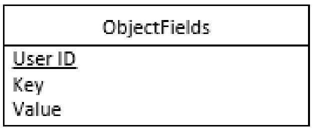


Fig. 2. Key-value data storage structure diagram

3. Tuple data model – each object has its tuple, contains key-value pair for each object property. This tuple also stored with a root object identifier (see Fig. 3).

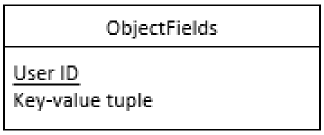


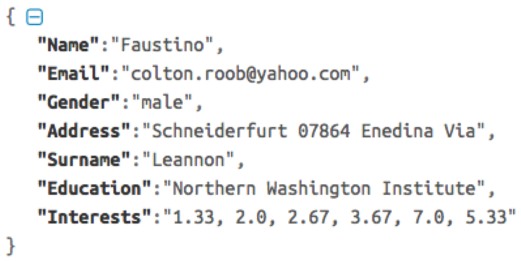
Fig. 3. Tuple data storage structure diagram

Performance measurement has to be done for three different database storage model they are relation, graph and document-oriented. The most common DBMS in their segments are PostgreSQL [7], Neo4j [8] and MongoDB [9] were chosen for analysis.

Two more data storage structures were added to the test according to selected DBMS specific.

4. Json data model – PostgreSQL has a special jsonb type for work with json objects. It stores it in the parsed binary format that helps to speed processing up.

Resulting json in the database consists of key-value pairs for each FOAF property. Fig. 4 is an example of how data stored in json



```
{ ⊟
   "Name":"Faustino",
   "Email":"colton.roob@yahoo.com",
   "Gender":"male",
   "Address":"Schneiderfurt 07864 Enedina Via",
   "Surname":"Leannon",
   "Education":"Northern Washington Institute",
   "Interests":"1.33, 2.0, 2.67, 3.67, 7.0, 5.33"
}
```

Fig. 4. Data stored in json

5. Connected graph data model – connection to root project is provided by edges with property key-value pairs, making connected tree to particular property value. How data is stored in a graph is shown on Fig. 5.

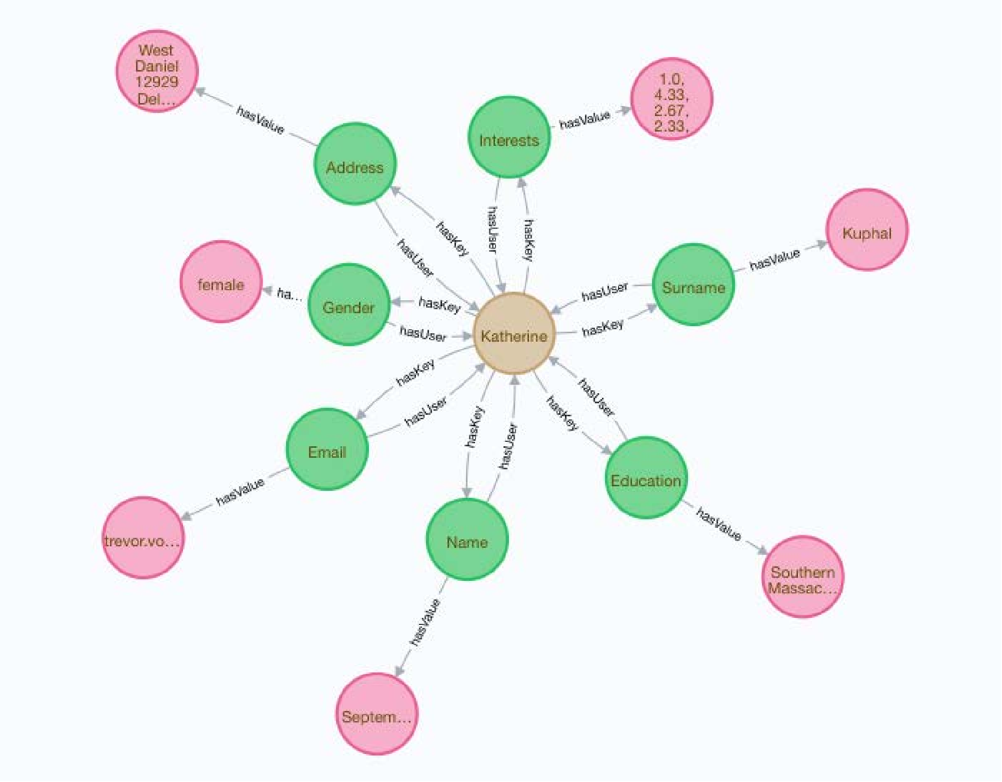


Fig. 5. Data stored in a graph

The table below shows data models distribution across the selected DBMSs.

Table 2. Data **models distribution.**

Data models	PostgreSQL	Neo4j	MongoDB
Field data model	+	+	+
Key-value data model	+	+	+
Tuple data model	+	+	+
Json data model	+	-	-
Connected graph data model	-	+	-

Benchmark architecture

Three modules “Test PostgreSQL”, “Test Neo4j”, “Test MongoDB” were implemented for performance measurement of DBMS. According to program specific, each test should implement three main object operations. The special interface should be created for this purpose, to make operation implementation in each test’s class necessary. Class subject area diagram is shown on Fig. 6.

The diagram represents following classes:

- Test – interface, containing main operation for performance measurement description;
- App – test launcher class;
- TestPostgreSQL – performance test implementation for PostgreSQL DBMS;
- TestMongoDB – performance test implementation for MongoDB DBMS;
- TestNeo4j – performance test implementation for Neo4j DBMS.

All test classes contain three general methods that were declared in Test. These methods are: *seed* – is used for generating random information and save it to selected database; *search* – is used for calling all search-methods in class; *clean* – is used to delete all test data from the database after a test run.

Each class also has specific methods used for searching data. Search queries are depended on data model presented in testing database. For example, search methods are: *testSearchByTuple, testSearchByKeyAndValueInRelations, testSearchSingleRepo.*

4 Benchmark Results

Performance test starts after launching the benchmark’s code on all described DBMS and includes tests with 1000, 10000, 100000 records in the database. During each iteration, a random data, generated by Java Faker library, added to the database. Each added object has seven main fields of FOAF representation. They are: *name, surname, address, email, gender, interests, education*. The example of the database object is on Fig. 7.

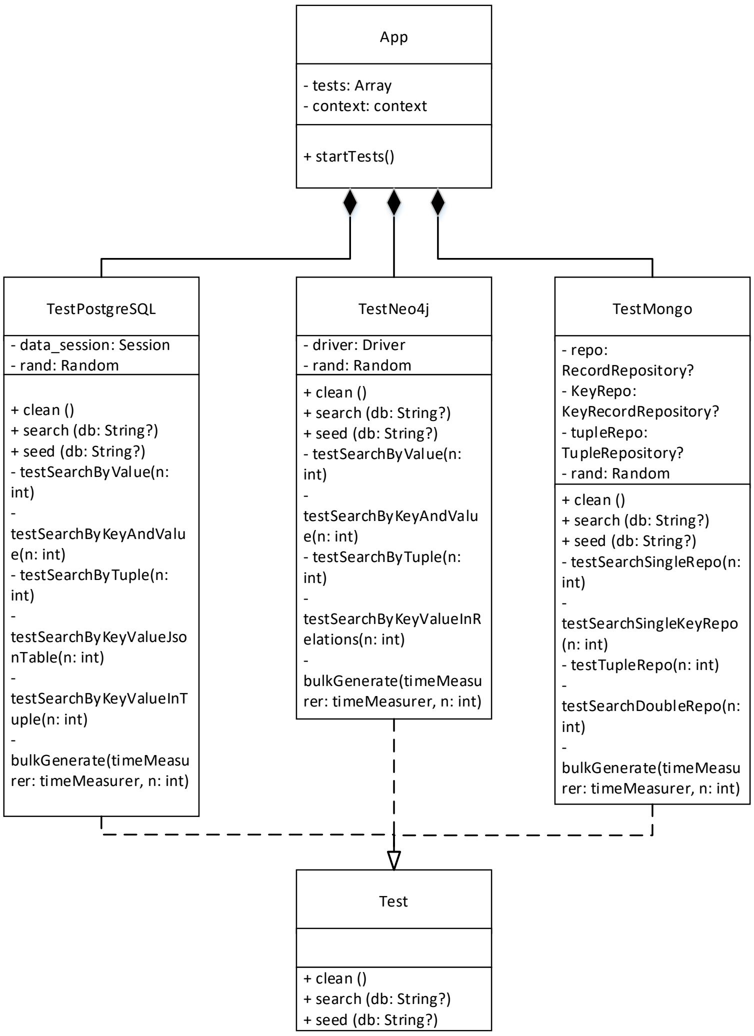

App
- tests: Array
- context: context
+ startTests()
TestPostgreSQL
- data_session: Session
- rand: Random
+ clean ()
+ search (db: String?)
+ seed (db: String?)
- testSearchByValue(n: int)
- testSearchByKeyAndValue(n: int)
- testSearchByTuple(n: int)
- testSearchByKeyValueJsonTable(n: int)
- testSearchByKeyValueInTuple(n: int)
- bulkGenerate(timeMeasurer: timeMeasurer, n: int)
TestNeo4j
- driver: Driver
- rand: Random
+ clean ()
+ search (db: String?)
+ seed (db: String?)
- testSearchByValue(n: int)
- testSearchByKeyAndValue(n: int)
- testSearchByTuple(n: int)
- testSearchByKeyValueInRelations(n: int)
- bulkGenerate(timeMeasurer: timeMeasurer, n: int)
TestMongo
- repo: RecordRepository?
- KeyRepo: KeyRecordRepository?
- tupleRepo: TupleRepository?
- rand: Random
+ clean ()
+ search (db: String?)
+ seed (db: String?)
- testSearchSingleRepo(n: int)
- testSearchSingleKeyRepo(n: int)
- testTupleRepo(n: int)
- testSearchDoubleRepo(n: int)
- bulkGenerate(timeMeasurer: timeMeasurer, n: int)
Test
+ clean ()
+ search (db: String?)
+ seed (db: String?)

Fig. 6. Class subject area diagram

user_id	user_data
762e2f20-b9eb-42aa-914b-af4e525bbb4b	{{Name, Anthony},{Surname, Paucek},{Address, “East Marlon 32552 Ethel Spurs”},{Email,kim.reynolds@gmail.com},{Gender, male},{Interests, “1.0, 4.33, 2.67, 2.33, 7.0, 4.33”},{Education, “West South Carolina Institute”}}

Fig. 7. Database object example

10000 different searching requests are made during tests. Each search operation relates to described data models. In Table 2 this relation are set.

Table 3. Search operations and data models relations

Search operations	Data models
Search by field	Field data model
Search by key and value	Key-value data model
Search by tuple	Tuple data model
Search by key-value in tuple	Tuple data model
Search by key-value in json	Json data model
Search by key and value in relations	Connected graph data model

The performance testing results are presented in Table 3. Accepted abbreviations in the table: P – PostgreSQL 11.2, N – Neo4j 3.5.3, M – MongoDB 4.0.3.

Table 4. Performance testing results (ms)

	1000 records			10000 records			100000 records		
DBMS	P	N	M	P	N	M	P	N	M
Seed	76	5676	274	70311	40276	2167	668076	426424	21412
Search by field	1	1	1	1	1	1	1	1	1
Search by key and value	1	2	1	1	10	1	1	83	1
Search by tuple	1	1	1	1	5	1	1	47	1
Search by key-value in tuple	1	-	2	7	-	21	72	-	243
Search by key-value in json	1	-	-	8	-	-	81	-	-
Search by key and value in relations	-	8	-	-	64	-	-	176	-
Single field search in repo with dynamic key	-	-	1	-	-	6	-	-	55
Clean	9	21	190	11	6578	2756	49	2745	20252

As it is seen from the table, PostgreSQL requires more time to seed information in the database, but it has the best results in searching scenarios on every tested database size. The PostgreSQL loss on seeding information is not critical, as in real application it is impossible to get such load in a small period of time, the database size will grow constantly.

According to this research PostgreSQL and tuple data storage structure are the most suitable for developing complex object matching system.

Conclusion

This work is in progress. In the future, we will extend a number of operations with data and will change model structures. In addition, some additional matching specifics operations will be added. One of the issues, which is not considered in this paper, is benchmarking of big volumes of data in a computational cluster. In that case, huge difference in performance results might be found. This part we also will consider in further work.

References

1. TCP benchmark description, http://www.tpc.org/information/benchmarks.asp, last accessed 2019/05/27
2. Benchmarks description, https://www.cse.wustl.edu/~jain/cse567-08/ftp/db/index.html, last access 2019/05/30
3. Bristlecone, https://github.com/shutterstock/bristlecone/tree/master/src/com/continuent/bristlecone/benchmark, last access 2019/05/30
4. CIS benchmark, https://www.cisecurity.org/cis-benchmarks/, last access 2019/05/30
5. Yahoo! Cloud Serving Benchmark, https://github.com/brianfrankcooper/YCSB, last access 2019/05/30
6. FOAF specification, http://xmlns.com/foaf/spec/, last accessed 2019/05/27
7. PostgreSQL documentation, https://postgrespro.ru/docs/postgresql, last accessed 2019/03/19
8. Cypher documentation, https://neo4j.com/docs/cypher-manual/current/, last accessed 2019/03/19
9. MongoDB documentation, https://docs.mongodb.com/manual/, last accessed 2019/03/19.

Language Integrated Query as a Canonical Data Model for Virtual Data Integration

Vladimir Klyuchikov

Lomonosov Moscow State University, Moscow, Russia
kluchvlad@gmail.com

Abstract. Nowadays data using by organizations in different business areas are very heterogeneous. This raises the issue of data integration. Two main classes of data models used for data representation can be distinguished: relational (SQL) and non-relational (NoSQL) data models. Data models of the classes differ a lot, for instance, relational models are applied for structured data, and NoSQL models mostly applied for semi-structured data. For the issues of data integration it is required to find a model that can unify relational and NoSQL models. A candidate for such unifying data model is Language Integrated Query – LINQ. The aim of this work is to validate that LINQ can successfully serve as unifying data model in data integration systems intended to integrate both relational and NoSQL data sources.

Keywords: Virtual Data Integration, Canonical Data Model, Language Integrated Query

1 Introduction

Currently, there is an exponential increase of the volume of experimentally obtained data in science and industry. The data can be obtained from various sources. For example, researchers can get scientific data from sensors during the experiments. Credit banks capture transactions each borrower and generate a credit history, that can be used in advance. The list of users and personal information, feedbacks and stories in social networks are also a data that continuously increase volume. The number of organizations that get the emerged data in different areas is also large. Data that is stored in various sources like web log files, web pages, documents, etc. possess different levels of structurization: structured, semi-structured, and unstructured data. Data in sources can be presented in various data models. Two large classes of data models can be distinguished: relational (mostly SQL) and NoSQL data models. NoSQL is a group of models with flexible schemas that can be further classified in four main categories: key-value, column-oriented, document and graph data models. These models are intended to represent semi-structured data or even schema-less data. NoSQL models are implemented in respective database management systems. To deal with heterogeneity of data the respective data integration methods and tools are required.

Frequently, the data sources are scattered and getting them directly from sources and processing is quite problematic. That data can be localized in *data lakes* [1] that are repositories for large quantities and varieties of structured or unstructured data. In data lakes the data are stored as-is: no initial structuring or transforming data are presumed. Different kinds of analytics – from dashboards and visualizations to big data processing, real-time analytics, and machine learning – are provided over data lakes to guide better decisions. Data lakes can be created for specific cases, such as analytics, machine learning, real-time data movements or on-premises data movements.

Even if the problem of localization of the initial data is solved, the problem of data integration from different sources remains urgent. Data integration requires the following conceptual specifications: *a global schema* and *mappings binding the global schema and source schemas*. The global schema is the integrated schema serving as a unified representation of schemas of participating data sources. Data integration methods are implemented within data integration systems (DIS).

Two kinds of data integration can be distinguished: materialized and virtual integration. *Materialized data integration* usually proceeds within *data warehouses* [13]. Each source can possess a schema that differs from the warehouse (global) schema. The data are reshaped into global schema using Extract-Transformation-Load (ETL) processes (that implement conceptual mappings) and stored (materialized) into the warehouse database.

Virtual DISs developed for concrete subject areas are called *subject mediators* (mediators, in short). To have access to “fresh” information, a virtual integration system is preferred to a warehouse since it avoids having to propagate updates of the data source to the warehouse. The process of answering user queries in virtual DIS is performed as follows [3]:

- a user poses a query in terms of global schema;
- the query is rewritten into a set of subqueries, each subquery is formulated in terms of some source schema;
- each subquery is passed to a specific *wrapper* of the relevant data source to be executed there;
- answers returned from the wrappers are collected, combined and returned to the user.

Three main techniques for definition of conceptual mappings between global and local schemas for virtual integration are known: Global-as-View (GAV) [14], Local-as-View (LAV) [14] and Global-Local-as-View (GLAV) [4, 8].

The canonical model plays a role of a unifying model, in which the source data models can be represented without loss of information [25].

Various kinds of data models with different semantics are used as canonical models: relational models [9] and their extensions [17], object models [3], XML [15], hypergraph models [23], Web Ontology Language (OWL) [18], RDF [24] etc. However, nowadays one of the properties that the canonical model should possess is the ability to unify the SQL and NoSQL source data models. One of the promising models that hopefully has this property is the Language Integrated Query (LINQ) that is a part of C# language standard ([6], *12.17 Query expressions*). Meijer et al. in [16] shows the duality of SQL and NoSQL models using an area of mathematics called category theory. LINQ is used by Meijer as the representation language to illustrate the duality. LINQ does not require strict data typing – if in relational models the data is strictly typed, in non-relational values are dynamically typed. Also, LINQ does not require preliminary data normalization, which is required in SQL and which, in this case, is not used in NoSQL models. Therefore, based on research in [16], it can be assumed that LINQ can be successfully applied as a canonical model in virtual integration of both SQL and NoSQL data sources. This work aims to confirm this assumption.

To achieve the goal, the following problems have to be solved: (i) selection of concrete data models to be unified in LINQ; (ii) implementation of DIS prototype based on the Global-As-View approach and LINQ as the canonical data model, and (iii) evaluation of DIS via use case in some subject area. In this paper the related work is shortly overviewed and the current progress of DIS prototype implementation is reported and illustrated by a use case.

The remainder of this paper is structured as follows. In Section 2 related work is discussed. The phases of the DIS prototype development and description of its components are shown in Section 3. The initial steps on evaluation of the DIS prototype by a use case are described in Section 4. Finally, Section 5 draws conclusions and points out future work.

2 Related Works

During recent years quite a number of different systems have been developed to implement virtual data integration. Various canonical models and conceptual specification approaches were applied.

In 2000 the Agora [15] system based on Local-As-View approach was developed. In Agora the XML is used as a canonical model. The system is intended to integrate relational or XML sources. The queries are posed using XQuery language.

In [2] the Automed system is presented. This DIS is based on a hypergraph data model and implements Both-As-View approach. In addition to the models accepted by Agora as source data models, Automed can also accept flat files as a data sources for processing and accessing it using the AutoMed Intermediate Query Langauge (AIQL).

A virtual data integration approach is also used in [1], where the authors considers RDF-based Data Integration Framework, and in [22], where the authors show how the Ontology Web Language (OWL) can be used as a canonical data model.

In [3] a virtual data integration system is presented and used for problem solving in the field of astronomy. As a canonical model the SYNTHESIS [11] language that is a combined object and frame data model. The SYNTHESIS canonical data model was intended to unify wide spectrum of data models like XML, relational, RDF etc. For instance, in [21] the main principles of conceptual mapping of array data model (ADM) into the SYNTHESIS language are analyzed and illustrated and in [19] the mapping of the RDF language into the SYNTHESIS is considered.

In [14] the MetaMed system used for integration of medical data is proposed. The extracted metadata are stored using the RDF and are structured by OWL ontologies.

For posing queries the SPARQL language is used. The authors integrate clinical documents and laboratory results that presented in DASTA format and imaging examination as DICOM files.

In [17] the possibility of using of SQL++ language as a canonical model for the relational and JSON data sources is analyzed. The authors discuss the use of unifying language in FORWARD virtual database query processor. SQL, NoSQL, SQL-on-Hadoop and NewSQL databases are integrated.

In fact, none of known existing DISs support the integration of the whole range of the NoSQL resources. However, with Mejer's research in [16] we can suggest that LINQ can serve as a unifying data model for a wide range of NoSQL models. LINQ is a query expression language within the C# language, and there is an advantage of LINQ over the SQL model: each class in LINQ may either have scalar properties, or contain arbitrary values, including other rows (or nested collections), that are typical for NoSQL models. Meijer et al. demonstrate the principles of interpretation of LINQ queries using category theory. They correlate data structures of SQL and NoSQL models using a notion of *duality* from category theory to conclude that NoSQL is a *dual* to SQL, and NoSQL can be called *coSQL*.

3 LINQ as a Canonical Data Model: Validation Steps

Validation that LINQ can be used as a canonical data model consists of three steps:

1. selection of source data models to be unified in LINQ;
2. implementation of a DIS prototype based on the GAV approach with LINQ as the canonical data model;
3. evaluation of DIS prototype using a concrete use case.

For validation LINQ as a canonical data model there were chosen five heterogeneous data models: SQLite as a relational model and four NoSQL data models. It is worth to admit that from a set of various NoSQL model classes the four main classes with different storage structure and data manipulation facilities were chosen. These classes are key-value, column-oriented, document and graph data models:

- *Key-Value Model.* The simplest model in representation is a key-value model. The representation consists of composition of a *key* and a *value*. A key-value NoSQL implemented systems allow either simple data types (e.g., numerals and strings) or the use of lists and sets of values of simple types. The databases based on this model do not support complex queries to be performed on the data stored in database, but only the search keys. The relationships in terms of reference keys and provides no referential integrity constraint are not supported in the key-value model [7].
- *Column-oriented Model.* Since the column model is organized in terms of columns and rows, it can be called as a modification of relational model. The modification is that the tables in column database can contain not only scalar values, but also nested tables. In other words, the rows do not store a tuple, but a set of attributes of the same type, while the set of attributes of a column contains the information from a given instance. Such feature allows queries to be performed more efficiently, although when recovering a complete instance, it may become more costly [7].

- *Document Model.* The databases based on the document model store the grouped data entities in document as an object that are composed by keys and values. Mainly, the documents are usually serialized in JSON syntax. The keys are generated randomly by the database or manually at persistence time. The document database allows complex queries involving different collections of documents. In the model is necessary that document has a database reference to another database, but it does not guarantee referential constraint [7].
- *Graph Model.* In the graph model the data items are connected by relationships by means of a graph structure. The graph model consists of (i) nodes, that correspond to data instances; (ii) edges, that refer to maintained relationships among node instances; and (ii) properties that relate to data instances. Edges contain linked input and output nodes. That feature guarantee referential integrity by ensuring that input node always makes reference to the output node. The access keys to the nodes are automatically set by the system. However, it is possible to establish unique constraints for other node properties [7].

The following DBMSs are chosen to be integrated in this work: SQLite[1] database based on a relational model, Accumulo[2] as a key-value store, HBase[3] as a column store, MongoDB[4] database based on document model and Neo4j[5] graph database.

Mentioned databases are planned to be integrated within a DIS prototype. The prototype should contain three main components: query rewriting component, DBMS wrappers supporting selected data models, and data merging component. The whole architecture that is planned to be realized is shown in Fig. 1.

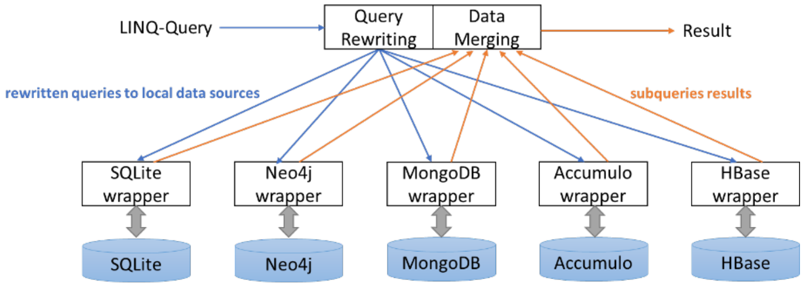


Fig. 1. The architecture to validate LINQ as a canonical model

Query rewriting is the first phase of DIS prototype operation. A query posed by a user should be rewritten into five subqueries [10]. After that rewritten subqueries are transformed into the source query languages. The wrappers for relational data models are integrated in some development environments, such as Microsoft Visual Studio or

[1] https://www.sqlite.org/index.html
[2] https://accumulo.apache.org/
[3] https://hbase.apache.org/
[4] https://www.mongodb.com/
[5] https://neo4j.com/

LinqPad[6], the default algorithm is LINQ-to-SQL[7] query transformation. The wrappers for chosen NoSQL models should be additionally developed. The development of these wrappers is considered as a future work. Transformed subqueries are passed to source DBMSs and results are directly extracted from the source databases. Since subqueries return results separately, the obtained data should be merged together in data merging component and then presented to the user.

4 Application of LINQ as a Canonical Data Model: a Use Case

Application and evaluation of LINQ as a canonical data model is illustrated in this section with a use case and consist of the following steps:

- a use case subject domain selection;
- definition of the global schema and a set of analytical queries that should illustrate the main LINQ constructs;
- selection of data sources to be deployed into source DBMSs;
- definition of conceptual mappings (based on the Global-as-View approach), that link the local sources’ schemas and the global schema;
- illustration of query rewriting and data merging processes.

Each step is described further in Sections 4.1–4.5.

Subject Domain Selection

As the domain for the use case the urban statistics were chosen including transports, demography, environment, immigration, etc. The examples of the subject domain entities and their attributes for transports and demography are illustrated in Table 1.

Various regional agencies or ministries can use the regional urban statistics to analyze the indicators and metrics. The indicators can be used for creating the development programs to improve the life quality in the region.

Global Schema and Analytical Query Example

After selection of the subject domain, the global schema for the domain should be defined. The global schema is used for aggregating all data obtained from heterogeneous data sources.

The global schema should be defined using the canonical model. Since the LINQ is a constituent of the C# standard, the entities (classes) and attributes should be defined using C#. An example for the definition of the classes *births* and *deaths* is shown in Table 2.

[6] https://www.linqpad.net/

[7] https://docs.microsoft.com/en-gb/dotnet/framework/data/adonet/sql/linq/

Table 1. Entities and attributes of the domain

Entities	Attributes
Demography	
births	year, district_code, district_name, neighborhood_code, neighborhood_name, gender, number
deaths	year, district_code, district_name, neighborhood_code, neighborhood_name, age, number
population	year, district_code, district_name, neighborhood_code, neighborhood_name, gender, age, number
unemployment	year, month, gender, demand_occupation, number
immigration_by_nationality	year, district_code, nationality, number
most_frequent_names	order, name, gender, decade, frequency
most_frequent_baby_names	order, name, gender, year, frequency
Transports	
accidents	id, district_name, neighborhood_name, street, weekday, month, day, hour, part_of_the_day, mild_injuries, serious_injuries, victims, vehicles_involved, longitude, latitude
bus_stops	code, transport, longitude, latitude, bus_stop, district_name, neighborhood_name
transports	code, transport, longitude, latitude, station, district_name, neighborhood_name

Table 2. The class declaration in the global schema

class births {
int year;
string district_name;
string neighborhood_name;
string gender;
int number;
string city;
}
	class deaths {
	int year;
	string district_name;
	string neighborhood_name;
	string age;
	int number;
	string city;
	}

In order to validate that LINQ can be used as a canonical data model, it is necessary to define a certain set of queries that covers the capabilities of the LINQ language. As an example, consider a LINQ query that contains filter (*where* clause), *join* operation, aggregation functions *sum*, sub-queries and mathematical operation in *select* clause:

```
var demogr =     from b in births
            join d in deaths on b.year equals d.year
            where b.year<=2017
            select new {y = b.year, brt = b.sum(n=>n.number),
                            dth = d.sum(n=>n.number),
                            diff = b.sum(n=>n.number) - d.sum(n=>n.number),
                            city = b.city};
```

This query should return the annual number of births and deaths from two separated classes (both classes should be joined) and the difference between births and deaths (rate of natural increase). The example uses *births* and *deaths* classes as data sources.

This paper considers only a fragment of a global schema that should be used in analytical queries. A definition of the whole global schema is a future work.

Data Sources

The raw data were extracted from the open data banks or previously developed databases. Data on events and statistics in Barcelona[8] are stored in SQLite relational database. In Neo4j (graph DBMS) a database with the events and statistics in Madrid[9] was created, the data on Bilbao[10] is going to be created in MongoDB (document DBMS). Data and statistics on Malaga[11] are going to be presented in HBase (column store), data on Seville[12] are going to be stored in Accumulo (key-value store).

As the query from Section 4.2 requires *births* and *deaths* entities, the parts of diagrams from relational and graph databases are illustrated in Fig. 2. The postfix *_rel* means that the entity is from relational database. Postfix *_gr* denotes entities from graph database.

On the left side of the Fig. 2 a part of the relational schema is presented: *births_rel* and *deaths_rel* entities. Each entity has a list of attributes that have explicit conducted data types. On the right side a visualization of the graph database is presented. The blue nodes are marked with *births_gr* label and orange nodes are marked with *deaths_gr* label. Since Neo4j does not show all labels of the nodes, there were shown only one label for each node. Other stored data of each label are hidden. In blue nodes the *gender* labels are shown, and in orange nodes the age category of deaths is pointed (the *age* label). Blue and orange nodes are linked by *brt_dth* relationships if *year*, *district_name* and *neighborhood_name* attribute values of *births_gr* and *deaths_gr* entities are equal.

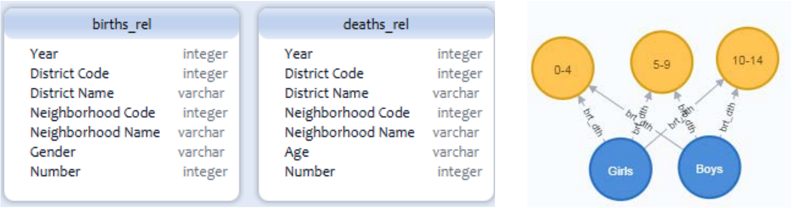
[8] https://opendata-ajuntament.barcelona.cat/en

[9] https://datos.madrid.es/portal/site/egob/

[10] https://www.bilbao.eus/opendata/es/formatos

[11] https://datosabiertos.malaga.eu/dataset

[12] http://sevilla-idesevilla.opendata.arcgis.com/datasets



births_rel	
Year	integer
District Code	integer
District Name	varchar
Neighborhood Code	integer
Neighborhood Name	varchar
Gender	varchar
Number	integer

deaths_rel	
Year	integer
District Code	integer
District Name	varchar
Neighborhood Code	integer
Neighborhood Name	varchar
Age	varchar
Number	integer

Fig. 2. The diagrams in relational (left) and graph (right) models for the example

The local schema definitions for document, column and key-value models are planned for the future work.

Conceptual Mappings

To define the conceptual mappings, the correspondences of entities and attributes between local and global schemas should be established first. The examples of correspondences between local (relational) and global schema elements are illustrated in Fig. 3.

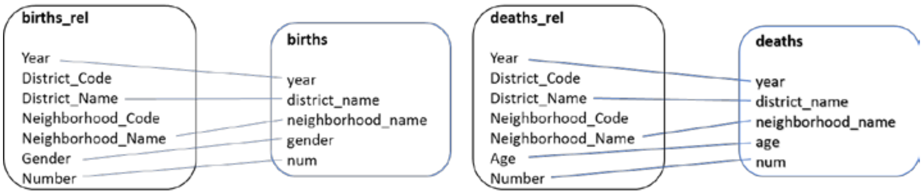


Fig. 3. Correspondences of the relational database (black) and global schema (blue)

According to GAV approach [14], the global schema entities should be represented as views over the local schema using the canonical data model. The correspondences can be produced either manually by an expert or automatically based on various approaches: machine learning approach [5], linguistic processing [26] etc.

For example, the conceptual mappings between *births* and *deaths* entities in global schema and *births_rel* and *deaths_rel* entities in local schema (SQL model) can be defined in the following way:

```
var births =   from b in births_rel
                 select new {year = b.Year, district_name = b.District_Name,
                    neighborhood_name = b.Neighborhood_Name,
                    gender = b.Gender, num = b.Number, city = "Barcelona"};

var deaths =   from d in deaths_rel
                 select new {year = d.Year, district_name = d.District_Name,
                    neighborhood_name = d.Neighborhood_Name,
                    age = d.Age, num = d.Number, city = "Barcelona"};
```

City is not the attribute of any local source and its value is generated within views. The value of *city* attribute is a constant that identifies the data source – the city to which the extracted data refer. As it is described earlier, the database based on relational model stores data on Barcelona.

Conceptual mappings for the graph database look almost the same, the only difference is that *births_gr* and *deaths_gr* entities are used instead of *births_rel* and *deaths_rel*.

Conceptual mappings can be constructed manually by an expert or in semi-automated way. This paper discusses a simple example in which conflict situations do not arise during construction of conceptual mappings. In general, various types of conflicts such as data type mismatch or structural conflicts requiring combination of several attributes of a local schemas into an attribute of the global schema can occur. Dealing with conceptual mappings in case of conflicts is a future work.

Query Rewriting, Transformation and Data Merging

Query Rewriting. First, the query from Section 4.2 should be rewritten. The query rewriting is processed using the views from Section 4.4 on the basis the GAV approach [14]. For instance, the rewritten query to the relational database is formed by replacing *births* and *deaths* by the bodies of views from Section 4.2 as nested queries:

```
var demogr =    from b in (from b in births_rel
                          select new {year = b.Year, district_name = b.District_Name,
                             neighborhood_name = b.Neighborhood_Name,
                             gender = b.Gender, num = b.Number, city = "Barcelona"})
                join d in (from d in deaths_rel
                          select new {year = d.Year, district_name = d.District_Name,
                             neighborhood_name = d.Neighborhood_Name,
                             age = d.Age, num = d.Number, city = "Barcelona"})
                on b.year equals d.year
                where b.year<=2017
                select new {y = b.year, brt = b.sum(n=>n.num),
                             dth = d.sum(n=>n.num),
                             diff = b.sum(n=>n.num)- d.sum(n=>n.num),
                             city = b.city};
```

The rewritten query for the graph database looks almost the same. The only difference is that in the rewritten query for the graph database that the value of the *city* attribute is *"Madrid"* instead of *"Barcelona"*.

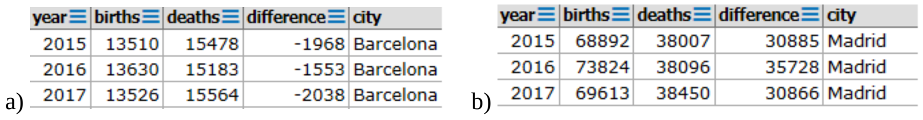
Query Transformation. To extract the data from local sources rewritten queries should be transformed into the source query languages. For instance, the rewritten LINQ query should be transformed in relational data model wrapper as follows:

```
SELECT b.year as y, sum(b.num) as brt, sum(d.num) as dth,
    sum(b.num) - sum(d.num) as diff, b.city
FROM
    (SELECT b.Year as year, b.District_Name as district_name, b.Neighborhood_Name
as neighborhood_name, b.Gender as gender, b.Number as num, “Barcelona” as city
FROM births_rel b) b
    INNER JOIN
    (SELECT d.Year as year, d.District_Name as district_name, d.Neighborhood_Name
as neighborhood_name, d.Age as age, d.Number as num, “Barcelona” as city
    FROM deaths_rel d) d ON b.year = d.year
    WHERE b.year <=2017
    GROUP BY b.year, b.city
```

Considering that the attribute value *city* = *“Madrid”* in the graph data source, the rewritten LINQ query is transformed into Cypher graph query language (supported by Neo4j) as follows:

```
    MATCH (b:births_gr)
    WITH [{year: b.Year, district_name: b.District_Name, neighborhood_name:
b.Neighborhood_Name, gender: b.Gender, num: b.Number, city: "Madrid"}] as b
    UNWIND b as b
    MATCH (d:deaths_gr)
    WITH [{year: d.Year, district_name: d.District_Name, neighborhood_name:
d.Neighborhood_Name, age: d.Age, num: d.Number, city: "Madrid"}] as d
    UNWIND d as d
    MATCH (b)--(d)
    RETURN b.year as y, sum(b.num) as brt, sum(d.num) as dth, sum(b.num)-
sum(d.num) as diff, b.city
```

Data Merging. Finally, the results extracted from sources should be merged in data merging component. For instance, the query to the relational database returns the result that is shown on Fig. 4(a), the query to the graph database returns the result that is shown on Fig. 4 (b). The result presented to a user in this case is just a merge of 4(a) and 4(b) results.



a)

year	births	deaths	difference	city
2015	13510	15478	-1968	Barcelona
2016	13630	15183	-1553	Barcelona
2017	13526	15564	-2038	Barcelona

b)

year	births	deaths	difference	city
2015	68892	38007	30885	Madrid
2016	73824	38096	35728	Madrid
2017	69613	38450	30866	Madrid

Fig. 4. The query results from relational (a) and graph (b) databases

5 Conclusions and Future Work

This paper presents an approach for validation that Language Integrated Query (LINQ) is able to serve as a canonical data model for virtual data integration in the world of SQL and NoSQL databases. The architecture of DIS to validate LINQ as a canonical model is presented. A use case intended to illustrate key steps of data integration is provided. The subject domain of the use case is defined, conceptual mappings between source schemas and the global schema are illustrated as well as query rewriting and transformation for the relational and Neo4j data models.

Future work includes development of wrappers for chosen NoSQL models as well as query rewriting and data merging components of the prototype DIS. Complete use case description including global and local schemas and their conceptual mappings accompanied by a set of comprehensive analytical queries and their results is also a future work.

Acknowledgements

This work is supervised by Sergey Stupnikov, lead research scientist at the Federal Research Center "Computer Science and Control" of Russian Academy of Sciences.

References

1. Amini, A., Saboohi, H., and Nematbakhsh, N.: An RDF-based data integration framework. In: National Electrical Engineering Conference (NEEC) 2008, Najafabad (2008).
2. Boyd, M., Kittivoravitkul, S., Lazanitis, C., McBrien, P., and Rizopoulos, N.: AutoMed: a BAV data integration system for heterogeneous data sources. In: Persson A., Stirna J. (eds) Advanced Information Systems Engineering. CAiSE 2004. Lecture Notes in Computer Science **3084**. Springer, Berlin, Heidelberg (2004).
3. Briukhov, D.O., Vovchenko, A. E., Zakharov, V.N., Zhelenkova, O.P., Kalinichenko, L.A., Martynov, D.O., Skvortsov, N.A., and Stupnikov, S.A.: The middleware architecture of the subject mediators for problem solving over a set of integrated heterogeneous distributed information resources in the hybrid grid-infrastucture of virtual observatories. Informatics and Applications **2** (1), 2–34 (2008).
4. Briukhov, D., Kalinichenko, L., and Martynov, D.: Source registration and query rewriting applying LAV/GLAV techniques in a typed subject mediator. In: Proceedings of the 9th Russian Conference on Digital Libraries, RCDL'2007, 253–262. Pereslavl, Russia (2007).
5. Bulygin, L.: Combining lexical and semantic similarity measures with machine learning approach for ontology and schema matching problem. In: Proceedings of the XX International Conference "Data Analytics and Management in Data Intensive Domains" (DAMDID/RCDL'2018), 245–249, Moscow (2018).
6. European Computer Machinery Association. Standard ECMA-334: C# Language Specification, 5th edition, December 2017. https://www.ecma-international.org/ publications/files/ECMA-ST/ECMA-334.pdf
7. Freitas, M., Souza, D., and Salgado, A.: conceptual mappings to convert relational into NoSQL Databases. In: Hammoudi, S., Maciaszek, L.A., Missikoff, M.M., Camp, O., Cordeiro, J. (eds.) 18th International Conference on Enterprise Information Systems (ICEIS 2016), **1**, 174–181. SCITEPRESS, Rome (2017).
8. Friedman, M., Levy, A., and Millstein, T.D.: Navigational plans for data integration. In: Proceedings of the National Conference on Artificial Intelligence (AAAI), 67–73. AAAI Press/The MIT Press (1999).
9. Haas, L.M., Lin, E.T., and Roth, M.A.: Data integration through database federation. IBM Systems Journal **41** (4), 578–596 (2002).

10. Hai, R., Quix, and C., Zhou, C.: Query rewriting for heterogeneous data lakes. In: Benczúr A., Thalheim B., Horváth T. (eds) Advances in Databases and Information Systems. DBIS 2018. Lecture Notes in Computer Science **11019,** 35–49. Springer, Cham (2018).
11. Kalinichenko, L.A., Stupnikov, S.A., and Martynov, D.O.: SYNTHESIS: A language for canonical information modeling and mediator definition for problem solving in heterogeneous information resource environments. IPI RAN, Moscow (2007).
12. Khine, P. and Wang, Z.: Data lake: a new ideology in big data era. In: Guchi, K., Chen, T. (eds.) 2017 4th Annual International Conference on Wireless Communication and Sensor Network (WCSN 2017) **17**, Wuhan (2017).
13. Kimball, R. and Ross, M.: The Data Warehouse Toolkit. 3nd edn. John Wiley & Sons, Inc., Indianapolis, IN (2013).
14. Lenzerini, M.: Data integration: a theoretical perspective. In: Proceedings of the 21st ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, 233–246. Madison, Wisconsin, USA (2002).
15. Manolescu, I., Florescu, D., Kossmann, D., Xhumari, F., and Olteanu, D.: Agora: living with XML and relational. In: El Abbadi, A., Brodie, M., Chakravarthy, S., Dayal, U., Kamel, N., Schlageter, G., Whang, K.-Y. (eds.): Proceedings of 26th International Conference on Very Large Data Bases (VLDB 2000), 623–626. Cairo (2000).
16. Meijer, E., Bierman, G.: A co-Relational Model of Data for Large Shared Data Banks. Microsoft Research. ACMqueue **3** (9), 1–19 (2011).
17. Ong, K. W., Papakonstantinou, Y., Vernoux, R.: The SQL++ Unifying Semi-structured Query Language, and an Expressiveness Benchmark of SQL-on-Hadoop, NoSQL and NewSQL Databases. CoRR, abs/1405.3631 (2014).
18. Rodríguez-Muro, M., Kontchakov, R., and Zakharyaschev, M.: Ontology-Based Data Access: Ontop of Databases. In: Alani H. et al. (eds): The Semantic Web – ISWC 2013. ISWC 2013. Lecture Notes in Computer Science **8218**, 558–573. Springer, Berlin, Heidelberg (2013).
19. Skvortsov, N.A.: Mapping of RDF Data Model into the Canonical Model of Subject Mediators. In: Smirnov, V., Stupnikov, S. Proceedings of the 15th All-Russian Scientific Conference “Digital Libraries: Advanced Methods and Technologies, Digital Collections” (RCDL 2013), 95–101 (2013).
20. Stupnikov, S. and Kalinichenko, L.: Extensible Unifying Data Model Design for Data Integration in FAIR Data Infrastructures. In: Manolopoulos Y., Stupnikov S. (eds): Data Analytics and Management in Data Intensive Domains (DAMDID/RCDL 2018). Communications in Computer and Information Science **1003**, 17–36. Springer, Cham (2019).
21. Stupnikov, S.: Unification of an array data model for the integration of heterogeneous information resources. In: Znamenskij, S., Kogalovsky, M. Proceedings of the 14th All-Russian Scientific Conference "Digital libraries: Advanced Methods and Technologies, Digital Collections" (RCDL 2012), 42–52. Pereslavl-Zalessky (2012).
22. Tan, P., Madnick, S., and Tan, K.: Context mediation in the Semantic Web: handling OWL Ontology and Data Disparity Through Context Interchange. In: Bussler, C., Tannen, V., Fundulaki, I. 2nd international workshop on semantic web and databases 2004, LNCS **3372**, 140–154. Springer, Berlin, Heidelberg (2005).
23. Theodoratos, D.: Semantic integration and querying of heterogeneous data sources using a hypergraph data model. In: 19th British National Conference on Databases (BNCOD), 166–182. Springer, Heidelberg (2002).
24. Vcelak, P., Kratochvil, M., Kleckova, J., and Rohan, V.: MetaMed – Medical meta data extraction and manipulation tool used in the semantically interoperable research information

system. In: Chen, Q., Huan, J., Xu, Y., Zhang, T., Wang, L. (eds.) 5th International Conference on Biomedical Engineering and Informatics (BMEI 2012), pp. 1270–1274. IEEE, Chongqing, China (2012).

25. Zakharov, V.N., Kalinichenko, L.A., Sokolov, I.A., and Stupnikov, S.A.: Development of canonical information models for integrated information systems. Informatics and Its Applications **1** (2), 15–38 (2007).
26. Zhang, Y., Wang, X., Lai, S., He, S., Liu, K., Zhao, J., and Lv, X.: Ontology matching with word embeddings. In: Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data, 34–45 (2014).

Classification of Human Actions Using Task fMRI Images

Dmitrii Sergeev[1]

[1]Moscow State University, Moscow, Russia
SerDimIgor@gmail.com

Abstract. In the past few years, the topic of brain signal analysis has become very popular in neuroscience. There are several approaches to imaging the brain. Classification of human activities with task fMRI is an important part of finding effective connectivity in human brain. This article is devoted to developing of an approach to constructing a classifier for human actions. Detailed definition of the basic notions used in analyzing fMRI images is provided. A review of datasets and methods for classifying fMRI images is presented with recommendations. Also, brief description of major international projects involved in brain analysis is provided. In conclusion, workflow and way forward to implementations is examined with description of proposed libraries to use for analysis, filtering, preprocessing, reading and writing fMRI and fitting classification models with it.

Keywords: task fMRI analysis, data intensive analysis, human action classification.

1 Introduction

Neuroimaging is the common name for several methods that allow visualization of the structure, functions, and biochemical characteristics of the brain [1]. At the same time, neuroimaging techniques do not require surgical intervention and direct contact with the internal organs, since these technologies made it possible to non-invasive visualization of the structure and functionality of the brain, becoming a powerful tool for research and for medical diagnostics with the development of technology and computational methods.

Functional neuroimaging is used to measure aspects of the brain to understand the relationship between the activity of certain areas of the brain with specific mental functions.

There are several approaches to collect data about human brain for latter analysis:

- Computed tomography (uses a series of x-rays aimed at the head from a large number of different directions);
- Diffuse optical tomography (uses infrared radiation, measures the optical absorption of hemoglobin);
- Optical Signal modified by an event (using infrared radiation);
- Electroencephalography (EEG);
- magnetic resonance imaging (MRI) (uses magnetic fields and radio waves without using ionizing radiation);

- functional magnetic resonance imaging (fMRI).

Most brain analysis nowadays is implemented on the basis of fMRI images. Functional magnetic resonance imaging is based on the paramagnetic properties of hemoglobin and makes it possible to see changes in the blood circulation of the brain depending on its activity [2]. The essence of the method is that when certain parts of the brain work, the blood flow in them increases. Changes in blood flow are recorded, and images can tell which parts of the brain are activated when performing certain actions. fMRI image is a 4-dimensional array of voxels (spatial and time). This type of images allows to analyze the activity of various parts of the brain at some time point.

Over the past decades, researchers have managed to accumulate a large amount of fMRI data. fMRI images have a complex descriptive structure and require large resources for their storage, such as high-performance computing systems. Besides that, the amount of data surpasses tens of terabytes of data, requiring special compute-intensive platforms to deal with these datasets. These facts underline the multidisciplinary nature of neuroscience and the need to develop IT methods for it.

There are several types of relationships in the brain – structural, functional and effective connectivity [3]. Effective connectivity, which describes the amount of information transmitted by information flows in the presence of any stimulus or the absence of incentives per unit of time, is among the most interest in analyzing brain images [3].

Classification of human actions using task fMRI images is an important part of analyzing effective connectivity. As an example, in Human Connectome Project participants were asked to perform seven tasks related to the following categories: Emotion, Gambling, Language, Motor, Relational, Social and Working Memory. Based on the task fMRI data obtained [4], a mathematical model is built relating images to a specific task. However, the model lacks accuracy and does not deal with data-intensive platforms.

This work is aimed at development of methods and tools to process large datasets in neurophysiology domain, build classification model to analyze effective connectivity of the brain using task fMRI images. The research is carried out as thesis for Masters Program “Big data: infrastructures and problem-solving techniques” under the department of Computational Mathematics and Cybernetics of Moscow State University.

2 Problem Statement and Formalization of Application Domain

Specification of the application domain is depicted on Fig. 1. Effective connectivity describes the causal interaction between units of connectivity (usually brain regions). It is described by the amount of information transmitted over information flows in the presence of any stimulus or absence of incentives per unit of time. The connected unit (region of interest) transmits information signal to another connected unit by information flow, receives information from another unit by information flow.

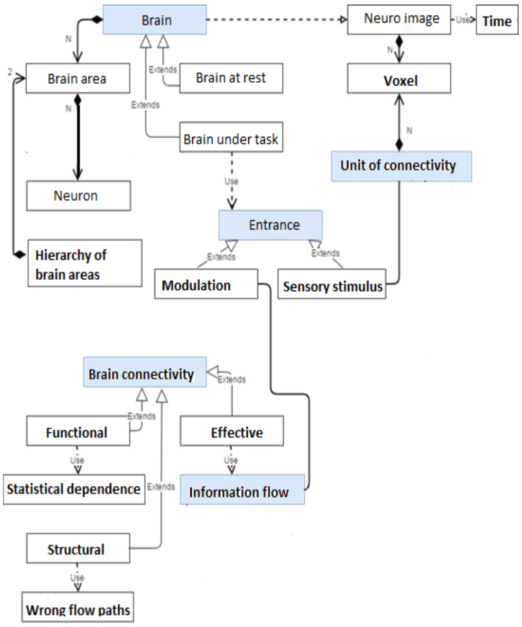


Fig. 1. Specification of neurophysiology application domain

A problem statement is formulated as follows: present an approach for dealing with large incoming datasets of fMRI, preprocess them, build classification model for preprocessed dataset and validate it on some data. Classification task formulates as follows: using task fMRI images relate them into seven groups of task, which a person was doing during that session.

3 Related Works

Classification Methods

One of the pioneering works on the classification of signals of the human brain is based on testing statistical hypotheses [5]. Using the t-test, the signal is classified into two classes. There are actual problems of binary classification, for example, to distinguish Alzheimer's patients from healthy people, solved with the help of t-test [6, 7]. The t-criterion has a number of advantages and disadvantages. The advantages include: ease of calculation; ease of interpretation; resistant to emissions; works with even a small amount of data. The disadvantages of this method include: assumptions that the data have a normal distribution; residues are independent and have a normal distribution.

Later works are based on linear classification methods such as: support vector machine (SVM [8, 9]), general linear models (GLM), etc. Such classifiers are suitable for solving problem of binary classification. However, the use of linear classification methods imposes significant restrictions on the dataset: it must be linearly separable. In [5], researchers analyze the brain signal using the support vector machine (SVM [8, 9]). The advantage of this approach is that linear models are easily interpreted, are trained with small samples, and are not prone to overfitting. Linear models have several disadvantages: they do not approximate complex surfaces; dataset must be independent.

Works on analyzing MRI images based on using neural networks [10–12] began to appear relatively recently. The main advantages are that neural networks allow to build complex separating surfaces, significantly increasing the quality of the model. The second essential advantage of neural networks is that it allows to implement multi-class classification without significantly complicating the model. Disadvantages are that models are prone to overfitting and require vast computational resources. Also, sometimes the dimension of the learning model is too large.

In [4] researchers analyze the signal from the brain using deep neural networks. Multiple architectures are built with two and more hidden layers and the quality of work of different architectures is compared. Experiments are performed on a dataset from the Human Connectome project. In [13], the authors take ready-made convolutional neural net (CNN (LeNet-5)) and successfully classify functional MRI data of Alzheimer's subjects. Accuracy on test dataset reached 96.85%. Usage of CNN allows to extract useful tags from images and approximate complex structures. Recent studies show that modern architects of convolutional neural networks classify the image more qualitatively than humans.

Based on the result of the study, it is recommended to use CNN.

Related Neurophysiology Projects and Dataset Description

Human Connectome Project. The main goal of the Human Connectome Project (HCP) [14] is to describe the structural and functional connection in the brain of a healthy young adult. Based on the data collected by HCP, a large number of studies are annually conducted to extract functional and structural dependencies between different parts of the brain. The database contains information about more than 1200 participants. The Human Connectome Project brings together a large group of researchers all around working in the field of neuroscience. All data can be downloaded from the project website for free after registration.

Human Connectome Project has a separate task fMRI dataset published. Each participant during the fMRI session was asked to perform some tasks from the following groups:

- Emotion processing: participants were asked to map several images to each other.
- Gambling: participants were asked to play a simple card game.
- Language processing: after listening to a short audio file, participants were asked to answer simple questions.
- Motor: participants were asked to move the divided body part.
- Social cognition: short video was presented to the participants and asked to answer the question whether the movements of objects in the clips are related to each other in some way.

Each subject has corresponding behavioral data, including age, weight, etc. Also, each task fMRI has a corresponding design file, which stores meta information about the experiment, the number and name of the experimental conditions recorded, and an indication of the path to the timing files. Timing files contain the time of the stimulus (onset) and its duration (duration). In addition, functional data, masks, files with the time of appearance of stimuli and their duration, files with data about experiments (design files) are stored. The dimension of one image is (91, 109, 91, 274), i.e. total 902629 voxels with 274 values.

Other Projects. The Human Brain Project (HBP) is a large research project to study the structure and analyze the functional connectivity between different parts of the brain. The project involves hundreds of scientists from 26 countries and 135 partner institutions. The goal of this project is to create a joint research infrastructure to enable researchers around the world to develop knowledge in the field of neurobiology, computer technology, and medicine related to the brain.

The BRAIN Initiative project was created at the initiative of the White House in 2013. This project was created as a private-public research initiative. A large number of neuroscientists from 30 different countries are involved in this project. At the first stages of the project, researchers will analyze the activity of neurons in mice and other animals, and at later stages of the project a functional map of dependencies of various parts human brain will be built. It is assumed that these studies will help researchers discover the secrets of brain disorders such as Alzheimer's and Parkinson's, depression.

1000 Functional Connectomes Project is a database of functional MRI images taken at rest. The purpose of this project is to collect fMRI images during rest. When visualizing the brain during rest, random low-frequency oscillations of large amplitude occur, which correlate in different functionally related areas. Based on the data obtained, researchers build maps of interaction between different areas of the brain. The database contains data from more than 1,400 participants collected independently in 35 international centers

4 Proposed Approach

Workflow

Workflow for proposed approach is depicted on Fig. 2. First, input data is preprocessed using NIPY [15] library. Next, regressors and contrasts (artifacts for handling task fMRI) are constructed. Next, using these artifacts and preprocessed data, the classification model is built. There are several libraries to work with CNN. Later, model is validation against testing dataset.

It is assumed that preprocessing step and construction of contrasts and regressors are built with PySpark [16] library in distributed manner.

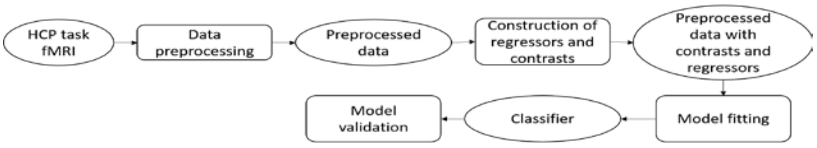

HCP task fMRI
Data preprocessing
Preprocessed data
Construction of regressors and contrasts
Preprocessed data with contrasts and regressors
Model fitting
Classifier
Model validation

Fig. 9. Workflow for constructing classifiers for human activity

Causality Model as Input for Classification

One of the main ideas of the approach, which differs it from other, is to use output of Dynamic Causal Modeling (DCM) [17] as the input to classification model. Dynamic causal modeling (DCM) is a general Bayesian structure for drawing conclusions about hidden neural states based on measurements of brain activity. DCM provides a posteriori estimates of neurobiological interpreted values, such as the effective strength of synaptic connections between neuronal populations and their context-dependent modulation (i.e., how experiment factors influence these values). In other words, with the help of DCM, it can be understood how a specific change in conditions during an experiment affects the activation of brain area.

DCM is stated as (linear or non-linear) differential equations. They describe hidden dynamics of neural populations. DCM models seek to ensure being neurophysiological interpreted.

The idea of using DCM as input for further classification is following. DCM is not a theoretical simulation of neuronal processes in its pure form, but a method that includes both a theoretical calculation (model prediction) and a validation on real data (implemented using Bayesian inversion). The key feature of the DCM method is its dependence on experimental data. Its equations take into account the influence of experimental manipulations on the dynamics of the system: the experimental conditions are included in the model as input data that either controls the local responses of the system or changes the connections. So, e.g., if a person was watching a video during fMRI, DCM will tend to seek the activation of the brain area responsible for visual cognition. This knowledge cam vastly increase classification accuracy.

Data Neuroimaging Format

Task fMRI neuroimage is a four-dimensional array of voxels, three dimensions describe the position of voxels in space, the fourth – in time. A voxel has an index in the three-dimensional spatial array of the fMRI neuro-image and has n values for each t (t=1...m) of the fourth dimension of the fMRI neuro-image (i.e., is a time series).

NIFTI [18] allows to store data in several ways: (1) as in ANALYZE in 2 files (1 file is a header file with the extension. hdr; 2 file – the data itself in .img format); (2) or all in one file with the extension. nii. NIFTI also supports working with compressed data (.gz). The first 4 measurements out of 7 are predefined to represent spatial and temporal coordinates (1–3 spatial, 4 temporal, 5–7 adjustable).

Header structure has size of 348 bytes. Some header fields are:

- Information about data collection (Dim info): char dim_info – stores the directions of frequency, phase coding, the direction in which the volume increased when receiving data;
- Image dimensions: short dim [8] contains information about image dimensions. dim [i] represents the length of the i-th dimension;
- Intent-fields: short intent_code is a code showing the statistical nature of the data, some codes require additional parameters, which are either indicated in the float intent_p * fields (if applicable to the picture as a whole), or form the 5th dimension (if these parameters are different for each voxel). The readable intent name can be stored in the char intent_name [16] field;
- Data type: int datatype shows the type of stored data; short bitpix contains information on the number of bits per voxel;

- Slice acquisition: char slice_code, short slice_start, short slice_end and float slice_duration store information about the fmri time distribution, and should be used together with char dim_info containing fieldslice_dim. The short slice_start and short slice_end fields indicate which layer is the first and last for a particular mri. Layers out of range are considered added to the file (and not received with mri, usually contain 0). The float slice_duration field indicates the amount of time needed to produce a single layer;
- Voxel dimensions: The float pixdim [8] contains the dimension of each voxel, by analogy with short dim [8]. But the values in float pixdim [0] must be equal to –1 or 1;
- Voxel Offset (Voxel o set): The int vox_o ff set field indicates the beginning of the data itself from the beginning of the file (for data contained in .nii file), for a pair of files (.hdr / .img), the field must contain 0 if there is no additional data other than the picture in .img not contained;
- Data scaling: The values stored in each pixel can be linearly scaled in different units. (fields float scl_slope and float scl_inter);
- Display range (Data display): For files that store scalar data, the cal_min and cal_max fields define the intended display range when the image is opened;
- Measurement units: Both temporal and spatial units used in dim [i] (i=1..4) (and for pixdim) are stored in the char xyzt_units field. Bits 1-3 are used for spatial measurements, 4–6 – for temporary, 7–8 – are not used;
- Image orientation information (Orientation information): In NIFTI, it is possible to uniquely store orientation information. The file standard assumes that the voxel coordinates correspond to the center of this voxel. It is assumed that the system of world coordinates is "RAS +". The format represents 3 different methods of mapping voxel coordinates (i, j, k) to world (x, y, z). The main one is that the world coordinates are determined by scaling the voxel size.

Image Analysis and Processing Tools

The NIPY library consists of several parts that enable the user to perform not only simple operations with fMRI images such as reads and writes, but also analysis algorithms. This library includes the following projects:

- Nipype provides a unified interface for working with fMRI images;
- NiBabel is module that allows you to work with a large number of medical and neurovizualizable file formats such as GIFTI, NIfTI1, NIfTI2, CIFTI-2, MINC1, MINC2, AFNI BRIK/HEAD, MGH and ECAT as well as Philips PAR/REC. This library allows you to both read and write the listed file formats. This library has a Python interface that makes it quite simple and easy to use. The library’s website provides detailed installation information and examples with explanations on the use of the library
- PyMVPA is a set of algorithms that are intended for statistical image analysis. In this package, implemented algorithms for classification, clustering and regression, created unified interfaces for interacting with standard data analysis libraries such as scikit-learn, shogun, MDP, etc

OpenCV [19–21] is first of all this computer vision library, there are several thousand high-performance image processing algorithms implemented in this library. This library is distributed under the BSD license, therefore the code of this library can be modified and used in commercial projects. The OpenCV library has a modular structure. Researchers use this library to pre-process MRI images and extract functions from MRI images.

Recently, a lot of articles appeared trying to classify fMRI images using convolutional neural networks. There are many different libraries and software products that implement neural network architectures. The Keras library is one of the most popular. This library is written in the Python programming language, with operations performed on TensorFlow. For a training of a convolutional neural network, a huge number of trained images are required. Keras contains within itself the architecture of popular convolutional neural networks, which were trained in ImageNet [22–24].

5 Conclusion

Classification problem is stated for task fMRI. Specification of application domain is provided. A review of datasets and methods for classifying fMRI images is presented with recommendations. Also, brief description of major international projects involved in brain analysis is provided. In conclusion, workflow and wayforward to implementations is examined with description of proposed libraries to use for analysis, filtering, preprocessing, reading and writing fMRI and fitting classification models with it.

Acknowledgements

This work is supervised by Dmitry Kovalev, Federal Research Center “Informatics and Control” of Russian Academy of Sciences.

References

1. Duncan, J.: Neuroimaging methods to evaluate the etiology and consequences of epilepsy. Epilepsy Research, 131–140 (2002).
2. Shulman, R.G., Rothman, D.L., Behar, K.L., and Hyder, F.: Energetic basis of brain activity implications for neuroimaging. Trends Neurosci, 489–495 (2004).
3. Schlösser, R., Gesierich, T., Kaufmann B., Vucurevic G., Hunsche S., and Gawehn, J.: Altered effective connectivity during working memory performance in schizophrenia: a study with fMRI and structural equation modeling. NeuroImage, 751–763 (2003).
4. Koyamada, S., Shikauchia, Y., Nakaea, K., Koyamaa, M., and Ishiia, S.: Deep learning of fMRI big data: a novel approach to subject-transfer decoding. Stat.ML, arXiv:1502.00093v1 (2015).
5. Mahmoudi, A., Takerkart, S., Regragui, F., Boussaoud, D., and Brovelli, A.: Multivoxel Pattern Analysis for fMRI Data: A Review. Computational and Mathematical Methods in Medicine (2012).
6. Friston, K., Frith, C., and Liddle, P.: Comparing functional (PET) images: the assessment of significant change. Journal of Cerebral Blood Flow and Metabolism, 690–699 (1991).
7. Hall, D. and Miller, H.: The Theory of Stochastic Processes. 1nd Edition. Routledge (1977).
8. Cortes, C. and Vapni, V.: Support-vector networks. Machine Learning, 273–297 (1995).
9. Vapnik, V.: The Nature of Statistical Learning. 2nd Edition. Springer (1995).

10. Karnowski, T.: Deep machine learning a new frontier in artificial intelligence research. Computational Intelligence Magazine, 13–18 (2010).
11. Grady, C., Sarraf, S., and Saverino, C.: Age differences in the functional interactions among the default, frontoparietal control and dorsal attention networks. Neurobiology of Aging (2016).
12. Shelhamer, E., Donahue, J., Karayev, S., Long J., and Girshick, R.: Convolutional architecture for fast feature embedding. Proceedings of the ACM International Conference on Multimedia, 675–678 (2014).
13. Sarraf, S. and Tofighi, G.: Classification of Alzheimer's disease using fMRI data and deep learning convolutional neural networks. cs.CV, arXiv:1603.08631v1 (2016).
14. Barch, D.M., Burgess, G.C., Harms, M.P., Petersen, S.E., Schlaggar, B.L., and Corbetta, M.: Function in the human connectome: task-fMRI and individual differences in behavior. Neuroimage, 169–189 (2013).
15. NIPY – neuroimaging software. https://nipy.org/
16. Drabas, T. and Lee, D.: Learning PySpark. 1nd Edition. Packt Publishing (2017).
17. Friston, K.: Dynamic causal modeling and Granger causality comments on: The identification of interacting networks in the brain using fMRI: Model selection, causality and deconvolution. Neuroimage **58**, 303–305. (2011).
18. Gorgolewski, K., Auer, T., Calhoun, V., Craddock, C., Das, S., and Duff, E.: The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. Scientific Data (2016).
19. OpenCv – Open Source Computer Vision Library. https://opencv.org/
20. Bradski, G. and Kaehler, A.: Learning OpenCV: Computer Vision with the OpenCV Library. 1nd Edition. O'Reilly Media (2008).
21. Garrido, G. and Joshi, P.: OpenCV 3.x with Python by example: make the most of OpenCV and Python to build applications for object recognition and augmented reality. 2nd Edition. Packt Publishing (2018).
22. Keras – open-source neural-network library. https://keras.io/
23. Gulli, A. and Pal, S.: Deep Learning with Keras: Implementing deep learning models and neural networks with the power of Python. 1nd Edition. Packt Publishing (2017).
24. Williams, A.: Deep Learning with Keras: Introduction to Deep Learning with Keras. 2nd Edition. CreateSpace Independent Publishing Platform (2017).

Comparison of Male and Female Nonlinear Brain Functional Connectivity

Egor Tirikov [1]

[1] Moscow State University, Moscow, Russia
em.tirikov@gmail.com

Abstract. In this paper, linear models, genetic programming and multilayer perceptron were considered for studying the nonlinear functional connectivity of the brain. The study of functional connectivity is important, since the results obtained can later be used to study such diseases as Parkinson's or Alzheimer's disease. The advantages and disadvantages of the considered methods were described, as well as further research plans, where gender differences in fMRI data will be explored. Also, preliminary results was provided, which demonstrate nonlinear relationship between brain regions.

Keywords: resting-state fMRI, nonlinear functional connectivity, data intensive analysis

1 Introduction

Today, in many fields of science it is necessary to process large amounts of semi-structured data. Neuroinformatics, which lies in the intersection of neurophysiology and informatics, is a cross-disciplinary domain of science that studies methods and tools for analyzing human brain activity and interaction. It is a well-known data-intensive domain of science. The amount of collected data in neuroinformatics is estimated at order of petabytes [1]. Therefore, complexity of using conventional approach to analysis, methods and processing tools is high and different specialized solutions have to be specifically designed for processing such large datasets. Furthermore, not only volume, but also different types, forms and formats of datasets pose a problem. As an example, electroencephalography (EEG), magnetic electroencephalography (MEG) and functional magnetic resonance imaging (fMRI) are all different brain signal techniques used to analyze brain activities [2].

There are three types of brain region interaction: functional connectivity, structural connectivity and effective connectivity [3, 4]. The study of functional connectivity is of great importance, as the obtained results are used to study Parkinson's disease [5], attention-deficit/hyperactivity disorder [6], Alzheimer's disease [7], etc. For example, in [5] it is stated that advanced Parkinson's disease reduces functional connectivity between the brain regions. Knowing exactly what changes occur in the brain during Parkinson's disease helps to better diagnose on early stages and apply appropriate treatment. fMRI measures brain activity by detecting changes in blood flow. There are two types of fMRI: task-fMRI and resting-state fMRI [8]. Primarily, resting-state fMRI data is used to analyze functional connectivity [9]. Resting-state fMRI is collected for patients at rest; usually patient is asked to close his/her eyes and not to focus on anything specific.

There are two types of functional connectivity: linear and nonlinear. In most cases researches study linear functional connectivity. Linear functional connectivity implies, that target brain region depends from others linearly, i.e.

$$y = w_1 x_1 + \cdots + w_n x_n,$$

where y is a value of target brain region, $x_1, \dots, x_n$ are values of other brain regions. Otherwise, it is non-linear. Though simple and useful in some studies [10, 11], linear model does not always correspond correctly to measurements. In [12, 13] it is shown that functional connectivity has nonlinear dependence between brain regions, showing that the problem of studying nonlinear functional connectivity is of relevance.

It is known [14–16] that functional connectivity differs for men and women. These articles are focused on linear functional connectivity, though important, it does not provide any details if more complex dependencies in the brain also differ for men and women. It becomes possible to make a more subtle diagnosis for both of these two groups, to diagnose diseases at earlier stages and to develop a more suitable treatment if it is known that such difference exists and is meaningful.

This article is devoted to developing an approach of constructing nonlinear functional connectivity in terms of analytical equations to study brain activity difference for men and women. The article is structured as follows: section 2 introduces formalization of the application domain. Section 3 overviews methods, which are used to compute nonlinear functional connectivity and presents recommendations. Section 4 describes workflow, libraries and touches some implementation issues. Section 5 concludes the article.

2 Related Works

Available Datasets

There are multiple datasets in neuroinformatics, among them there are 1000 Functional connectivity project (FCP) [17], Human connectome project (HCP) [1] and Human Brain Project (HBP) [18].

The organizers of FCP collected 1200 data sets of resting-state fMRI from 33 independent sources. For each dataset, information about age, sex of subjects and image processing center is provided. There is a huge difference between age groups, number of samples, frequencies and slices for these datasets.

HCP is a project that was launched in 2009. There are three directions in HCP Project: HCP young adult 1200, HCP lifespan Studies and Connectomes Related to Disease Studies. First project studies brain connectivity within healthy brain of young adult. Second project studies difference in brain connectivity between different age groups. Last project studies difference in brain connectivity between healthy and diseased brain.

The goal of the HCP project is to build a network map that is supposed to explain the anatomical and functional connections inside the brain of a healthy person.

HBP project began in 2013 and is designed for 10 years. There are six research platforms: Neuroinformatics, Brain Simulation, High Performance Analytics and Computing, Medical Informatics, Neuromorphic Computing and Neurorobotics. This is currently the largest project for brain research. The goal of this project is the development of scientific infrastructure in neurophysiology, medicine and computer technology. This project not only research human brain, but also rodents' brain and other species. It also investigates ethical issues arising from the study of the brain.

HCP dataset includes not processed and preprocessed fMRI data [19]. Preprocessed data is data with remove head movement and resizing of images. There are two types of fMRI available: 3T fMRI and 7T fMRI [20].

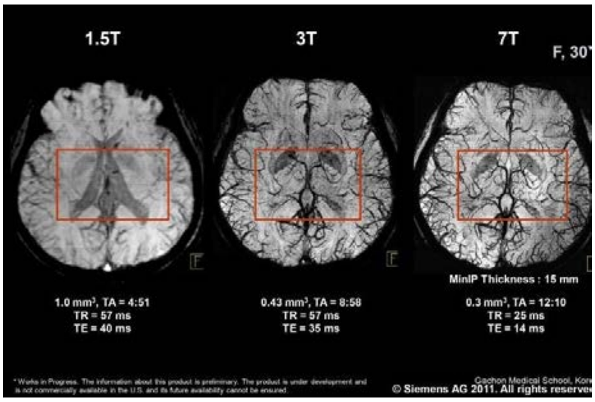

1.5T
3T
7T
F, 30
MinIP Thickness : 15 mm
1.0 mm³, TA = 4:51
TR = 57 ms
TE = 40 ms
0.43 mm³, TA = 8:58
TR = 57 ms
TE = 35 ms
0.3 mm³, TA = 12:10
TR = 25 ms
TE = 14 ms
* Works in Progress. The information about this product is preliminary. The product is under development and is not commercially available in the U.S. and its future availability cannot be ensured.
Gachon Medical School, Kor
© Siemens AG 2011. All rights reserve

Fig. 1. Difference between different types of fMRI

It is planned to use 3T fMRI, because there are more people images than for 7T fMRI (1032 people vs 138 people). Four experiments were done for each person. Each experiment lasted 14.4 minutes, timestep was 0.72 seconds. fMRI image is a 4D image (spatial and time coordinates), which uses NIFTI format [21].

Methods for Searching Nonlinear Functional Connectivity

General linear models. General linear model (GLM) is a well-known procedure to compute statistical linear models. It may be written as $y = w_1x_1 + \cdots + w_nx_n$.

It is one of the most popular method in neurophysiology [10, 11, 22, 23]. Its popularity is explained by the fact that the method has low computational complexity. It can be seen that assumption, that variables are linearly dependent is made. It should be noted, that GLM can be used for constructing nonlinear relationship with some modification. For this purpose, functions ϕ_i are defined, where ϕ_i is nonlinear combination of input variables, so the resulting function is following:

$$y = w_1\phi_1(x_1, \dots x_n) + \cdots + w_m\phi_m(x_1, \dots x_n).$$

The disadvantage of this approach is that these functions need to be defined in advance. Since it is impossible to sort out all combinations of functions, it is likely to overlook meaningful functions.

Genetic Programming. Genetic programming is a method that helps to restore non-linear functional dependence. This method is based on the idea of biological evolution. At the beginning there is a set of functions (usually, they are set randomly). Then the iterative process begins, in which the functions are changed in any way and those having best approximation are selected.

For convenience, the function is represented as a tree. For example, on Fig. 2 a function $f(x, y) = (\sin x + 3y)\cos(xy)$ is depicted.

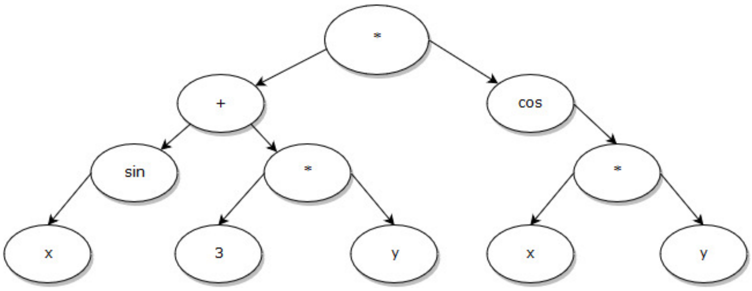


Fig. 10. Example of function in tree structure view

At each operation (selection, mutation and crossover) are iteratively repeated. In crossover operation two trees are taken, then random node in each tree for these functions is swapped. Mutation operation differs from crossover in that there is only one function involved. This operation consists on that a random node is selected and either it changes itself or the entire subtree that corresponds to that node.

This method does not require any assumptions about the functional dependences in advance; however, the complexity of genetic programming algorithm grows exponentially with search space increase. It poses a problem, because fMRI data is a high dimensional data. Some researches [24] try to bypass this problem by combining deterministic approach and genetic programming. They first built a simple model with a large number of signs, and then selected a few best ones. On these selected traits, they already apply the method of genetic programming. This approach has its advantages and disadvantages. The authors conducted an experiment, where they applied a genetic algorithm on the original features and on those already selected. It is shown that that with the same number of iterations, the algorithm with the selected features produces smaller error. The disadvantage of this approach is that by selecting signs, dependence information between several regions is lost, since simpler model can consider them not important. Other approach is to decrease search space with PCA/ICA [14].

Multilayer Perceptron. Another method for computing analytical form of functional connectivity is multilayer perceptron (MLP). MLP is a class of feedforward networks, consisting of at least three layers: input, hidden and output. Fig. 3 depicts a simple example of MLP. In analytical form it is $y = 4f(-x_1 + 3x_2) - 2f(2x_2 + x_3)$. Function f called activation functions. The activation function is commonly used:

1. RELU (rectified linear unit) function: $y = \max(0, x)$;
2. Identity function: $y = x$;
3. Logistic function: $y = \frac{1}{1+e^{-x}}$;
4. Tanh function: $y = \frac{e^x - e^{-x}}{e^x + e^{-x}}$.

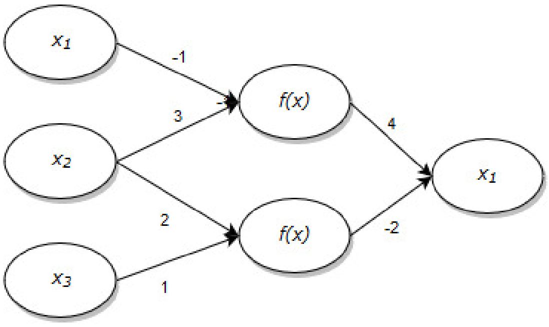


Fig. 3. Example of multilayer perceptron

This model allows moving away from assumptions about the form of functional dependence. The disadvantage is that this model requires much more computing time than simpler models. Another disadvantage is that if the constructed model is large enough, the analytical form is difficult to read and analyze.

Functional Connectivity Difference for Men and Women

In [14] the approach for computing functional linear connectivity is presented, though it does not provide more complex relations, which are of great interest. In this article [15], the authors caused negative emotions in men and women with the help of the olfactory system. All subjects were divided by gender. Using statistical tests, it was found that the activation of neurons in men and women is different. In another article [16], the authors also studied gender differences in fMRI, but at rest. Statistical tests were also used, and differences were found in some areas of the brain. The disadvantage of these two works is that here only a difference between some regions of the brain was found, but the reasons for these changes were not shown. The aim of this work is to search for nonlinear functional connections between brain regions in analytical form. If it can be getting an analytical form of relations between regions of the brain, then in the future it can be use various mathematical methods in order to understand exactly how one region depends on others**.**

3 Implementation

Workflow

Workflow is depicted on Fig. 4. First, using HCP preprocessed resting fMRI dataset, regions of interest are extracted. Most popular method for getting information about brain regions (often brain regions name regions of interest (ROI)) is atlas (atlases define mapping of voxels to brain regions). Recently, it also became possible to use machine learning methods to extract ROI time series [25]. This task is performed using atlases from NiLearn [26] and NiPy [27] Python packages.

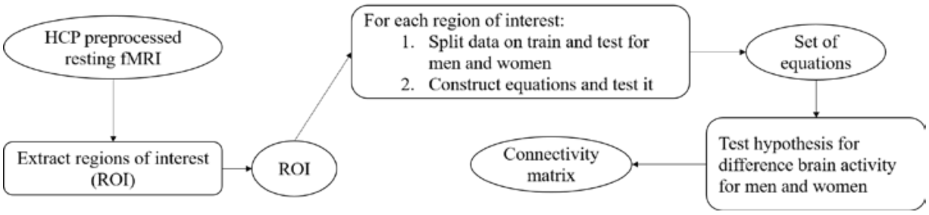

HCP preprocessed resting fMRI
Extract regions of interest (ROI)
ROI
For each region of interest:
1. Split data on train and test for men and women
2. Construct equations and test it
Set of equations
Test hypothesis for difference brain activity for men and women
Connectivity matrix

Fig. 11. Workflow

For each region of interest following procedure is applied: 1) data is split into train and test datasets for men and women; 2). set of equation is constructed using some algorithm and is validated on test data. As each region of interest is processed independently, each procedure is packaged into PySpark [28] job.

After that, equations are concatenated together. Statistical testing is invoked later to produce gender connectivity matrix, as in [14].

Preliminary Results

This section provides a comparison of the three models described above. Computational part was executed on machine with two Intel Xeon e5-2670 v2, 96 gigabytes of RAM and three GPUs Nvidia Titan X Pascal with 36Gb of video RAM total.

The R^2 score metric was used for model quality assessment. The results are depicted on Fig. 5.

L1 regularization is used in the construction of linear regression to reduce number of dependent regions.

During the execution of genetic programming, an additional feature which comes from the linear model is used. Unless it is used, the results are worse than linear regression. This could be due to the fact that there are too many free variables and not enough iterations to find global optimum. Genetic programming gave better results than linear regression. The multilayer perceptron gives the worst result, so additional heuristics are needed to improve its performance.

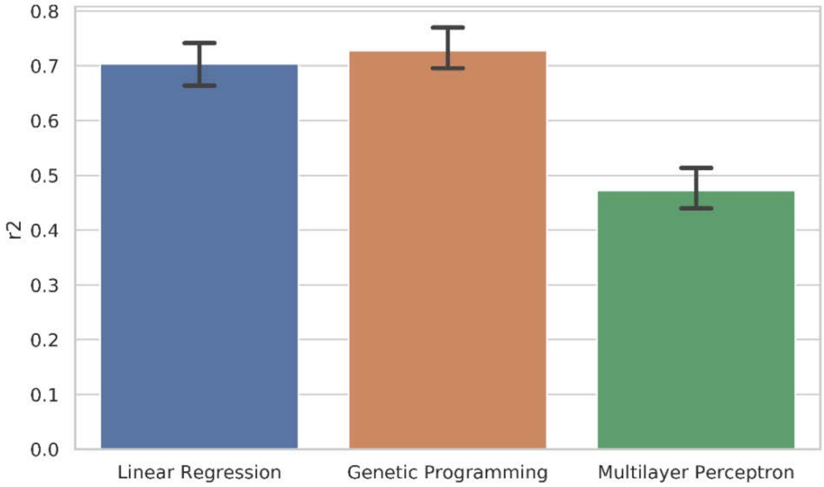


Fig. 4. Model Comparison

Following formulas are obtained.

Linear regression:

$$x_0 = 0.05x_2 + 0.43x_3 + 0.008x_4 - 0.003x_6 - 0.05x_7 - 0.01x_9 + 0.2x_{11} - 0.05x_{12} + 0.05x_{20} - 0.04x_{23} + 0.04x_{25} + 0.09x_{27} + 0.12x_{29} + 0.03x_{30} + 0.07x_{31} + 0.04x_{32} - 0.03x_{34} + 0.09x_{40} - 0.01x_3^2 - 0.005x_{20}^2 + 0.02x_3x_6;$$

Multilayer perceptron:

$$h_1 = -0.005x_1 + 0.005x_2 + 0.29x_3 - 0.02x_4 - 0.02x_5 - 0.0008x_6 - 0.006x_7 + 0.02x_8 - 0.02 - 0.02 - 0.006x_{11} - 0.02x_{12} + 0.02x_{13} + 0.003x_{14} - 0.02x_{15} + 0.007x_{16} + 0.07x_{17} - 0.01x_{18} + 0.002x_{19} - 0.02x_{20} + 0.015x_{21} - 0.04x_{22} + 0.03x_{23} - 0.02x_{24} + 0.01x_{25} - 0.03x_{26} + 0.04x_{27} - 0.007x_{28} + 0.01x_{29} - 0.01x_{30} + 0.03x_{31} - 0.007x_{32} - 0.04x_{33} - 0.06x_{34} + 0.03x_{35} + 0.007x_{36} + 0.05x_{37} + 0.02x_{38} + 0.05x_{39} - 0.006x_{40} + 0.04x_{41} - 0.005x_{42} - 0.02x_{43} + 0.002x_{44} + 0.03x_{45} - 0.002x_{46} + 0.05x_{47};$$

$$h_2 = 0.003x_1 - 0.05x_2 + 0.03x_3 - 0.1x_4 + 0.06x_5 - 0.13x_6 - 0.03x_7 - 0.06x_8 + 0.007x_9 - 0.03x_{10} - 0.06x_{11} - 0.02x_{12} + 0.01x_{13} + 0.21x_{14} - 0.004x_{15} - 0.09x_{16} + 0.06x_{17} + 0.31x_{18} + 0.02x_{19} + 0.01x_{20} - 0.002x_{21} + 0.13x_{22} - 0.04x_{23} + 0.07x_{24} + 0.01x_{25} + 0.02x_{26} - 0.01x_{27} - 0.2x_{28} - 0.002x_{29} + 0.004x_{30} + 0.009x_{31} + 0.001x_{32} - 0.09x_{33} + 0.01x_{34} - 0.01x_{35} - 0.02x_{36} - 0.18x_{37} + 0.02x_{38} - 0.05x_{39} + 0.07x_{40} + 0.04x_{41} + 0.02x_{42} + 0.008x_{43} - 0.03x_{44} - 0.01x_{45} + 0.09x_{46} - 0.06x_{47};$$

$$x_0 = 0.1218\frac{1}{1+e^{-h_1}} - 0.2706\frac{1}{1+e^{-h_2}}.$$

Genetic programming: $x_0 = x_{13}x_{14}x_{25}^2x_9 + y_1$, where y_1 is the same function as function for linear regression.

It can be seen that genetic programming algorithm produces the most readable form of result.

4 Conclusion

The research is done as thesis for Master's Program "Big data: infrastructures and problem-solving techniques" in the department of Computational Mathematics and Cybernetics of Moscow State University. Several methods are studied to find the nonlinear functional connectivity between regions of the human brain. Though genetic programming does not perform well in high dimensional space and needs proper features, it shows best results compared to linear regression and multilayer perceptron.

As future work, it is planned to 1) build analytical functions separately for men and women; 2) improve the results for multilayer perceptron, so that it can be used for obtaining analytical formulas; 3) test hypotheses for functional connectivity difference for men and women.

Acknowledgements

This work is supervised by Dmitry Kovalev, Federal Research Center **"**Informatics and Control**"** of Russian Academy of Sciences.

References

1. Human Connectome Project, https://www.humanconnectome.org/.
2. Lee, M., Joanna, V., Bruce, B., and Shapiro, K.: Neurobiology of Brain Disorders. 2nd edn. Academic Press (2014).
3. Haiqing, H. and Mingzhou, D.: Linking functional connectivity and structural connectivity quantitatively: a comparison of methods. Brain Connect **6** (2), 99–108 (2016).
4. Friston, K.: Functional and effective connectivity in neuroimaging: a synthesis. Human brain mapping, 2, 56–78 (1994).
5. Politis, M., Gennaro Pagano, and Flavia Niccolini: Chapter nine – imaging in Parkinson's disease. International Review of Neurobiology **132**, 233–274 (2017).
6. Loe-Heidi, I. and Feldman, M.: Attention-deficit and hyperactivity disorders. In 2th Neural Basis of International Encyclopedia of the Social & Behavioral Sciences. Elsevier (2015).
7. Damoiseaux, J.: Resting-state fMRI as a biomarker for Alzheimer's disease? Alzheimer's Research & Therapy **4** (2) (2012).
8. Shu Zhang, Xiang Li, Jinglei Lv, etc: Characterizing and differentiating task-based and resting state FMRI signals via two-stage sparse representations. Brain Imaging Behav. **10** (1), 21–32 (2016).
9. H. Lv, Z. Wang, E. Tong, etc.: Resting-state functional MRI: everything that nonexperts have always wanted to know. American Journal of Neuroradiology 39 (8), 1390–1399 (2018).
10. Soch, J., Meyer, A., and Haynes, J.: How to improve parameter estimates in GLM-based fMRI data analysis: cross-validated Bayesian model averaging. Neuroimage **158**, 186–195 (2017).
11. Eklund, A., Lindquist, M., and Villani, M.: A Bayesian heteroscedastic GLM with application to fMRI data with motion spikes. Neuroimage **155**, 354–369 (2017).

12. Pierre-Jean, L., Jean-Baptiste, P., Guillaume, F., and Silke Dodelline, G.: Functional connectivity: studying nonlinear, delayed interactions between BOLD signals **20** (2), 962–974 (2003).
13. Karanikolas, G., Giannakis, G., Slavakis, K. etc.: Multi-kernel based nonlinear models for connectivity identification of brain networks. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 6315–6319, IEEE (2016).
14. Kovalev, D., Priimenko, S., and Ponomareva, N.: Search for Gender Differencein Functional Connectivity of Resting State fMRI. Data analytics and management in data intensive domains, 190–196 (2017).
15. Kochab, R., Paulya, K., Kellermann, T. etc.: Gender differences in the cognitive control of emotion: An fMRI study. Neuropsychologia **45** (12), 2744–2754 (2007).
16. C. Xu, C. Li, H. Wu, etc.: Gender differences in cerebral regional homogeneity of adult healthy volunteers: a resting-state FMRI study. Biomed Research International (2015).
17. 1000 functional connectomes project, http://fcon_1000.projects.nitrc.org/.
18. Human Brain Project, https://www.humanbrainproject.eu/.
19. Glasser, M.F., Sotiropoulos, S.N., Wilson, J.A., etc.: The minimal preprocessing pipelines for the Human Connectome Project. Neuroimage **80,** 105–124 (2013).
20. Hyeong Cheol Moon, Hyeon-Man Baek, and Young Seok Park: Comparison of 3 and 7 Tesla magnetic resonance imaging of obstructive hydrocephalus caused by tectal glioma. Brain Tumor Research and Treatment **4** (2), 150–154 (2016).
21. Niftii file format, https://nifti.nimh.nih.gov/.
22. Jiansong Xu, Marc N. Potenza, Vince D. Calhoun, etc.: Large-scale functional network overlap is a general property of brain functional organization: Reconciling inconsistent fMRI findings from general-linear-model-based analyses. Neuroscience & Biobehavioral Reviews **71**, 83–100 (2016).
23. Chung, Moo K., Vilalta, Victoria G., Rathouz, Paul J., Lahey, Benjamin B., and Zald, David H.: Linear embedding of large-scale brain networks for twin fMRI. arXiv preprint arXiv:1509.04771 (2016).
24. Ilknur, Icke, Nicholas, A., Allgaier Christopher M., Danforth, Robert A., Whelan, Hugh P., Garavan, Joshua, and Bongard, C.: A Deterministic and Symbolic regression hybrid applied to resting-state fMRI data. Springer, 155–173 (2014).
25. Xiaomu Song, Lawrence P. Panych, and Nan-kuei Chen: Brain functional mapping using spatially regularized support vector machines, IEEE Signal Processing in Medicine and Biology Symposium (SPMB). IEEE (2015).
26. Nilearn library, https://github.com/nilearn/nilearn/
27. Nipy library, https://github.com/nipy/
28. PySpark library, https://spark.apache.org/

INFORMATION EXTRACTION FROM TEXT – I

Expert Assignment Method Based on Similar Document Retrieval

D.V. Zubarev[1], D.A. Devyatkin[1], I.V. Sochenkov[1], I.A. Tikhomirov[1], and O.G. Grigoriev[1]

[1] Federal Research Center “Computer Science and Control” of Russian Academy of Sciences, Moscow, Russia

Abstract. The paper describes the problem of expert assignment. Based on the analysis of methods that are currently used to solve this problem, the main shortcomings of these methods were identified. These shortcomings can be eliminated by analysing large collections of documents whose authors are potential experts. The article describes the method of compiling a ranked list of experts for a given document, using similar document retrieval. To evaluate the proposed method, we used a collection of grants applications from a science foundation. Experimental studies show that the more documents are available where experts are authors, the better the performance of the proposed method becomes. In conclusion, the current limitations of the proposed method are discussed, and future work is described.

Keywords: Scientific expertise, expert assignment, unstructured data analysis, text analysis, similar document retrieval

1 Introduction

A competent and objective examination of applications for grants and scientific publications is a prerequisite for scientific progress. But it requires a competent and objective selection of experts. Currently, in most cases, the appointment of experts is based on manually assigned codes from manually created classifiers or manually chosen keywords. Experts and authors independently assign codes or keywords to their profiles or documents (application for a grant, report on a grant, an article, etc.), and the appointment of an expert is carried out by comparing the assigned codes or keywords. Classifiers are rarely updated, so they quickly become obsolete, have uneven coverage of the subject area (one code can correspond to thousands of objects, and the other to dozens) and have all the other drawbacks of manual taxonomies. In addition, experts often assign themselves several codes, but their level of competence varies greatly between these codes [1]. If there are several dozen experts who correspond to the same code (which happens quite often), then the further choice will be extremely subjective and non-transparent (in fact, manual selection of an expert is performed). All this leads to insufficient compliance of the competence of the selected expert and the object of examination and, possibly, to the subjective choice of the expert. As a result, there are refusals of examination, or it is conducted incompetently and, possibly, subjectively. Therefore, it is important not to determine the formal coincidence of the expert interests and the expertise subject topic, but to use all possible information for accurate expert ranking.

Information about expert competence is accumulated in the documents in which he participated (scientific articles, scientific and technical reports, patents, etc.). This information is much more precise in determining the expert's knowledge area than the classifier codes or keywords. This article describes the method of searching and ranking experts for a given object of expertise using thematically similar documents retrieval. The method requires a database of experts and a large set of texts associated with experts. It is assumed that this method will become the basis for a whole class of methods that use unstructured information to select experts for the objects of expertise.

2 Related Works

Automating the search for an appropriate expert for examination has long been a subject of research. Researchers often narrow the research scope, for example, limiting themselves only to the appointment of experts to review articles submitted to the conference [2], or to select an expert who will answer user questions on the corporate knowledge base [3].

As a rule, expert assignment methods are divided into two groups [4]. The first group includes methods that require additional actions from experts or authors. For example, one of the methods involves the examination of the submitted abstracts of articles and the self-assessment of his readiness to consider any of the works in question. Another involves the selection by an expert of keywords that describe his competence from the list provided by the conference organizers and comparing these keywords from the expert with the keywords chosen by the authors of the article. These approaches are well-suited for small conferences but are not suitable for events in which several tens of thousands of participants take part. Even with relatively small conferences, the use of keywords is inappropriate if the number of topics for this event is large enough.

The second group includes methods that automatically build an expert's competence model based on his articles and / or other data and compare the resulting model with peer-reviewed articles submitted using the same model [5]. In this work, the name and surname of the expert were sent as a request to Google Scholar and CiteSeer. For the full texts of the articles found and the article submitted, the Euclidean distance was measured. This method does not take into account the namesakes, the dynamics of changes in the expert interests, a possible conflict of interests and requires significant computational resources. Another method [6] uses annotations and titles to classify articles according to topics predetermined by the conference organizers. However, it is not always possible to pre-determine a specific set of topics. The method presented in [7] uses bibliographic data from the reference list of the presented article. First names and surnames of authors are mined from bibliographic references, co-authors are determined for them using external resources (DBLP), etc. Thus, a co-authorship graph is constructed, on which a modification of the page ranking algorithm for identifying experts is performed. In [8] a special similarity measure that compares the reference lists are used to determine the proximity between expert publications and article submission. The comparison takes place under headings and authors, it also takes into account the case when the expert's articles are cited in the presented article. Bibliographic list comparison is a fairly effective operation, but it is difficult to assess the expert's competence only by bibliographic references, without using full texts. In [9] topic modelling is used to represent an object of examination and each document associated with an expert. Topic distribution of an expert is adjusted according to the time factor that is meant to capture the changes of research directions of an expert as time goes on. Cosine measure is used to measure the similarity between the expert's topic distribution and the topic distribution of the object under review. Furthermore, vector space model (with TF-IDF weighting scheme) is used to calculate an additional similarity score between experts and the object of examination. The final score for relevance is calculated using a weighted sum that takes into account the two previous scores. In the experimental studies of this work, the number of topics was chosen to be 100, which, according to the authors, reflects the real number of topics in information technology knowledge area. In this study, words are used as features.

In [10] a hybrid approach is used that combines full-text search (performed using ElasticSearch) over experts' articles and an expert profiling technique, which models experts' competence in the form of a weighted graph drawn from Wikipedia. The vertices of the graph are the concepts extracted from the expert's publications with TagMe tool. Edges represent the semantic relatedness between these concepts computed via textual and graph-based relatedness functions. After that, each vertex is assigned a score corresponding to the competence of the expert. This score is computed employing a random walk method. Concepts with a low score are removed from the expert's profile, due to the assumption that they cannot be used to characterize his competence. Also, the vertices are assigned a vector representation which is learned via structural embeddings techniques on concepts graph. At query time, the object of examination is parsed with TagMe tool, and embeddings are retrieved for extracted concepts, then they are averaged. As a result, a cosine measure is used to measure the similarity between averaged expert's vectors and vector that represents the object of the expertise. The final list of relevant experts is obtained via combining full-text search results and results of semantic profiles matching. It should be noted that impact of semantic profiles is rather small. According to the results of experimental studies conducted in this work, the increase in the quality assessment using the expert's semantic profile was 0.02, compared to the use of full-text search with the BM25 ranking function. This method was tested on a dataset [11], in which short phrases describe areas of knowledge (GT5). These phrases were used as queries (objects of expertise).

Thus, the existing methods for the expert search do not use all available information related to this task. Some methods are limited to processing only bibliographic lists or annotations with titles while ignoring the full texts of articles. Others are based on full text analysis of articles but they use ordinary full-text retrieval tools that apply to simple keyword search and are not effective for thematically similar documents retrieval. In addition, it should be noted that some of the methods described are computationally expensive since they do not use efficient means of indexation, and when selecting experts for each new object, it is necessary to repeat complex computational operations. In the approach proposed in this article, the main part of computationally intensive operations is performed only once.

3 Expert Assignment Method

The first step of the proposed method is the search for thematically similar documents for a given object of examination (application) [12]. Search is made on the collections of scientific and technical texts. These can be scientific articles, patents and other documents related to experts. The collections are pre-indexed. Before indexing the text undergoes a full linguistic analysis: morphological, syntactic and semantic [13, 14]. Indexes store additional features for each word (semantic roles, the syntactic links and so on) [15]. During indexing, several types of indexes are created, including an inverted index of words and phrases, which is used to search for thematically similar documents. Indexing is incremental; that is, after initial indexing, one can add new texts to the collection without re-indexing the entire collection [15].

When searching for thematically similar documents, the given document is represented as a vector, elements of which are TF-IDF weights of keywords and phrases. Phrases are extracted based on syntactic relations between words. This allows extracting phrases consisting of words that are not adjacent to each other but have a syntactic connection. For example, the phrases “images search” and “digital images” will be extracted for the fragment: “search for digital images”. The degree of similarity is calculated between the vector of the original document and the documents vectors from the index. Some similarity measure is used to calculate the degree of similarity (we tried cosine and hamming distance). The main parameters of the search method for thematically similar documents are presented in Table 1.

Based on the list of thematically similar documents, a list of candidates for experts is compiled. This is a trivial operation since the documents relate to the expert: the expert is one of the authors, he reviewed this article/application, etc.

After that, if there is the necessary meta-information, the experts are excluded from the list of candidates according to various criteria. For example, if meta-information about belonging to organizations is available for a peer-reviewed document and an expert, some experts could be filtered out because of a conflict of interest. At present, this step depends on the available meta-information, and it is related to the type of reviewed document. The experimental implementation of the method used several filters that are appropriate for grant applications:

1. All experts who are involved as participants in the given application are excluded.
2. All experts working in the same organizations as the head of the given application are excluded.

Table 1. The main parameters for thematically similar documents search method

Description	Name
The percent of words and phrases in the source document that determine the similarity of documents	`TOP_PERCENT`
The maximum number of words and phrases that are used to determine document similarity	`MAX_WORDS_COUNT`
The minimum number of words and phrases that are used to determine the similarity of documents	`MIN_WORDS_COUNT`
Minimum TF-IDF weight of a word or phrase included in the top keywords of the document	`MIN_WEIGHT`
The minimum value of the similarity score	`MIN_SIM`
The maximum number of similar documents for the source document	`MAX_DOCS_COUNT`

After that, the relevance of each expert to the object of expertise is calculated. The calculation takes into account the similarity of the documents (S_{sim}), with which the expert is associated, to the reviewed document, as well as several additional measures. In case if the expert has multiple documents, their ratings of similarity are averaged out. The set of additional measures depends on the type of the reviewed document. In the implemented method, one simple measure (S_{sci}) was used: the equality of the knowledge area code assigned to the expert and to the document under review (0 when the codes are not equal and 1 otherwise). The overall relevance score of the expert is calculated using the following formula:

$$W_{sim}\,S_{sim} + W_{sci}\,S_{sci},$$

where S_{sim}, S_{sci} are values of the measures described above, as W_{sim}, W_{sci} are weights with the condition $W_{sim} + W_{sci} = 1$. S_{sci} criterion was useful in ranking experts who were heads of interdisciplinary projects. An interdisciplinary project can relate to several scientific areas, but the head is an expert in only one area, so he should be ranked lower than the experts who have the same area of knowledge. The score of the relevance of each expert may lie in the interval [0;1]. After evaluation, experts are ranked in descending order of relevance.

4 Description of the Experimental Setup

4.1 Dataset Description

As a result of cooperation with the Russian Foundation for Basic Research, it was possible to conduct a series of experiments on the applications accumulated by the Foundation in various competitions held from 2012 to 2014. The Fund provided an API for indexing the full texts of applications. The application text included:

- summary of the project;
- description of the fundamental scientific problem the project aims to solve;
- goals and objectives of the study;
- proposed methods;
- current state of research in this field of science;
- expected scientific results;
- other substantive sections that are required by the competition rules.

For each application, a meta-information containing the following fields was provided:

- document identifier;
- the identifier of the head (principal investigator);
- identifier of the organization in which the head works;
- a list of the identifiers of participants (co-investigators);
- coded participants full names;
- publication year of the grant application;
- code of the field of knowledge which the application belongs to (Biology, Chemistry, etc.));
- main code and additional application codes;
- keywords of the application.

There was also presented impersonal information about the experts who reviewed the applications:

- expert identifier;
- identifier of the organization which the expert works in;
- expert keywords;
- code of the main area of knowledge of an expert;
- applications which the expert is the head in (list of identifiers);
- applications which the expert is the participant in (list of identifiers);
- applications reviewed by the expert (list of identifiers);
- applications the expert refused to review (list of identifiers).

The size of the collection of applications was about 65 thousand documents. Information was also received about 3 thousand experts, where the share of experts who were the head (principal investigator) of at least one application was 78%. At first, it was supposed to use only the applications of experts, in which they were principal investigators. However, it turned out that the share of such documents was about 9% among all grant applications. Moreover, most of the experts were associated with only one grant project. To increase the number of documents associated with the experts, we took into account applications in which the expert participated as co-investigator. We also used an external collection of scientific papers, which mainly consisted of articles from mathnet.ru and cyberleninka.ru, to search for additional experts publications. First, we looked for documents that confirm the support of grants with the participation of the expert (the grant identifier is usually written in the acknowledgments section). This provided us with about 4,000 additional documents. In addition, we performed a search for similar works for each expert application. To filter documents that are similar, but not related to experts, we compared the full names of the authors of the article with the full names of the applicants. If at least one full name corresponded, then we considered that this document is associated with an expert. Usually, there are no full names of the authors of the article, there are only short names (last name with initials), then there should be at least two matches with the short names of the applicants. We received about 30,000 new documents related to experts, using the search for similar documents.

Since the names of authors of papers are not structured and are presented as text, we parsed names into their individual components. We will briefly describe the parsing method. First, given the input string that contains the name of the author, the type of pattern is identified. Multiple patterns are supported:

1. The Slavic pattern includes several variations:
   a. Last name[,] First name [Patronymic];
   b. Last name Initials;
   c. First name [Patronymic] Last name;
   d. Initials Last name;
2. The Western pattern consists of several variations:
   a. First name [Middle]… Last name
   b. Last name, First name [Middle]…
   c. First name Initial [Initial]… Last name
3. Spanish pattern similar to the western one, except that there may be two last names:
   a. First last name [Second last name], First name [Middle]
   b. First name [Middle]… First last name [Second last name]
4. Asian pattern:
   a. Last name First name [First name]…

This classification is necessary because the parser can match full names with several patterns, e.g. 1.1 and 2.1. As a training set, we used the names of public persons and the country of their citizenship obtained from Wikidata dump, and also added the names with countries obtained from Russian patents (www1.fips.ru). We trained Fasttext classifier on that dataset and obtained 0.96 precision@1 on the test data. When a pattern is identified for the given input text, then all variations available for this pattern are tested. If there is only one matching option, the parsing is complete. If more than one option matches, for example, 1.1 and 1.3; we use the common first names dictionary to determine the right variation.

After performing these procedures, the share of experts with documents increased to 88%. In addition, the number of experts associated with only one document was significantly reduced, as can be seen in Table 2.

Table 2. Distribution of the number of documents associated with an expert including extra documents

Number of documents per expert	Number of experts	Number of documents per expert	Number of experts	Number of documents per expert	Number of experts	Number of documents per expert	Number of experts
1	115	11	50	21	24	31	9
2	101	12	45	22	18	32	14
3	108	13	39	23	26	33	13
4	94	14	39	24	24	34	9
5	83	15	34	25	12	35	7
6	76	16	34	26	19	36	6
7	79	17	36	27	17	37	11
8	64	18	31	28	14	38	3
9	66	19	28	29	16	39	9
10	55	20	19	30	9	40	11

4.2 Evaluation Methodology

To assess the proposed method, data from previous expert selections of applications for participation in the A-2013 competition was used (total of 10,000 applications, an average of 3 experts per application). For every application from this competition, a ranked list of experts (found experts) was compiled using the proposed method. Then this list of experts was compared with the list of experts assigned to the given application.

There are common metrics that are used for evaluation of the experts search methods [16, 17]. Some of these metrics are applicable only if expert assignment goes along with the expert search. Those metrics assess the uniformity of the expert load and the assignment of a certain number of experts to each object of expertise. Each expert search provides a pool of relevant experts for further assignment. Therefore, this task should be evaluated using other metrics. Classical information retrieval metrics are frequently used: MAP, NDCG@100. Using these metrics is justified if the test data contains a large number of relevant experts for each object of expertise. We used a data set of up to 3 relevant experts for the object of expertise. This number of experts is not enough to correctly interpret the assessment results for a large number of selected experts. (like 100). Therefore, recall was used for evaluation in order to determine what total share of relevant experts was in the pool of selected experts. Recall was calculated using the following formula:

$$Recall = \frac{F_{found}}{F_{total}},$$

where F_{found} is the number of found experts from among those that have been assigned to this application; F_{total} is the number of assigned experts that could be found by the method (i.e. only experts that have at least one associated document).

Micro averaging was used to calculate metrics for all applications (i.e. for all applications are summarized F_{found} and F_{total}, and based on this, the required metric was calculated). Also, recall was calculated separately for each knowledge area.

The standard way of measuring precision in this situation is not appropriate since it is not known whether the found expert that has not been assigned to this application is suitable. He might be suitable for this application but was not assigned because he was busy on other projects or for other reasons.

Therefore, to evaluate precision, the information on this application expertise refusals was used. There were about 2 thousand of refusals according to the provided data. The idea is, that the compiled experts list shouldn’t contain those who refused to review this application. The precision was calculated using the following formula:

$$Precision\ = 1 - \frac{R_{found}}{R_{total}},$$

where R_{found} is the number of found experts from those who refused to expertise the application; R_{total} is the number of refused experts that could be found by the method (i.e. experts with documents).

4.3 Parameters Optimization

Optimization of the algorithm parameters was performed on a separate sample collection of 700 applications. For optimization, we used a grid search of a single algorithm parameter with a fixed value of the remaining parameters. Optimization was performed to maximize recall. The results are presented in Table 3.

4.4 Experimental Results

We conducted multiple experiments with different similarity measures (cosine, hamming) and a different set of features (only words, words with phrases). Also, we used different datasets: at first, experts were associated with applications, in which they are the head (only-head); then additional documents were added (extra-docs). Table 4 shows the micro-averaged recall and precision on the top 150 for those experiments.

Hamming distance along with adding word phrases result in the best recall on two datasets. New documents addition associated with experts (extra-docs dataset) increased the value of recall, but decreased the precision by almost the same value. As we discussed earlier, MAP is not the best metric for this task. Its value depends on the experts ranking, but this ranking is better when smaller dataset is used (only-head). Using more docs associated with experts (extra-docs) gives greater recall but lesser MAP. It should be borne in mind that experts found should be distributed over several dozens or hundreds of applications, and several experts are usually appointed for each application. Therefore, each application requires a sufficiently large pool of relevant experts. Therefore, recall is a more important metric for this task than MAP.

Table 3. Values of method parameters after optimization

Name	Value
`TOP_PERCENT`	0.4
`MAX_WORDS_COUNT`	200
`MIN_WORDS_COUNT`	15
`MIN_WEIGHT`	0.03
`MIN_SIM`	0.05
`MAX_DOCS_COUNT`	500
W_{sci}	0.1
W_{sim}	0.9

Table 4. Evaluation results

	Only-head			Extra-docs		
	Recall	**Precision**	**MAP**	**Recall**	**Precision**	**MAP**
Cosine, only words	0.67	0.52	0.136	0.73	0.43	0.098
Cosine, with phrases	0.69	0.51	0.139	0.75	0.42	0.108
Hamming, only words	0.69	0.52	0.141	0.76	0.4	0.107
Hamming, with phrases	0.7	0.5	0.148	0.77	0.41	0.123

Since the result of the method is a ranked list of experts, it is possible to plot a graph of recall and precision (Hamming, with phrases) shown in Fig. 1.

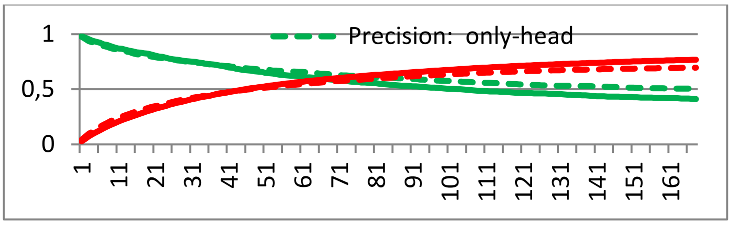


Fig. 1. The dependence of recall and precision on the rank

The graph shows that the maximum recall is achieved at 100–120 rank, after that, recall almost does not change. Also, this graph shows that ranking on smaller dataset (only-head) is better, because recall values on low ranks (1–40) are better than extra-docs dataset.

Recall values were also calculated for each area of knowledge separately, and the results are shown in Table 5.

Table 5. Recall for each knowledge area

	only-head	**extra-docs**
Mathematics informatics and mechanics	0,69	0,79
Physics and astronomy	0,71	0,81
Chemistry	0,74	0,79
Biology and medical science	0,74	0,75
Earth Sciences	0,76	0,87
Human and Social Sciences	0,77	0,92
Information technologies and computer systems	0,55	0,61
Fundamentals of Engineering	0,57	0,64

The table shows that the best results were obtained in the fields of Earth Science (5) and the Human and Social Sciences (6) – recall is about 90%. In other areas, good results were obtained (recall from 70% to 80%). Average results were obtained for the fields of Information technologies (7) and Fundamentals of engineering (8). In addition, the figure shows the increase of the number of documents related to experts has a positive impact on assignment of experts recall in all knowledge areas.

5 Conclusion

In this paper, the expert appointment method based on the analysis of text information was described, and the results of method evaluation experiments were presented. We proposed a new evaluation methodology and conducted experiments on the RFBR data set, which distinguishes this work from the previous ones. The method showed its viability, but it is necessary to improve it in order to increase the recall. Adding more expert-related texts improved the review somewhat, but not as dramatically as expected. According to Fig. 2, there are still many experts that have only one document authored by them. It may be viable to add documents related to these experts in the first place: scientific publications, scientific and technical reports, patents, etc. It is also possible to expand the list of documents indirectly related to the expert, with the exception of authorship, for example, articles that he reviewed. These documents should contribute to the overall score of relevance with a lower ratio since the expert has no direct relationship to the text, however, if he regularly reviews the papers of a certain topic, it should be taken into account when scoring.

In further experiments, it is also proposed to expand the set of criteria that affect the expert's assessment for a given object of examination, for example, add an expert rating calculated using a page ranking algorithm based on quotes from expert works.

Further studies are also expected to improve the methodology for evaluation of the expert’s assignment. In the technique proposed in the article, there are several shortcomings, namely: the dependence of recall on the original expert assignment, which could be subjective; the inability to assess the selected experts, which were not involved in the expertise of this proposal, which makes it impossible to calculate precision of the selection. Precision measurement based on refusals of expertise is also not optimal. Cases of refusal are 15 times less than cases of acceptance, and refusal can occur for other reasons than a mismatch between the competence of the expert and the subject of the application. However, the question of how to evaluate the work of the expert assignment method is currently unresolved. The involvement of external experts can significantly improve the quality of the evaluation, but it will require a large number of experts from different knowledge areas, who should be well acquainted with the expert community.

The proposed method can be used not only when appointing an expert for grant application of a scientific fund but also in the reviewer selection for any text object: articles in scientific journals, conference abstracts, patent applications, etc.

Acknowledgements

The research is supported by Russian Foundation for Basic Research (grant No. 18-29-03087).

References

1. V Rossijskom nauchnom fonde proshlo zasedanie ehkspertnogo soveta po nauchnym proektam [The Russian Scientific Foundation held a meeting of the expert council on scientific projects]. Available at: http://rscf.ru/ru/node/2367 last accessed 2019/08/16.
2. Dumais, Susan T. and Nielsen, Jakob: Automating the assignment of submitted manuscripts to reviewers. Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 233–244 (1992).
3. Balog, Krisztian, Leif Azzopardi, and Maarten De Rijke: Formal models for expert finding in enterprise corpora. Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval. ACM. 43–50 (2006).

4. Kalmukov, Yordan and Rachev, Boris: Comparative analysis of existing methods and algorithms for automatic assignment of reviewers to papers. arXiv preprint Available at: https://arxiv.org/pdf/1012.2019.pdf last accessed: 2019/05/11 (2010)
5. Pesenhofer, Andreas, Mayer, Rudolf, and Rauber, Andreas: Improving scientific conferences by enhancing conference management systems with information mining capabilities. Digital Information Management, 2006 1st International Conference on. IEEE, 359–366 (2006).
6. Ferilli, Stefano, et al.: Automatic topics identification for reviewer assignment. International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems. Springer, Berlin, Heidelberg, 721–730 (2006).
7. Rodriguez, Marko A., and Bollen, Johan: An algorithm to determine peer-reviewers. Proceedings of the 17th ACM conference on Information and knowledge management. ACM, 319–328 (2008).
8. Li, Xinlian and Watanabe, Toyohide: Automatic paper-to-reviewer assignment, based on the matching degree of the reviewers. Procedia Computer Science **22**, 633–642 (2013).
9. Peng, H. et al. Time-aware and topic-based reviewer assignment. International Conference on Database Systems for Advanced Applications. Springer, Cham, 145–157 (2017).
10. Cifariello, P., Ferragina, P., and Ponza, M.: Wiser: a semantic approach for expert finding in academia based on entity linking. Information Systems **82**, 1–16 (2019).
11. Berendsen, Richard, et al.: On the assessment of expertise profiles. Journal of the American Society for Information Science and Technology **64** (10), 2024–2044 (2013).
12. Sochenkov, I.V., Zubarev, D.V., and Tihomirov, I.A.: Eksplorativnyj patentnyj poisk [Exploratory patent search]. Informatika i ee primeneniya [Informatics and its Applications]. **12** (1), 89–94 (2018).
13. Osipov, Gennady, et al.: Relational-situational method for intelligent search and analysis of scientific publications. Proceedings of the Integrating IR Technologies for Professional Search Workshop, 57–64 (2013).
14. Shelmanov, A.O. and Smirnov, I.V.: Methods for semantic role labeling of Russian texts. Computational Linguistics and Intellectual Technologies. Proceedings of International Conference Dialog **13** (20), 607–620 (2014).
15. Sochenkov, I.V. and Suvorov, R.E.: Servisy polnotekstovogo poiska v informacionno-analiticheskoj sisteme (Chast' 1) [Full-text search in the information-analytical system (Part 1)]. Informacionnye tekhnologii i vychislitel'nye sistemy [Journal of Information Technologies and Computing Systems] **2**, 69–78 (2013).
16. Li, L., Wang, L., Zhang, Y.: A comprehensive survey of evaluation metrics in paper-reviewer assignment. Computer Science and Applications: Proceedings of the 2014 Asia-Pacific Conference on Computer Science and Applications (CSAC 2014), Shanghai, China, 27–28 December 2014. CRC Press, 2015. P. 281.
17. Lin, S. et al. A survey on expert finding techniques. Journal of Intelligent Information Systems **49** (2), 255–279 (2017).

Depression Detection from Social Media Texts

Maxim Stankevich[1], Andrey Latyshev[2], Evgenia Kuminskaya[3], Ivan Smirnov[4,5], and Oleg Grigoriev[6]

[1] Federal Research Center "Computer Science and Control" of RAS, Moscow, Russia
stankevich@isa.ru
[2] Limited Liability Company "RI Technologies", Moscow, Russia
andrey.latyshev@gmail.com
[3] Psychotherapy and Counselling Psychology FGBNU PI RAE, Moscow, Russia
evgenia.kuminskaya@gmail.com
[4] Federal Research Center "Computer Science and Control" of RAS, Moscow, Russia
[5] Peoples' Friendship University of Russia (RUDN University), Moscow, Russia
ivs@isa.ru
[6] Federal Research Center "Computer Science and Control" of RAS, Moscow, Russia
oleggpolikvart@yandex.ru

Abstract. Nowadays the problem of early depression detection is one of the most important in the field of psychology. Social networks analysis is widely applied to address this problem. In this paper, we consider the task of automatic detection of depression signs from textual messages of Russian social network VKontakte users. We describe the preparation of users' profiles dataset and propose psycholinguistic and stylistic markers of depression in text. We evaluate machine learning methods for detecting signs of depression from social media messages. The results of experiments show that psycholinguistic markers based features achieved 66% of F1-score on the binary classification task which is promising result in comparison with similar works.

Keywords: Depression Detection, Social Networks, Psycholinguistics.

1 Introduction

Nowadays the problem of early depression detection is one of the most important in the field of psychology. Over 350 million people worldwide suffer from depression, which is about 5% of the total population. Close to 800 000 people die due to suicide every year and it is statistically the second leading cause of death among people in 15–29 years old [1, 2]. At the same time, the major number of suicides associated with depression. Recent researches reveal that depression is also the main cause of disability and a variety of somatic diseases.

For example, F. I. Beliakov [3] in his paper summarizes the main results of recent depression, anxiety, and stress investigations and their relation to cardiovascular mortality. His overview shows that increased risk of death from cardiovascular diseases associates with depression and stress. P. G. Surtees et al. conducted a prospective study in the UK that based on the 8.5 years of observation [4]. This study provides that the presence of major depression is associated with a 3.5-fold increase in mortality from coronary heart disease (CHD). W. Whang et al. demonstrate that women with depression have an increase of fatal CHD by 49% in 9 years of follow-up [5]. These studies demonstrate that depression treatment and stress control, as well as early diagnosis and prevention of symptoms of psychological distress and mental disorders, can increase life expectancy.

Nevertheless, depression is still often falsely associated with a lack of willpower and unwillingness to cope with the “bad mood”. There is social stigmatization of this disease, and it is embarrassing to admit it for a person. As a result, people with depression often hide their condition, do not seek help in time, and aggravate the disease.

Online methods and social media provide an opportunity to privately detect the symptoms of depression in time. It would allow people to suggest measures for its prevention and treatment in the early stages. The report of the European branch of WHO (2016) paid special attention to the identification of signs of depression and the personalization of online methods of its prevention.

In this paper, we consider the problem of automatic detection of depression signs from textual messages of Russian social network Vkontakte users. We explored the ability of psycholinguistic and stylistic markers to predict depression from the text of messages. In Section 2, related works are reviewed, in Section 3 we present dataset of Vkontakte profiles, in Section 4 we describe our methods and feature engineering and in last sections, we present and discuss results of experiments.

2 Related Work

Instrumental possibilities of analyzing the behavior of users in social networks are actively developing. In particular, methods of computational linguistics are successfully used in analyzing the texts from social networks.

The computerized analysis method of texts LIWC (Linguistic Inquiry and Word Count) [6] allows assessing the extent to which the author of a text uses the words of psychologically significant categories. The method works on the basis of manually compiled dictionaries of words that fall into different categories: meaningful words (social, cognitive, positive/negative words, etc.), functional words (pronouns, articles, verb forms, etc.). LIWC is used for different languages, including Russian [7], but does not consider the specifics of the language, since it is simply a translation of dictionaries from English to Russian.

A. Yates et al. [8] used neural network model to reveal the risks of self-harm and depression based on posts from Reddit and Twitter and showed the high accuracy of this diagnostic method. The authors indicate that proposed methods can be used for large-scale studies of mental health as well as for clinical treatment.

Seabrook et al. [9] utilized the MoodPrism application to collect data about status updates and mental health of Facebook and Twitter users. It was found that the average proportion of words expressing positive and negative emotions, as well as their variability and instability of manifestation in the status of each user, can be used as a simple but sensitive measure for diagnosing depression in a social network. In addition, it was found that usefulness of the proposed method may depend on the platform: for Facebook users these features predicted a greater severity of depression, and lower for Twitter.

M. Al-Mosaiwi et al. [10] examined the usage of absolute words (i.e., always, totally, entire) in text writings from various forums devoted to different disorders: depression, anxiety, suicidal ideation, posttraumatic stress disorder, eating disorder, etc. It was found that the number of absolute words in anxiety, depression, and suicidal ideation related forums was significantly greater than in forums from the control group.

Most of the related studies investigate the relationship between mental health and English-speaking social media texts. As an exception, Panicheva et al. [11] and Bogolyubova et al. [12] investigated the relationship between so-called dark triad (Machiavellianism, narcissism, and psychopathy) and Russian texts from Facebook. Using the results of the dark triad questionnaire and profile data of Facebook users the authors conducted a correlation analysis to reveal informative morphological, lexical, and sentiment features.

The study of detecting an early risk of depression based on the experimental task Clef/eRisk 2017 described in the article [13]. The main idea of the task was to classify Reddit users into two groups: the case of depression and non-risk case. The study evaluates the applicability of tf-idf, embeddings, and bigrams models with stylometric and morphological features using Clef/eRisk 2017 dataset and reports 63% of F1-score for depression class.

It should be noted that the use of computational linguistics for analyzing text messages of social networks is mainly limited to lexical approaches. The syntactic-semantic analysis and psycholinguistics markers of the text are still not well evaluated on depression detection task. In this paper, we applied psycholinguistic markers, dictionaries and n-grams models to detect depression in social media texts.

3 Dataset

We asked volunteers from Vkontakte to take part in our psychological research and complete Beck Depression Inventory questionnaire [14]. This questionnaire allows to calculate depression score on 0–63 scale. Before answering questions, users gave access to their public pages under privacy constraints via Vkontakte application. We automatically collected all available information from public personal profile pages using Vkontakte API for the users who completed questionnaire. Posts, comments, information about communities, friends etc. were collected from January 2017 to April 2019 for each user. Overall, information from 1020 profiles were assembled to compile our dataset. All of the personal information that can reveal the identity of persons were removed from data collection.

The scope of our interest were textual messages, namely posts, written in Russian. Therefore, we focused on text messages written by Vkontakte users on their personal profiles and mainly operate with these messages. It is important to note, that social media data contains significant amount of noise and text volume for each user considerably vary from person to person. Before performing on depression detection task, we accurately cleaned the data. First, we applied constrains on required text volume and number of posts. Secondly, we analyzed scores from Beck Depression Inventory and divided our users into 2 groups: persons with score less then 11 were annotated as control group (users without depression signs); persons with score greater than 29 were annotated as depression group (users with depression signs). In this section, we describe these steps and provide statistics on the data. We refer to the data before any changings as *initial data*, to the data after cleaning as *cleaned data*, and to the data after depression risk grouping as *pre-classification data*.

The *initial data* contained information about 1020 persons who took Beck Depression inventory questionnaire. The distribution of the depression marker across users from *initial data* presented in Fig. 1.

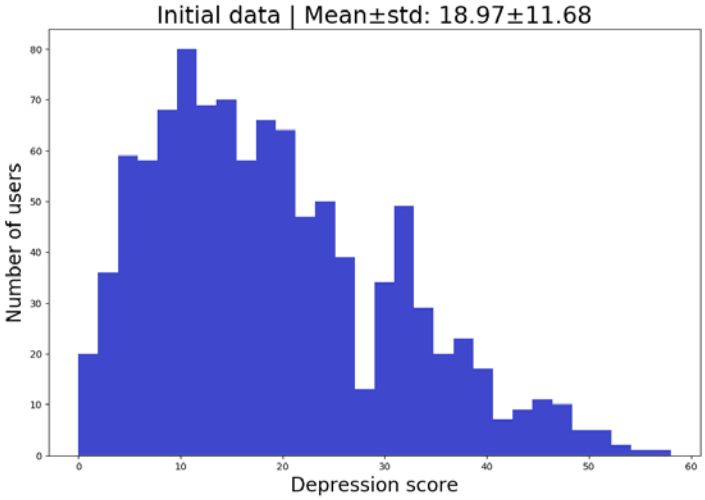


Fig. 1. Depression scores distribution in *initial data*

The mean age in the *initial data* is 25. The gender partition is unbalanced: 699 (68.53%) Females and 321 Males (31.47%). More statistics on the data provided in Table 1. It can be seen from the Table 1, that *initial data* is extremely noisy. Standard deviation values for post, sentence, and word counts are doubled in comparison with their mean values. It was also discovered that 155 of users from dataset did not provide any textual volume. The superficial analysis of the data revealed that data require adjustments and cleaning. As the next step, we performed several actions to adjust the data:

1. Removed all characters which are not alphabet or standard punctuation symbols from texts using regular expressions;
2. Removed all posts with more than 3000 characters or less than 2 words;
3. Removed all users with less than 10 posts or less than 1000 characters provided;
4. Set 100 as the maximum posts count limit for all users.

Applying these steps on the *initial data* yielded 531 user profiles which we annotate as *cleaned data*. We can note, that after data adjustments only 32872 users post left from initial 67.257 posts (see Table 1). We found that limitation of maximum post length is strongly necessary because the manual observation of the data revealed that the most of long posts (more than 3000 characters) were usually not authored by users themselves.

After adjusting steps, the mean depression score slightly decreased, from what we can make the assumption, that persons with higher level of depression write less text than person without depression signs. The gender distribution become even more unbalanced with 397 females (74.76%) and 134 males (25.23%). The histogram of posts count demonstrated on Fig. 2.

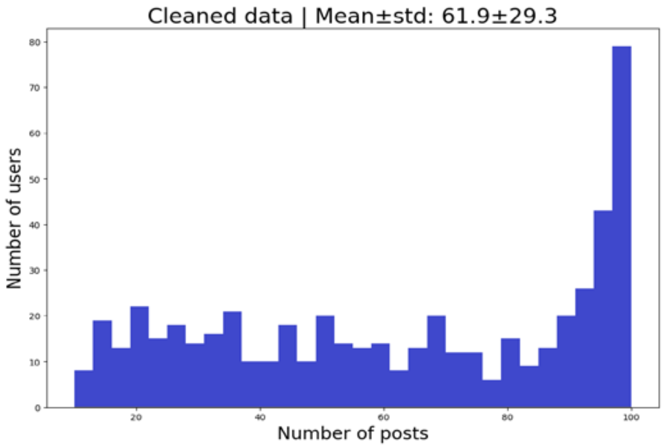


Fig. 12. Posts count distribution in *cleaned data*

After the data cleansing stage, we found this text volume much more suitable for applying natural language processing tools and performing any type of machine learning based evaluation. Anyhow, the depression scores provided by Beck Depression Inventory required some interpretation. We outlined 2 different ways how we can design our research. First one is the regression analysis using raw depression scores, which might be seen as the most appropriate and confident way. But in the other hand, this Russian-speaking social network data is novel, and currently there are no studies related to the depression detection task among Russian-speaking social networks. Most of the English-speaking social networks based depression tasks were designed as a binary classification problem: discover if person depressed or not. To make it possible to compare our results, we decided to perform the similar binary classification task on given data and compare our results with Clef/eRisk 2017 Shared Task [15].

Table 1. Dataset statistics on different data preparation stages. The numbers presented as mean value ± standard deviation

Observed data	*Initial data*	*Cleaned data*	*Pre-classification data*
Number of users	1020	531	248
Males	321 (31.47%)	134 (25.23%)	66 (26.61%)
Females	699 (68.53%)	397 (74.76%)	182 (73.39%)
Age	24.88±6.47	25.99±6.11	25.8±5.69
Depression score	18.97±11.68	17.99±11.04	17.4±15.28
Total number of posts	67257	32872	15238
Avg. posts count	65.93±103.85	61.9±29.3	61.44±29.65
Avg. words count	3114.67±8637.82	1438.01±1244.16	1441.56±1220.59
Avg. sentences count	189.96±492.78	148.98±101.69	148.63±102.81
Words per post	28.75±29.57	22.22±14.42	22.93±15.79
Words per sentence	9.61±4.43	8.98±2.76	9.08±2.72
Sentences per post	2.66±1.96	2.31±0.99	2.34±1.08

Table 2. Statistics between depression and control group

Group	*Depression group*	*Control group*
Number of users	92 (35.65%)	156(60.46%)
Males	19 (20.65%)	47 (30.12%)
Females	73 (79.34%)	109 (69.87%)
Age	25.67±6.43	25.87±5.21
Depression score	36.44±6.37	6.17±2.75
Total number of posts	5268	9970
Avg. posts count	57.26±30.13	63.91±29.07
Avg. words count	1328.15±1271.14	1508.44±1184.7
Avg. sentences count	138.61±113.08	154.54±95.75
Avg. Words per post	22.74±18.58	23.04±13.89
Avg. Words per sentence	8.88±2.71	9.19±2.72
Avg. Sentences per post	2.34±1.32	2.35±0.91

As the next step, we observed depression scores and discovered that we cannot simply divide our data by setting boarder value and annotating all users with depression score above this value as a risk group of depression and all user with the depression score bellow boarder value as a non-risk group. In order to form the *pre-classification data*, we annotated all persons with depression score less than 11 as non-risk group (control group). For a risk group we assembled the data of persons with depression scores above 29 (depression group). These values were discussed and proposed by the psychologist experts related to our study. The persons with depression score between these values were removed from observation.

Performing this step reduced the data population to the 248 users, where 156 were labeled as control group (without depression signs) and 92 users were labeled as belonging to the depression group. The general statistics on *pre-classification data* also presented in Table 1. The statistics between groups on the *pre-classification data* presented in Table 2.

It can be observed from Table 2 that users from depression group tends to write lesser amount of text in the Vkontakte social media. The values of average posts count, average words count, average sentence count are less than in the control groups. The length of posts and sentences are greater in control group. The gender partition is even more biased towards female in depression group.

4 Features and Methods

Before forming the feature sets, all user posts were concatenated into the one text for every user in dataset. We retrieved four groups of features from texts: morphological, syntactic, sentiment and psycholinguistic. We applied MyStem [16] for tokenization, lemmatization, and part-of-speech tagging, and Udpipe [17] for syntax parsing. The sentiment features were calculated using Linis-Crowd sentiment dictionary [18].

Psycholinguistic markers are linguistic features of text that represent psychological characteristics of author and may signal about his psychological disorders. For example, people in stress more frequently use in text pronoun “we” [19]. Psycholinguistic markers are calculated on morphological and syntactic information and in a manner correspond to the writing style of the author. We use more than 30 markers and the most significant of them are the following:

- Mean number of words per sentence;
- Mean number of characters per word;
- (N punctuation characters) / (N words);
- Lexicon: (N unique words) / (N words);
- Average syntax tree depth;
- (N verbs) / (N adjectives);
- (N conjunctions + N prepositions) / (N sentences);
- (N infinitives) / (N verbs);
- (N singular first person past tense verbs) / (N verbs);
- (N first person verbs) / (N verbs);
- (N third person verbs) / (N verbs);
- (N first person pronouns) / (N pronouns);
- (N singular first person pronouns) / (N pronouns);
- (N plural first person pronouns) / (N pronouns).

These psycholinguistic markers were previously utilized for the task of predicting depression from essay in Russian. They are described in more details at [20]. We extend psycholinguistic markers set with postags ratio and following social network specific features: uppercase characters ratio, average number of Vkontakte links per post, number of exclamation marks, number of “sad” and “happy” smiles.

We also formed two n-grams sets: tf-idf matrix computed on the unigrams and tf-idf matrix computed on the both unigrams and bigrams combined. N-grams that appeared less than in 1% of texts were removed from the feature sets. User's lexicon formed while tf-idf set preparation were extremely poor with 5742 unique tokens for unigrams and 10909 unique tokens for both unigrams and bigrams combined. We relate this fact to the specific of social network language. The writings contain a lot of slang and words with wrong spelling.

Another feature set were retrieved by utilizing dictionaries which was used for the task of detection verbal aggression in social media writings [21]. It is containing following dictionaries: negative emotional words, lexis of suffering, positive emotional words, absolute and intensifying terms, motivation and stressful words, invectives, etc. To calculate features, for every user we calculate the occurrences of words from different dictionaries in user's writings and divide this number on total user's words count.

As it was mentioned before, we designed depression detection task as binary classification. We evaluate 4 different sets of features: psycholinguistic markers (PM), unigrams (UG), bigrams (BG), and dictionaries(D).

5 Results of Experiments

To perform on the task, we utilized scikit-learn machine learning library [22]. Random forest and support-vector machines (SVM) models were used to perform evaluation on the data. All of the feature's sets were normalized and scaled. Hyperparameters of the classification algorithms were tuned by grid-search runs.

Table 3. Classification report

Dummy classifier					
Set	***Precision***	***Recall***	***F1***	***ROC AUC***	***F1-w***
-	45.23±2.38	30.43±8.13	36.0±8.48	48.29±3.76	50.48±6.17
Random Forest					
Set	***Precision***	***Recall***	***F1***	***ROC AUC***	***F1-w***
PM	59.80±6.21	59.80±6.21	54.47±3.66	70.91±6.81	67.98±3.03
UG	51.68±9.89	57.17±3.70	53.84±6.35	64.59±3.79	63.03±8.29
BG	49.64±6.67	58.47±6.06	53.12±3.16	63.18±2.68	61.65±5.95
D	46.21±5.52	56.30±7.20	50.66±5.80	58.07±6.33	59.90±4.88
PM-r	**62.60**±7.77	53.26±7.88	56.59±2.20	74.89±4.05	69.16±2.60
SVM					
Set	***Precision***	***Recall***	***F1***	***ROC AUC***	***F1-w***
PM	55.43±1.99	72.82±1.88	62.92±1.51	71.12±4.46	68.66±1.72
UG	45.63±7.94	83.69±13.53	57.57±3.41	67.72±3.61	49.79±13.77
BG	44.38±6.07	**85.86**±11.24	57.60±2.76	66.88±2.64	47.72±14.90
D	55.68±9.49	55.43±8.34	55.53±8.85	63.57±7.85	66.94±6.89
PM-r	58.40±2.99	77.17±1.88	**66.40**±1.33	**75.11**±3.24	**71.42**±2.21

Since the depression detection task is previously untested on the Russian-speaking social media data, we also demonstrate the accuracy yielded by random based dummy classifier. The metrics for evaluation is weighted mean F1-score of both control and depression group (F1-w) and ROC AUC score. To make it possible to compare our results with Clef/eRisk 2017 Shared Task results, we also demonstrate precision, recall and F1-score for depression class only. The evaluation metrics were calculated by averaging 5 runs of 4-folds cross-validation. The classification results presented in Table 3.

The evaluation revealed that Psycholinguistic markers performed well on the data. We initially assumed that some of the psycholinguistic markers could work poorly on the data because users usually write very short texts and the volume of concatenated posts cannot be compared to a logically connected text of the same size. This constrains are important for the specific of some psycholinguistic markers. We analyzed feature importance from several Random Forest runs in order to reduce the size of the PM feature vector which can possibly improve classification performance. The reduced version of PM (PM-r) was included in classification report.

The best result on the data yielded by SVM+PM-r model with 75.11% ROC AUC score, 71.42% weighted F1-score and 66.40% F1-score on depression class. The same feature set with the Random Forest algorithm also achieved decent results with 74.89% ROC AUC score and highest precision (62.60%) in our experiments.

The dictionaries based set demonstrated poor results in comparison with other sets. In other hand, considering the general complexity of the depression detection task these dictionaries demonstrated some positive results. These dictionaries should be redesigned and filtered which can make them useful as additional features for PM set.

The surprising result in our experiments is that n-gram and tf-idf based features did not perform well on the data. As it was mentioned before, we relate this fact to the great amount of slang, wrong spelling and another noise in social media language. We should focus this problem by applying term clustering. For example, we can use words embeddings as it was implemented in this work [13].

It should be noted, that we can compare our results with the results of Clef/eRisk 2017 Shared Task evaluation only with some restrictions. First, language of Clef/eRisk 2017 was English, while our data is in Russian. Secondly, the number of data samples and class ratio is different. Finally, depression class in Clef/eRisk 2017 Shared Task was assembled by manual expert examination of profiles from subforum related to the depression disorder. In our study, we operate only with the Beck Depression Inventory scores.

Despites this facts, best F1-score reported in Clef/eRisk 2017 overview [23] was 64% achieved by the model that utilized tf-idf based features on the data with LIWC and dictionary features. In our experiments tf-idf based features demonstrated 57.60% of F1-score with *SVM+BG* model. It is important to mention, that current state-of-art result on Clef/eRisk 2017 data is 73% of F1-score [24]. The best depression detection performance on our Vkontakte data is 66% of F1-score achieved by filtered version of psycholinguistic markers.

6 Conclusion

In the study we performed depression detection task among 1020 users of Russian-speaking social network Vkontakte based on their text messages. By analyzing Beck

Depression Inventory scores and processing the initial data we formed the sample of 248 users' posts collections with binary depression/control group labeling. We formed tf-idf and dictionary based feature sets and retrieved novel psycholinguistic features from users' writings. The experiments were performed using SVM and Random Forest classifiers and results were compared with Clef/eRisk 2017 Shared Task evaluation. The best result in our experiments is 66.40% of F1-score (75.11% of ROC AUC score) achieved by model that based on filtered psycholinguistic markers.

It was discovered that psycholinguistic markers performed well on the data and can be effectively utilized for the depression detection task. We found that Vkontakte textual data is extremely noisy which is resulted in the relatively poor classification results achieved by tf-idf based models. We assume that term clustering methods could improve performance of n-grams models. It is also clear, that dictionaries that we used for feature set should be redesigned and filtered.

Thus, the analysis of depression linguistic markers in social network posts is a promising area that can possibly make the prevention and treatment of depression more accessible to a large number of users. In the future work we planning to examine neural network models for the depression detection task and evaluate regression analysis on the data using Beck Depression Inventory scores.

Acknowledgments

This work was financially supported by the Ministry of Education and Science of the Russian Federation. Grant No. 14.604.21.0194 (Unique Project Identifier RFMEFI60417X0194).

References

1. Turecki, G. and Brent, D.A.: Suicide and suicidal behaviour. The Lancet **387** (10024), 1227–1239 (2016).
2. World Health Organization. https://www.who.int/mental_health/prevention/ suicide/suicideprevent/en/, last accessed 2019/08/19
3. Belialov, F.I.: Depression, anxiety, stress, and mortality. Terapevticheskii arkhiv **88** (12), 116–119 (2016).
4. Surtees, P.G., Wainwright, N.W., Luben, R.N., Wareham, N.J., Bingham, S.A., and Khaw, K.T.: Depression and ischemic heart disease mortality: evidence from the EPIC-Norfolk United Kingdom prospective cohort study. American Journal of Psychiatry **165** (4), 515–523 (2008).
5. Whang, W., Kubzansky, L.D., Kawachi, I., Rexrode, K.M., Kroenke, C.H., Glynn, R.J., and Albert, C.M.: Depression and risk of sudden cardiac death and coronary heart disease in women: results from the Nurses' Health Study. Journal of the American College of Cardiology **53** (11), 950–958 (2009).
6. Tausczik, Y.R. and Pennebaker, J.W.: The psychological meaning of words: LIWC and computerized text analysis methods. Journal of language and social psychology **29** (1), 24–54 (2010).
7. Kailer, A. and Chung, C.K.: The Russian LIWC2007 dictionary. Austin, TX: LIWC.net (2011).
8. Yates, A., Cohan, A., and Goharian, N.: Depression and self-harm risk assessment in online forums. arXiv preprint arXiv:1709.01848 (2017).

9. Seabrook, E.M., Kern, M.L., Fulcher, B.D., and Rickard, N.S.: Predicting depression from language-based emotion dynamics: longitudinal analysis of Facebook and Twitter status updates. Journal of Medical Internet Research **20** (5), e168 (2018).
10. Al-Mosaiwi, M. and Johnstone, T.: In an absolute state: Elevated use of absolutist words is a marker specific to anxiety, depression, and suicidal ideation. Clinical Psychological Science **6** (4), 529–542 (2018).
11. Panicheva, P., Ledovaya, Y., and Bogolyubova, O.: Lexical, morphological and semantic correlates of the dark triad personality traits in russian facebook texts. In 2016 IEEE Artificial Intelligence and Natural Language Conference (AINL) (pp. 1–8). IEEE (2016, November).
12. Bogolyubova, O., Panicheva, P., Tikhonov, R., Ivanov, V., and Ledovaya, Y.: Dark personalities on Facebook: Harmful online behaviors and. Computers in Human Behavior **78**, 151e159 (2018).
13. Stankevich, M., Isakov, V., Devyatkin, D., and Smirnov, I.: Feature Engineering for Depression Detection in Social Media. In ICPRAM, 426–431 (2018).
14. Beck, A.T., Steer, R.A., and Brown, G.K. Beck depression inventory-II. San Antonio **78** (2), 490–498 (1996).
15. Losada, D.E. and Crestani, F.: A test collection for research on depression and language use. In International Conference of the Cross-Language Evaluation Forum for European Languages, 28–39. Springer, Cham (2016, September).
16. MyStem Homepage, https://tech.yandex.ru/mystem, last accessed 2019/08/19
17. Straka, M. and Straková, J. Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with udpipe. In Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, 88–99 (2017, August).
18. Koltsova, O.Y., Alexeeva, S., and Kolcov, S.: An opinion word lexicon and a training dataset for russian sentiment analysis of social media. Computational Linguistics and Intellectual Technologies: Materials of DIALOGUE 2016 (Moscow), 277–287 (2016).
19. Pennebaker, J.W. The secret life of pronouns. New Scientist **211** (2828), 42–45 (2011).
20. Stankevich, M., Smirnov, I., Kuznetsova, Y., Kiselnikova, N., and Enikolopov, S.: Predicting Depression from Essays in Russian. Computational Linguistics and Intellectual Technologies, DIALOGUE **18**, 637–647 (2019).
21. Devyatkin, D., Kuznetsova, Y., Chudova, N., and Shvets A.: Intellectual analysis of the manifestations of verbal aggressiveness in the texts of network communities [Intellektuanyj analiz proyavlenij verbalnoj agressivnosti v tekstah setevyh soobshchestv]. Artificial Intelligence and Decision Making, (2), 27–41 (2014).
22. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., and Vanderplas, J.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, (12), 2825–2830 (2011).
23. Losada, D.E., Crestani, F., and Parapar, J.: CLEF 2017 eRisk Overview: Early Risk Prediction on the Internet: Experimental Foundations. In CLEF (Working Notes) (2017).
24. Trotzek, M., Koitka, S., and Friedrich, C.M.: Utilizing neural networks and linguistic metadata for early detection of depression indications in text sequences. IEEE Transactions on Knowledge and Data Engineering (2018).

Development and Implementation of the Algorithm for Automatic Analysis of Metrorhythmic Characteristics of Russian Poetic Texts

V. B. Barakhnin[1], O. Yu. Kozhemyakina[1], and I. V. Kuznetsova[2]

[1] Institute of Computational Technologies SB RAS, Novosibirsk, Russia
[2] Novosibirsk State University, Novosibirsk, Russia
bar@ict.nsc.ru

Abstract. This paper presents the implementation of the program module responsible for the analysis of the structural level – the definition of the poem's metrorhythmics (meter, number of feet, and rhyme) in Russian poetic texts. The algorithm for determination of meter and number of feet takes into account the problem of the ambiguity of the placement of emphasis in homographs, possible omissions of schematic emphasis (pyrrhic), overlay of over schematic emphasis (spondee), which are solved by method "by analogy". The algorithm to identify the cases of shifting the emphasis from one part of speech to another (proclitic) is described. The algorithm of rhymes search is presented, the result of which is the definition of the stanzas of the poem.

Keywords: analysis of poetic texts, meter definition, metrorhythmic analysis, rhyme identification.

1 Introduction

The compilation of metric reference books to the corpus of poems is an important task of Russian literary researches. At the moment, the development of information technologies allows to automate the analysis of poetic texts, which in turn will reduce the amount of routine work of philologists. To solve this problem, it is necessary to develop algorithms for automating the analysis of the structural level of the poetic text, including the following metrorhythmic characteristics: meter, number of feet (foot – the basic unit of measurement of accentual-syllabic meter), rhyme.

Among similar works for other languages the system "SPARSAR" [1] can be identified, which provides automatic analysis of poetic texts in English and Italian. This system performs analysis at the quantitative, syntactic and semantic levels using NLP (Natural language processing). The obtained data are visualized in the form of developed schemes that allow comparing the works of one poet with each other and the works of different poets. The works of William Shakespeare, Thomas S. Eliot and Sylvia Plaza were used as a test sample. The program determined the size of their works with an accuracy of 90%. The system analyzed 500 poems, then the expert checked the accuracy of the analysis of 50 randomly selected poems. 5% of errors were detected.

In work [2] the static methods for the analysis, generation and translation of rhythmic poetry were used. The authors used machine learning without a teacher (unsupervised learning) to identify the patterns of stress placement in the number of poems to supplement in doubtful cases the proposed rhyme model. The authors also conducted experiments on the generation of English-language love lyrics and translations of Italian poetry into English with the preservation of the desired rhythmic scheme. The authors used 5 Shakespeare sonnets (70 lines) as a test sample; 81.4% of lines (57 lines) were correctly classified by metrorhythmics.

In [3], the authors have classified the texts according to the meter. They used an open source “the Scandroid” [4] to extract features (the program contains an algorithm for determining the stress in words and its own dictionary with accentuation of words – exceptions) and machine learning to classify poems by meter. Also in this work was implemented a module that defines the rhyme using the dictionary “The Carnegie Melon University Pronouncing Dictionary” [5] as a source of information about the pronunciation. To determine the accuracy of the work the sample of 205 poems was used, 88% of the words were correctly divided into syllables, the number of feet was determined with an accuracy of 99%.

In [6] the corpus of poems was analyzed using the “connectionist model” of poetic meter. It was shown that the prosodic picture of the poetic text is individual, and it is possible to determine the author by it, as well as that it reflects the aesthetics of the period of the author's creative work. As a test sample, a number of 1000 lines (100 lines by ten different authors) was used. The software package for statistical data processing SPSS Statistics was applied for the analysis [7].

In the work [8] it is described how to create software to determine the quantitative characteristics of the style of American poets and the visualization of a collection of poems in relation to each other. To visualize the obtained metrics, the authors used the principal component analysis and Classical Multidimensional Scaling.

It should be noted that it is impossible to design a universal system of automatic analysis of meter and rhythm, suitable at least for a group of more or less similar languages, because each language requires the development of its approaches, taking into account its structure. Such an experiment was conducted for similar in structure (Latin and Greek) languages, but even in this case, the study revealed features of languages, because of which their joint analysis is impossible [9, p. 52–54].

Finally, although in the analysis of the lower levels of the structure of the Russian verse the simplest mathematical approaches have been used for a long time – for example, numerous studies of the statistics of the types of Russian rhyme (including those applied to the temporal dynamics), generalized in [10], but often the collection of statistical information is still carried out almost manually (except for content analysis).

Some studies describing an integrated approach to automating the characteristics of Russian poetic texts (for example, [10]), affect, as a rule, very specific genres of poetry – for example, folk poetry, structural characteristics of which, such as metric, themes, etc., are significantly different from the corresponding structures in “literary” verse, or are of a rather private nature: the quantitative analysis of semantic associations of pentameter on the material of a number of Russian poets studied by M. L. Gasparov [11], the effect of metrorhythmic on semantics in the work of I. A. Brodsky – by M. Yu. Lotman [12], the metric halo's of "Black shawl" by A. S. Pushkin – by M. Wachtel [13], etc. The similar studies in relation to Czech poetry were conducted by M. Chervenka (see, for example, [14]).

For the analysis of the structural level of Russian poetic texts there are no practically implemented systems (at least in the open access), except for the pilot project of the system [15, 16], developed at the Institute of Computing Technologies of SB RAS.

The algorithm from [17] is at the core of this system, but it has several disadvantages, for example, does not take into account unequal meter of the poem, while in syllabic-tonic versification, there are meters, like the free amphibrach, choree, pentameter, characterized by a different number of feet in lines. In addition, the usage of the algorithm in its "pure" form, without taking into account the possible ambiguities of automatic accentuation, gives a relatively high percentage of errors in determining of the number of feet.

For these reasons, it was decided to implement the algorithm from [18], which includes a more strict classification of poems by meter. However, it does not affect such problems as ambiguous accentuation of homographs and clitics, so it also needs some modification. This study describes the development and implementation of the algorithm for analyzing the structural level: meter, number of feet, and rhyme; the modifications of the algorithm [18], features of its implementation and results are presented. The results of the work of two systems are compared: the one, which is developed on the basis of a modified algorithm from [18], and the one which is already existing on the basis of the algorithm from [17].

2 The Algorithm for Determining the Meter and Number of Feet

Let's describe the steps of the algorithm for determining the meter and number of feet, presented in [18]. The essence of this algorithm is to compare the rhythmic variants of the verse of the studied poetic text with a set of rhythmic patterns from a certain repertoire of metrorhythmic variants of the verse. This algorithm consists of five steps:

1. The pre-processing of a text. The lines of poetic text are numbered (St(n)), and PT={St(n)}, n=1, 2, …, N, where N is the total number of lines. All punctuation marks are deleted.
2. The accentuation is carried out on the basis of the dictionary A.A. Zaliznyak [19], which contains the accented word forms.
3. Each word of poetic text is divided into syllables and translated into a sequence of characters "*c*" and "*C*", denoting unstressed and stressed syllables respectively. The spaces between words are removed to display the syllabic scheme Sl:

$$\mathrm{Sl} =$$

$$= c_1 \dots c_{m(0)} C c_1 \dots c_{m(1)} \dots C_i c_1 \dots c_{m(i)} \dots C_{k-1} c_1 \dots c_{m(k-1)} C_k c_1 \dots c_{m(k)}, \quad (1)$$

where $c_{m(i)}$ is the unstressed syllable of the *i*-th word, *i*∈[0,*k*], *Ci* is the stressed syllable of the *i*-th word at *i*∈[0, *k*], *k* is the number of stressed syllables;

4. Scheme (1) is converted into syllabic rhythmic scheme Rs:

$$\text{Rs} = c^{r(0)} C_1 C^{r(1)} \dots C_i c^{r(i)} C_{i+1} \dots C_{k-1} c^{r(k-1)} C_k c^{r(k)}, \quad (2)$$

where *k* is the number of stressed syllables in the string, *R* is the number of all syllables in a line, *r(i)* is the interaccent interval, where *i*∈[1,*k*–1], *r(0)* and *r(k)* is the anacrusis and the clause. After that, the parameters *k, r(i), R-r(k)* are extracted, what further determine the type of rhythmic scheme of a poem.

5. The selection of the terms of classification of poetic text by metrorhythmic based on existing principles of versification. Depending on whether the parameters *k, r(i), R-r(k)* take constant or non-constant values for different foot lines, the authors formulate 5 classification conditions that can be implemented in Russian versification (Table 1).

Table 1. Classification of poetic text by metrorhythmic schemes

	R-r(k)	*k*	*r(i)*	Classification
1	≠const	≠const	arbitrarily	1. Metric verse in unequal feet 2. Dismetric verse
2	≠const	=const	arbitrarily	*k* - accentual verse in unequal/equal feet
3	≠const	=const	=const	*k* - foot syllable-tonic verse in equal feet
4	=const	=const	=const	*k* - foot syllable-tonic verse in strict equal feet
5	=const	≠const	≠const	isosyllabic verse in equal feet (proclitic, pyrrhic)

The first condition describes the often encountered type – syllabic-tonic versification with violation of the number of feet in the line – free syllabic-tonic versification. The second and third conditions of classification describe verses with specific metrorhythmics, and in this work will not be considered. The fourth condition classifies the verses with ideal meter without breaking the rhythm caused by omissions of schematic emphasis or overlay of over schematic emphasis. This kind of metrorhythmics are rare, because most poems contain alternating male and female rhymes, which automatically violate the condition of the constancy of the parameter *R - r(k)* (see Table 1). Most poetic texts contain disturbances in the rhythm; and are described in the algorithm of the fifth condition for classification as isosyllabic poems.

In the framework of this research the implementation and testing of the first and fifth conditions according to the classification of poetic text in metrorhythmic schemas was carried out.

3 Modification of the Meter and Number of Feet Detection Algorithm

This paper proposes a number of modifications to optimize the algorithm to improve the accuracy in the analysis of poetic texts.

Modification 1. The algorithm [18] assumes an "ideal" accentuation of words and completely ignores the existence of problems related to the omissions of schematic emphasis (pyrrhic). An example of pyrrhic can be shown in the quatrain of "Eugene Onegin":

Мой дядя самых честных правил,
Когда не в шутку занемог,
Он уважать себя заставил
И лучше выдумать не мог.

That quatrain will translate into:

cC cC cC cC c
cC cC cc cC
cc cC cC cC c

Only the first line of the scheme is strictly maintained (iambic tetrameter), and in others there are three accents, the one is missing. In the case of missing of the metric stress there is a special auxiliary foot of two unstressed syllables – pyrrhic (cc), which can replace the foot of iamb and of choree (e.g «Нет, не черкешенка она» from the "Answer to F.T." by A.S. Pushkin).

The algorithm [3] does not consider the overlay of over schematic emphasis (spondee), an example of which is illustrated by line «Швед, русский, колет, рубит, режет» from the "Poltava" by A.S. Pushkin with the scheme: cc cC cC cCc, the transfer of the emphasis from one part of speech to another (proclitic: «уронили мишку нА пол») and homographs («чЕстных», «честнЫх»).

These problems are solved by the method "by analogy", the idea of which in relation to this problem was expressed in [20]. The essence of the method is the following: lines and stanzas with ambiguous accent arrangement are compared with lines and stanzas, in the words of which the stress is placed unambiguously, and the choice of accent is made, providing the unity of metric characteristics for the whole poem.

To implement this method, when the poetic text is translated into a sequence of characters "c" and "C" (denoting unstressed and stressed syllables, respectively) in words with an ambiguous arrangement of accent, the positions of all possible variants of the accent are denoted by the symbol "x". Thus, the text is represented as a table of characters "C", "c", "x" of dimension n (the number of lines of poetic text) on m (the line with the maximum length).

Further, to eliminate the ambiguity of the accent arrangement in each line of the table, the element "x" is searched and the column of the table is taken by the index of this element. In this column, the most common single element is looked for and its value is assigned to the element "x".

The algorithm from [18] does not consider the proclitics (the pulling the accent on the preposition, for example, in the poem "Teddy Bear" by A. Barto – «уронили мишку нА пол»), so the database of proclitics on the basis of the dictionary of A. Zaliznyak [19] has compiled, To resolve the ambiguities associated with proclitics. It contains the information on the variants of accentuation of combinations of some words and prepositions. The text is analyzed for the presence of prepositions. If there is a preposition in the text, then the search for a combination of this preposition in conjunction with the word standing to the right of it is carried out. Upon detection of this combination in the database of proclitics, the information about the variants of accentuations in this combination is retrieved. In the case of ambiguous variants of the arrangement of accents, we again resort to the method "by analogy".

Modification 2. The algorithm from [18] is sensitive to the parameters which it receives (*R*–*r*(*k*), *k*, *r*(*i*)), what leads to an incorrect definition of the meter and number of feet. Therefore, for the parameters the inaccuracies have been introduced: the parameters considered to be constant (=const), if they are constant for at least 90% of the lines of the poem.

Modification 3. The step of the algorithm from [18] is worked out in detail, specifically the fifth condition, the classification of poems according to metrorhythmic, which takes into account pyrrhic and spondee (*R*–*r(k)*=const, *k*≠const, *r(i)*≠const). If the text satisfies this condition, then further clarification to determine the meter and number of feet of the poetic text is made as follows. After each word is divided into syllables and translated into a sequence of characters "c" and "C", the syllabic pattern is compared to the pre-compiled patterns, which are the foot of the intended meter. Using statistical evaluation the most suitable pattern from which the most suitable meter follows is revealed, that is the pattern with minimal difference from the syllabic scheme corresponds to a certain meter.

4 Approach to the Implementation of the Module Definition of the Rhyming Lines

In the article [18], in the algorithm for determining rhyme of poetic text it is suggested to seek rhyming lines in a poem, using the web app "Big rhyming dictionary" [21]. This app takes a word and returns a set of words rhyming with it.

Because the sending of requests to the web application for each word is a long process and the extraction of all the words and sets, and rhyming words to them with their following conservation in the database, is the process requiring big resources, in this paper we used an alternative way to search rhymed lines.

The rhyme search algorithm is implemented for reasons of the possibility of rhyme formation: the lines are rhymed if the last words in the line have the same position of the stressed syllable and the endings are phonetically coincided.

To identify the phonetically matched endings, we used data about the endings from the article [22]. It contains a pair of letter combinations, reflecting the sounds of rhyming verse endings from the literature of the 18–19 century:
[('и', 'ы'), ('и', 'ый'), ('ы', 'ый'), ('и', 'е'), ('и', 'ий'), ('у', 'уй'), ('ой', 'о'), ('кий', 'ки'), ('ей', 'е'), ('ай', 'о'), ('ой', 'а'), ('ей', 'и'), ('ий', 'е'), ('и', 'ьи'), ('и', 'ья'), ('ьи', 'ья'), ('и', 'ье'), ('е', 'ье'), ('к', 'г'), ('х', 'к',), ('г', 'х'), ('а', 'о'), ('е', 'и'), ('ья', 'ье'), ('ьи', 'ье'), ('ом', 'ым'),

('ит', 'ет'), ('ин', 'ен'), ('ий', 'а'), ('ой', 'а'), ('ый', 'а'), ('о', 'у'), ('уг', 'ок'), ('ах', 'ых'), ('е', 'ы'), ('ив', 'ов'), ('и', 'ой'), ('и', 'а'), ('я', 'и',), ('а', 'ы'), ('ы', 'у'), ('я', 'е'), ('ы', 'о'), ('ый', 'о'), ('ы', 'ой'), ('у', 'ой'), ('у', 'ый'), ('ы', 'ей'), ('ешь', 'ишь'), ('он', 'ен'), ('ел', 'ол'), ('ей', 'ой'), ('ом', 'ем'), ('ть', 'дь'), ('д', 'т'), ('ор', 'ер'), ('ом', 'им')]

With their usage, the algorithm for determining rhyme was developed, which consists of the following sequential steps:

1. The splitting of the text into stanzas.
2. The line numbering and extracting the last word from each line of the stanza.
3. The definition of accentuation for each word. The search for sets of words with the same accentuation.
4. The search for sets of words whose endings are rhymed (phonetically matched)
5. The search for the intersection of sets from p. 4 and 5 – the obtained rhyming lines.
6. The representation of rhymes in letter form: each set of rhymed lines is assigned a letter of the Latin alphabet, and the male – a lowercase letter, the female – a capital letter. On the basis of the received designations the letter sequence of lines in a stanza is made.

5 Obtained Results

As a test sample we have used a corpus of lyrical works of Alexander Pushkin (156 poems from the period of 1818–1825), pre-marked by meter, foot, rhyme with the help of reference book [23]. The algorithms for the definition of the meter and number of feet were tested (we modified the algorithm from [18] and implemented the algorithm from [17]).

Table 2. The comparing of algorithms for determining the metrorhythmic characteristics

Modified algorithm from [18]		Algorithm from [17]	
Meter	Number of feet	Meter	Number of feet
95,5%	95,5%	66,6%	57%

The accuracy of the determination of meter and number of feet in the modified algorithm [18] increased in comparison with the algorithm from [17] (Table 2). The difference in the work of the algorithms is due mainly to the fact that the first of the presented algorithms recognizes the metric versification with unequal feet, while the second incorrectly classifies it. Also, the sensitive parameters of the first algorithm made it possible to increase the accuracy in determining the number of feet in comparison with the already existing algorithm.

Among the disadvantages of this algorithm for determining the meter and stop we can underline: the limited database compiled on the basis of the dictionary A. Zaliznyak (a lack of some words), the replacing the letter “ё” by “е” in the test sample, what is allowed by the rules of the Russian language, but is critical for this algorithm, these problems are partially offset by the usage of the method “by analogy”. Further researches will address these shortcomings.

The main percentage of errors is due to the fact that there are unfinished poems in the corpus, in which one or more words are missing, they are replaced by supplemented ones, hereupon the standard algorithm that provides a full line does not work correctly. The further purpose of the study will be the identification of such lines to be excluded from the General analysis.

Also, this algorithm does not distinguish a meter with unequal feet and complex meter. This will also be done in a further study.

The module of the rhyme definition in modified algorithm of [18] has identified correctly 95% of the strophic patterns.

6 Conclusions

The article compares the work of two algorithms for the analysis of the structural level of Russian poetic texts: a relatively simple one from [17], which was used by us in the implementation of the pilot project of the system [16], and a more advanced algorithm from [18], modified by us, first of all, in order to eliminate possible ambiguities of automatic accentuation. It is shown that the modified algorithm from [18] gives a higher accuracy in determining the meter and number of feet, and also determines the strophic pattern with an accuracy of 95 %.

Further researches will be aimed at identifying typical problem situations that generate the errors in the work of the algorithm, and their subsequent elimination.

Acknowledgements

The study was carried with the support of the Russian Science Foundation (project No. 19-18-00466).

References

1. Delmonte, R.: A Computational approach to poetic structure, rhythm and rhyme. In: Proceedings of the First Italian Conference on Computational Linguistic in Pisa University Press, Vol. 1, P. 144–150 (2014).
2. Greene, E., Bodrumlu, T., and Knight, K.: Automatic analysis of rhythmic poetry with applications to generation and translation. In: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. P. 524–533 (2010).
3. Tanasescu, C., Paget, B., and Inkpen, D.: Automatic classification of poetry by meter and rhyme. In: Proceedings of AAAI 2016. University of Ottawa (2016). Available at: https://www.aaai.org/ocs/index.php/FLAIRS/FLAIRS16/paper/viewFile/12923/12883 (Date accessed 06.07.2019).
4. The Scandroid. Available at: http://charlesohartman.com/verse/scandroid/index.php (Date accessed 06.07.2019).
5. The Carnegie Melon University Pronouncing Dictionary. Available at: http://www.speech.cs.cmu.edu/cgi-bin/cmudict (Date accessed 06.07.2019).
6. Hayward, M.: Analysis of a corpus of poetry by a connectionist model of poetic meter. Poetics **24** (1), 1–11 (1996).

7. IBM SPSS Statistics. Available at: https://www.ibm.com/ru-ru/products/spss-statistics (Date accessed 06.07.2019).
8. Kaplan, D.: Computational analysis and visualized comparison of style in American poetry. Unpublished undergraduate thesis (2006). Available at: https://faculty.missouri.edu/~kaplandm/pdfs/KaplanBlei2007_ComputationalPoetryStyle_long.pdf (Date accessed 06.07.2019).
9. Mittmann, A.: Escansão automático de versos em português. Tesis (Doctorado), Universidade Federal de Santa Catarina. (2016).
10. Samoilov, D.: Book about Russian rhyme. M.: Hudozhestvennaya literatura (1982) (in Russian).
11. Gasparov, M.L.: Meter and meaning. About one of the mechanisms of cultural memory. M.: RSUH (1999) (in Russian).
12. Lotman, M.: "On the death of Zhukov" (1974). How does the poem of Brodsky work. In: The researchs of the Slavists in the West. M.: Novoe literaturnoe obozrenie. P. 64–76 (2002) (in Russian).
13. Wakhtel, M.: "Black shawl" and its metric halo. In: Russian verse: metric, rhythm, rhyme, stanza. M.: RSUH. P. 61–80 (1996) (in Russian).
14. Chervenka, M.: Meaning and verse. In: Works on poetics. M.: Yazyki slavyanskoy kultury (2011) (in Russian).
15. Barakhnin, V. and Kozhemyakina, O.: About the automation of the complex analysis of Russian poetic text. CEUR Workshop Proceedings **934**, 167–171 (2012) (in Russian).
16. Analysis of the poetic texts online. Available at: http://poem.ict.nsc.ru/ (Date accessed 06.07.2019) (in Russian).
17. Kozmin, A.V.: Automatic verse analysis in Starling system. Computer linguistics and intellectual technologies: Proceedings of the international conference "Dialogue 2006". M.: Publishing center of RSUH. P. 265–268 (2006) (in Russian).
18. Boikov, V.N., Karyaeva, M.S., Sokolov, V.A., and Pilschikov, A.I.: About automatic specification of the verse in the information-analytical system. CEUR Workshop Proceedings **1563**, 144–151 (2015) (in Russian).
19. Zaliznyak, A.A.: Grammatical dictionary of the Russian language. The changing word forms: about 10,000 words. 2-e ed. M.: Russian language (1980) (in Russian).
20. Barakhnin, V.B., Kozhemyakina, O.Yu., and Zabaykin, A.V.: The algorithms of complex analysis of Russian poetic texts for the purpose of automation of the process of creation of metric reference books and concordances. CEUR Workshop Proceedings **1536**, 138–143 (2015) (in Russian).
21. Big rhyming dictionary. Available at: http://rifmovnik.ru/docs.htm (Date accessed 06.07.2019) (in Russian).
22. Zhirmunsky, V.M.: Rhyme, its history and theory. Petrograd: Academia (1923) (in Russian).
23. Lapshina, N.V., Romanovich, I.K., and Yarkho, V.I.: Metrical Handbook for Pushkin's poems. M.; L.: Academia (1934) (in Russian).

Towards a Semantically Annotated Corpus of Educational Mathematical Texts in Russian

Olga Nevzorova[1,2], Alexander Kirillovich[1],
Konstantin Nikolaev[1], and Kamilla Galiaskarova[1]

[1] Kazan (Volga region) Federal University, Kazan, Russia
[2] Tatarstan Academy of Sciences, Kazan, Russia

onevzoro@gmail.com, alik.kirillovich@gmail.com,
konnikolaeff@yandex.ru, galias-alsu@yandex.ru

Abstract. We discuss a semantically annotated corpus of educational mathematical texts in Russian. The objective of our research is to create test collections for automatic formalization of educational mathematical documents. The corpus includes mathematical assertions extracted from educational math textbooks. We manually annotated each assertion as the formula representation in LaTeX and created the formalization of the formula in OpenMath. Symbols used in OpenMath representations are defined in OntoMath^{Edu}, a new educational mathematical ontology.

Keywords: Mathematics, Corpus, Ontology, OpenMath, OntoMath^{Edu}.

1 Introduction

Most of mathematical knowledge is currently recorded in the form of informal documents, consisting of natural language text mixed with formulas in presentation markup. The meaning of such documents is accessible to human readers, but not to machines. In order to this meaning can be machine-actionable, the documents have to be formalized and represented in a form that computers can act on. In practice, full formalization is not necessary, and in fact representation of same semantics only can be enough. This “flexiformalization” paves the way to intelligent mathematical knowledge management applications such as semantic search services, recommender systems, etc. [1, 2]

We study the math assertions in math textbooks for secondury schools. Many of such assertions have the form of plain natural language text but not math statements on formal math language. Our objective is to create a translator of math assertions represented in the form of natural language text to formula representations. These representations we are planning to use in content markup. This development, in turn, requires training and test collections.

In this paper we consider a experimental semantically annotated math corpus, that consists of math assertions extracted from educational math documents. Each assertion is manually annotated as the formula representation in LaTeX and later we create the formalization of this formula in OpenMath [3]. Symbols used in OpenMath representations are defined in OntoMath^{Edu} (https://github.com/CLLKazan/OntoMathEdu), a new educational mathematical ontology [4]. We believe that this ontology will serve as a Linked Open Data hub for mathematical education. Concepts of the ontology contain labels in English, Russian and Tatar and will be interlinked with the external lexical resources from the Linguistic Linked Open Data (LLOD) cloud [5], first of all, WordNet [6], BabelNet [7], RuThes Cloud [8] and Russian-Tatar Thesaurus [9].

The rest of the paper is organized as follows. In Section 2, we briefly review some projects of building formal and informal mathematical corpora. In Sections 3 and 4 we describe the corpus and the process of its construction. In conclusion, we outline the directions of future work.

1 Related Works

In this section we briefly describe informal, formal and parallel informal/formal mathematical corpora.

Informal corpora. arXiv (https://arxiv.org/) is the largest informal mathematical corpus in the world. Its content is represented in LaTeX format. arXMLiv (https://kwarc.info/projects/arXMLiv/) [10] contains arXiv collection, automatically converted to XML, HTML 5 and Content MathML, and making it is more suitable for machine processing.

Formal corpora. The Mizar Mathematical Library (http://mizar.uwb.edu.pl/ library/) is the largest corpus of fully formalized mathematics.

Parallel informal/formal corpora. One of the largest manually-created parallel informal/formal corpora is based on the Flyspeck Project. Flyspeck [11] (https://github.com/flyspeck/flyspeck) is a project, which gives a formal proof of the Kepler conjecture in the HOL Light proof assistant. This project is based on the informal book [12] in LaTeX. Approximately 500 formal statements have been aligned with their informal counterparts. The corpus is available by a user-friendly wiki interface [13].

In [14] Kaliszyk et al. lunched a project aimed at automatic translation of informal mathematical texts into formal ones on base of machine learning methods, trained on aligned informal/formal mathematical corpora. In the subsequent works they pesented several synthetic informal/formal corpora as well as translators trained on them. For example, in [15] they presented a neural network translator from informalized LaTeX-written Mizar texts into the formal Mizar language. The training corpus has been generated by transformation of Mizar to natural language LaTeX text on the basis of the existed method developed for presenting the Mizar articles in the journal *Formalized Mathematics*. In [16, 17] they presented a system for parsing ambiguous formulas from the Flyspeck project. The training informal/formal corpus has been constructed by ambiguation of formal statements from the HOL Light theorems in Flyspeck.

The Formal Abstracts (https://formalabstracts.github.io/) is ongoing project, aiming at formalization of the main results of informal mathematical documents (for example, formalization of the main theorem of a research paper). This formalization is also intended to be used in machine learning tasks.

For our knowledge, there is not neither parallel informal/formal mathematical corpus for Russian nor parallel educational mathematical corpus, so the development of such corpus is needed.

2 Corpus Description and Construction

The corpus is organized as a collection of records. Each record includes the following three fields:

- Russian sentence, extracted from educational textbooks.
- Formula representation of this statement in LaTeX format.
- Formalization of this formula in OpenMath format, where OntoMath^{Edu} ontology is used as a OpenMath content dictionary.

When building the corpus, the following tasks are successively solved.

2.1. Natural Language Statements Extraction

At the first step, we manually extract Russian sentences from education textbooks. We use the secondary school geometry books for 7th–9th grades. The extracted statements are classified according to the following simple classification scheme:

- Class 1: Statements of equality

  a. with complex statement in the left part and simple right part (e.g. positive integer). Example: “Сумма градусных мер двух острых углов прямоугольного треугольника равна 90°” (“The sum of the degree measures of two acute angles of a right triangle is 90°”)
  b. with comparison between equivalent components. Example: “Площадь прямоугольника равна произведению его смежных сторон” (“The area of the rectangle is equal to the product of its adjacent sides”)

- Class 2: Statements of inequality. Example: “Каждая сторона треугольника меньше суммы двух других сторон” (Each side of the triangle is less than the sum of the other two sides)
- Class 3: Definitions of mutual arrangement (e.g. perpendicularity). Example: “Диагонали квадрата взаимно перпендикулярны” (“The diagonals of the square are mutually perpendicular”)
- Class 4: Composite statements (several formulas in one statement provided with “AND” preposition). Example: “Средняя линия трапеции параллельна основаниям и равна их полусумме” (“The middle line of the trapezoid is parallel to the bases and equal to their half-sum”)

- Class 5: Conditional statements. Example: “Если угол одного треугольника равен углу другого треугольника, то площади этих треугольников относятся как произведения сторон, заключающих равные углы” (“If the angle of one triangle is equal to the angle of another triangle, the area of these triangles are the product of the parties, enclosing equal angles”)

2.2. Statements Explication

In the extracted statements, many concepts are mentioned only implicitly due to metonymy, ellipsis, etc. For example, for the statement “The sum of the angles of a convex n-gon is (n-2)*180°” it is assumed that the units of measurement for angles are used in this sum, rather than the angles themselves. Therefore, in the second stage we explain implicit concepts in the extracted statements. Table 1 contains examples of original statements and their explanations.

Table 1. Examples of statements explication

Original (Russian)	Explicated (Russian)	Original (English)	Explicated (English)
Сумма углов выпуклого n-угольника равна (n-2)*180°	Сумма градусных мер углов выпуклого n-угольника равна (n-2)*180°	The sum of the angles of a convex n-gon is (n-2)*180°	The sum of the degree measures of the angles of the convex n-gon is (n-2)*180°
Средняя линия трапеции параллельна основаниям и равна их полусумме	Средняя линия трапеции параллельна её основаниям и её длина равна полусумме длин оснований	The middle line of the trapezoid is parallel to the bases and equal to their half-sum	The middle line of the trapezoid is parallel to its bases and its length is equal to half the sum of the base lengths

2.3. Concepts Annotation

At the third step, we annotate math concepts in the extracted statements. The concepts are annotated in terms of OntoMath^{Edu} ontology. For example, the statement “The middle line of the trapezoid is parallel to the bases and equal to their half-sum” contains the following classes of OntoMath^{Edu} ontology: *Middle line*, *trapezoid*, *base*, etc.

2.4. Representation of Statements as Formulas

At the next stage, we represent the statements as the formulas in LaTeX. Table 2 contains examples of this representation as formula statements.

2.5. Formalization of the Formulas in OpenMath

At the final step, we formalize formulas in OpenMath format. We use OntoMath^{Edu} ontology as a content dictionary in this formalization.

Table 2. Examples of statements and its formula representation

Statement (Russian)	Formula representation (Russian)	Statement (English)	Formula representation (English)
Сумма углов выпуклого n-угольника равна (n-2)*180°	∠A_1+∠A_2+…+∠A_n =(n-2)*180°, где A_1…A_n – выпуклый n-угольник; ∠A_1, ∠A_2, …, ∠A_n – углы выпуклого n-угольника	The sum of the angles of a convex n-gon is (n-2)*180°	∠A_1+∠A_2+...+∠A_n=(n -2)*180°, where A_1...AP – convex n-gon; ∠A_1, ∠A_2, ..., ∠A_n – angles of a convex n-gon
Сумма двух острых углов прямоугольного треугольника равна 90°	∠ABC+ ∠BAC = 90°, где ABC – прямоугольный треугольник; ∠BCA – прямой угол	The sum of the two acute angles of a right triangle is 90°	∠ABC+ ∠BAC = 90°, where ABC is a right triangle; ∠BCA – straight angle

3 Conclusion

In this paper we presented a semantically annotated corpus of educational math texts in Russian. The corpus consists of natural language statements, extracted from an educational textbook. Extracted statements were manually complemented by its representation as LaTeX formulas and OpenMath formal representation. As a OpenMath content dictionary we used OntoMath^{Edu} ontology.

The corpus now is still on the development stage, so our immediate goal is to release the first working version.

After that we are going to adopt it in the development of the components of a new digital educational platform, which is intended for solving such tasks as automatic knowledge testing; automatic recommendation of educational materials according to an individual study plan; and semantic annotation of educational materials. In particular, the corpus is intended to be used for training an automatic translator from Russian educational documents to its formal representation, as well as a test collection for an ontology-based mathematical information extraction tool. Also, the corpus can be used to verbalize a formal mathematical document as a natural language text in Russian. Additionally, we are going to use it for enrichment of OntoMath^{Edu} ontology.

The corpus will be published at the Linked Open Data (LOD) cloud.

Acknowledgements

This work was funded by the subsidy allocated to Kazan Federal University for the state assignment in the sphere of scientific activities, grant agreement no. 1.2368.2017, and Russian Foundation for Basic Research and the government of the region of the Russian Federation, grant No. 18-47-160007.

References

1. Kohlhase, M.: The Flexiformalist Manifesto. In: Voronkov, A., et al (eds.) Proceedings of the 14th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC 2012), 30–35. IEEE (2012). doi:10.1109/SYNASC.2012.78
2. Kohlhase, A. and Kohlhase, M.: Towards a flexible notion of document context. In: Protopsaltis, A., et al (eds.) Proceedings of the 29th ACM international conference on Design of communication (SIGDOC 2011), 181–188. ACM (2011). doi:10.1145/2038476.2038512
3. Buswell, S., Caprotti, O., Carlisle, D.P., Dewar, M.C., Gaëtano, M., and Kohlhase, M.: The OpenMath Standard, Version 2.0. The OpenMath Society (2004). https://www.openmath.org/standard/om20-2004-06-30/omstd20.html
4. Kirillovich, A., Nevzorova, O., Falileeva, M., Lipachev, E., and Shakirova, L.: OntoMath^{Edu}: towards an educational mathematical ontology. In: Workshop Papers at 12th Conference on Intelligent Computer Mathematics (CICM-WS 2019). CEUR Workshop Proceedings (forthcoming)
5. McCrae, J.P., Chiarcos, C., Bond, F., Cimiano, P., Declerck, T., de Melo, G., Gracia, J., Hellmann, S., Klimek, B., Moran, S., Osenova, P., Pareja-Lora, A., and Pool, J.: The open linguistics working group: developing the linguistic linked open data cloud. In: Calzolari N., et al. (eds.) Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016), 2435–2441. ELRA (2016).
6. McCrae, J.P., Fellbaum, C., and Cimiano, P.: Publishing and Linking WordNet using lemon and RDF. In: Chiarcos C. et al. (eds.) Proceedings of the 3rd Workshop on Linked Data in Linguistics (LDL-2014), 13–16. ELRA (2014).
7. Ehrmann, M., Cecconi, F., Vannella, D., McCrae, J., Cimiano, P., and Navigli, R.: Representing multilingual data as linked data: the case of BabelNet 2.0. In: Calzolari N., et al. (eds.) Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014), 401–408. ELRA (2014).
8. Kirillovich, A., Nevzorova, O., Gimadiev. E., and Loukachevitch, N.: RuThes cloud: towards a multilevel linguistic linked open data resource for russian. In: Różewski, P. and Lange, C. (eds.) Proceedings of the 8th International Conference on Knowledge Engineering and Semantic Web (KESW 2017). Communications in Computer and Information Science, vol. 786, pp. 38–52. Springer, Cham (2017). doi:10.1007/978-3-319-69548-8_4
9. Galieva, A., Kirillovich, A., Khakimov, B., Loukachevitch, N., Nevzorova, O., and Suleymanov, D.: Toward domain-specific russian-tatar thesaurus construction. In: Proceedings of the International Conference IMS-2017, pp. 120–124. ACM (2017). doi:10.1145/3143699.3143716
10. Stamerjohanns, H., Kohlhase, M., Ginev, D., David, C., and Miller, B.: Transforming large collections of scientific publications to XML. Mathematics in Computer Science, 3 (3), 299–307 (2010). doi:10.1007/s11786-010-0024-7
11. Hales, T. C.: Introduction to the Flyspeck project. In: Coquand, T., Lombardi, H. and Roy, M.-F. (eds.) Mathematics, Algorithms, Proofs. Dagstuhl Seminar Proceedings, vol. 05021. IBFI (2006).
12. Hales, T.: Dense Sphere Packings: A Blueprint for Formal Proofs. Cambridge University Press (2012).
13. Tankink, C., Kaliszyk, C., Urban, J., and Geuvers, H.: Formal mathematics on display: a wiki for Flyspeck. In: Carette, J., et al (eds.) Proceedings of Intelligent Computer Mathematics: MKM, Calculemus, DML, and Systems and Projects 2013 (CICM 2013). Lecture

Notes in Computer Science, vol. 7961, pp. 152–167. Springer (2013). doi:10.1007/978-3-642-39320-4_10

14. Kaliszyk, C., Urban, J., Vyskočil, J., and Geuvers, H.: Developing corpus-based translation methods between informal and formal mathematics: project description. In: Watt S. M., et al (eds.) Proceedings of the International Conference on Intelligent Computer Mathematics (CICM 2014). Lecture Notes in Computer Science, vol. 8543, pp. 435–439. Springer, Cham (2014). doi:10.1007/978-3-319-08434-3_34
15. Wang, Q., Kaliszyk, C., and Urban, J.: First experiments with neural translation of informal to formal mathematics. In: Rabe, F., et al (eds.) Proceedings of the 11th International Conference on Intelligent Computer Mathematics (CICM 2018). Lecture Notes in Computer Science, vol. 11006, pp. 255–270. Springer, Cham (2018). doi:10.1007/978-3-319-96812-4_22
16. Kaliszyk, C., Urban, J., and Vyskočil, J.: Learning to parse on aligned corpora (Rough Diamond). In: Urban, C. and Zhang, X. (eds.) Proceedings of the 6th International Conference on Interactive Theorem Proving (ITP 2015). Lecture Notes in Computer Science, vol. 9236, pp. 227–233. Springer, Cham (2015). doi:10.1007/978-3-319-22102-1_15
17. Kaliszyk, C., Urban, J., and Vyskočil, J.: Automating formalization by statistical and semantic parsing of mathematics. In: Ayala-Rincón, M. and Muñoz, C. (eds.) Proceedings of the 8th International Conference on Interactive Theorem Proving (ITP 2017). Lecture Notes in Computer Science, vol. 10499, pp. 12–27. Springer, Cham (2017). doi:10.1007/978-3-319-66107-0_2
18. Atanasyan, L., Butuzov, V., and Kadomcev S.: Geometry, 7–9 grades: textbook for general-education schools. Prosveshenie, Moscow (2010).

INFORMATION EXTRACTION FROM TEXT – II

The Software Environment for Multi-aspect Study of Lexical Characteristics of Text

Elena Sidorova and Irina Akhmadeeva

A.P. Ershov Institute of Informatics Systems, Siberian Branch of the Russian Academy of Sciences, Acad. Lavrentjev avenue 6, 630090 Novosibirsk, Russia
{lsidorova,i.r.akhmadeeva}@iis.nsk.su

Abstract. The software environment for multi-aspect study of the lexical characteristics of the text is considered. The proposed environment provides tools and features allowing automatically building a dictionary based on a text corpus of interest. The created toolkit focused on lexical units acting as markers and indicators of higher level objects. The considered environment allows solving various text analysis tasks; because it integrates various tools for conducting language research and supports customization of vocabularies to a problem area. This toolkit includes interfaces for developing vocabularies and a system of features. To study the contexts of the use of terms, concordance construction tools are provided. Concordances allow the researcher to test his or her hypothesis about the functionality of a particular lexical unit. To describe more complex constructions to be extracted, a user can apply search patterns, supported by a user-friendly language. Using these patterns allows us to develop lexicographic resources containing not only the traditional vocabularies and stable inseparable lexical phrases, but also language constructs that have a more complex structure.

Keywords: domain vocabulary, terminology, concordance, search pattern

1 Introduction

A plain text, as it is a source of information, and one of the most important means for communication needs to be thoroughly studied. It is necessary for both evaluating “quality” of what has been written and automatic text processing along with supporting information retrieval services. Studying language phenomena and modeling text understanding processes taking place at the different language levels are in the focus of contemporary research in computational linguistics.

In order to work out these problems, it is usual to apply a variety of knowledge written in a formalized form. Among them are widely known thesauri such as WordNet and RusNet, explanatory combinatorial dictionaries, annotated corpora of texts (for example, The Russian National Corpus www.ruscorpora.ru), and other resources. Serving as an instrument for describing a subject vocabulary, thesaurus allows us to characterize terms and their connections from the point of view of peculiarities of use in this subject domain [1]. Another way of studying the linguistic phenomena is to use corpora of texts. A text corpus is the source and tool of multi-aspect lexicographic works [2]. The use of specialized methods, such as a frequency analysis of a vocabulary in the corpus, construction of concordances on various grounds, can help in automating the work of experts on a formal structures research, initial filling of dictionaries, and construction of linguistic models on the basis of an annotated corpus of texts. Despite the widely demanded functionality, there are no known analogues of the specialized set of customizable components that integrate lexicographic research methods for Russian and provide semantic markup of terms, statistical analysis, and construction of concordances. As for other languages, similar functionality is presented on such platforms as GATE (https://gate.ac.uk) or CLARIN portal (www.clarin.eu). Components developed by various groups of researchers from different countries and for different languages are presented in these resources. As well as a method of integrating components into a chain of calculations is proposed.

Literature overview [3–5] shows that many researchers having a task to extract terminology from a large text collection usually choose to combine linguistic and statistical methods. For extraction of lists of candidate terms that satisfy the specified linguistic conditions, the method of search patterns describing classes of language expressions is used. Depending on the type of language information taken into account, the patterns used in various works are divided into grammatical, lexico-grammatical [3, 5] and lexico-syntactic patterns [7, 8]. Extraction of candidate terms is accompanied by calculation of statistics and weights for filtering and sorting a result list. The list of candidate terms includes not only special concepts established in this field, but also numerous general scientific, peripheral and author’s terms that, as shown in [12], are characterized by a high degree of variation of the language form. In this situation, an expert assessment stage is needed, at which the ranked lists are presented to the expert for selecting true terms.

This paper concerns describing various supporting tools for studying lexical characteristics of a text based on corpora. Combining proposed tools allowed us to develop an environment for creating problem-oriented vocabularies and provide the end user with various possibilities to study language phenomena.

2 Requirements for the Text Research Support Environment

The development of linguistic models and the creation of resources of sufficient quality require scrupulous manual labor, supported by software tools. The software environment should provide the expert with various tools to create necessary knowledge bases and carry out case studies.

We formulate the requirements for the system for multi-aspect study of lexical characteristics of text as follows:

1. The system should be able to automatically fill vocabularies based on text corpora;
2. The system user should be able to customize and add various attributes for vocabulary terms;

3. The system should be able to carry out lexical analysis (segmenting a text, and extracting terms that are presented in the vocabulary);
4. The system should keep statistical and combinatorial properties of language phenomena found in texts;
5. The system should be able to build a concordance of terms and provide the user with corresponding visualization tools.

We developed a system including basic research tools that follows (Fig. 1): an interface for developing a dictionary and creating a group of features, tools for automatic generation of lexical content of a dictionary by the corpus of texts and calculating quantitative characteristics of found terms, concordance construction tools for studying contexts of lexical units.

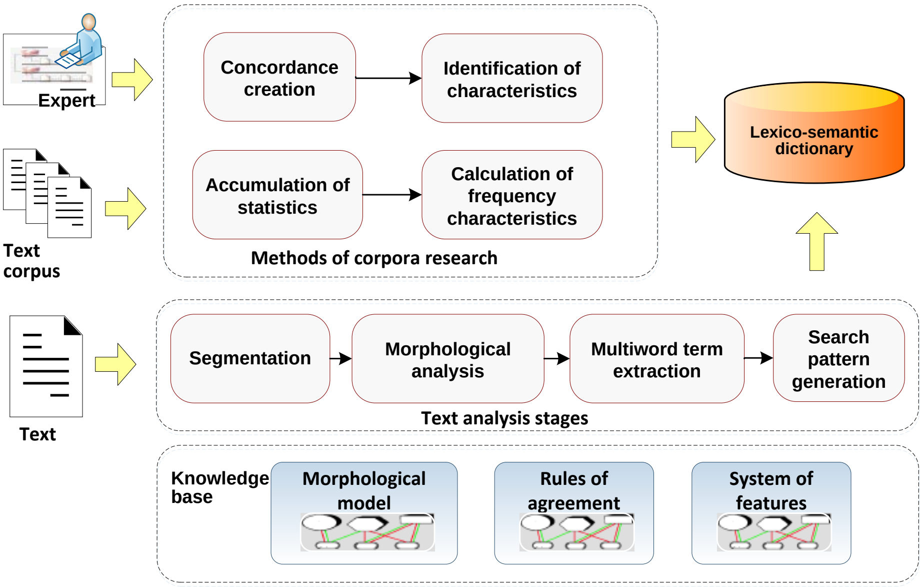

Expert
Text corpus
Concordance creation
Identification of characteristics
Accumulation of statistics
Calculation of frequency characteristics
Methods of corpora research
Lexico-semantic dictionary
Text
Segmentation
Morphological analysis
Multiword term extraction
Search pattern generation
Text analysis stages
Knowledge base
Morphological model
Rules of agreement
System of features

Fig. 1. The environment for study of lexical characteristics of text

3 Knowledge Representation Model

The considered lexicographic knowledge model includes three main components. The dictionary defines the lexical model of the sublanguage under consideration, which defined by the problem area. Grammar provides search and retrieval of lexical units from texts. The set of user-defined pragmatically-oriented features supports recording of observations, and is focused on further support of automated text processing methods.

A representative problem-oriented corpus of texts lies in the research basis. The main tools providing research support are following:

- search for examples of using vocabulary terms;
- build a variety of contexts (concordances);
- calculate frequencies, co-occurrences, distributions, etc.

3.1 Lexical Model

In our approach, the dictionary entry contains all information that necessary either for extracting terms from text or for supporting the subsequent stages of the text analysis.

A problem-oriented dictionary is a volume of vocabulary organized according to a semantic (thematic/genre/etc.) principle, considering a certain set of basic formal relationships. Formally, the dictionary is defined as a system of the form:

$$V=\{W, P, M, G, S, F_w, F_p\},$$

where W is a set of lexemes, where each lexeme is mapped to the entire set of its lexical forms; P is a set of multiword terms defined as a pair of a form <N-gram, structure type>. The N-gram specifies sequence of lexemes, and the structure type defines the head of the phrase and rules for matching N-gram elements.

M is a morphological model of language. It defines morphological classes and features.

G is a set of agreement rules which are used to extract multiword terms.

S is a problem-oriented set of features, terms could be marked with.

$F_w=W\rightarrow 2^{M\times S}$, $F_p=P\rightarrow 2^{G\times S}$ is a function that maps terms to sets of features.

The morphological representation the system provides is designed in such a way that it could be customized depending on the specific problem the user is working on. He or she can define his or her own set of features and classes, and ensure they are integrated in the basic morphological representation. A morphological class is defined by a part of speech, a set of lexical features (for example, animacy or gender for nouns) and a type of paradigm. It is a rather rare case when one would need to change class. For example, it is necessary when using additional specialized dictionaries of terms (dictionaries of names, geographical locations) or there is a need to include words of another language in the dictionary.

The description of morphological information includes the following concepts: morphological attribute, class, part of speech, and type of paradigm.

The morphological attribute is described by the name N_i and the set of its values X_i: $<N_i, X_i>$ (for example, <*Gender, {masculine, feminine, neuter}*>). Part of speech is also an attribute, but since it must always be present, it was decided to create a separate entity for this purpose. Attributes within each class are divided into derivational, inherent to all forms of the lexeme of this class, and inflectional, distinguishing forms of one lexeme.

The paradigm type determines its length and matches each element of the paradigm with a set of attribute values (for example, for a “simple” adjective it is a triple <case, number, gender>). Such elements are strictly ordered, which makes it possible to use a compact form of writing in a tree-like structure, the vertices of which are subsets of the attribute values $<A_i, X_i>$. A pair of functions f: $n\rightarrow X_{i1} * ... * X_{ik}$, g: $X_{i1} * ... * X_{ik}\rightarrow n$ provides a conversion of the inflectional paradigm to a set of attribute values, and vice versa. So each lexeme is assigned a paradigm from the paradigm table, and each paradigm is assigned a type of paradigm describing its structure.

The morphological class includes a part of speech, a set of derivational lexical features $x_{ij} \in X_i$ (for example, animation or gender in nouns) and a type of paradigm describing attributes of word forms.

Another important feature of the system is the support of multiword terms (phrases) formed according to the shallow syntactic analysis based upon a fixed set of rules. Most of the multiword terms include from two to four words and are formed using the rules of the following type:

- A+N (*“аналоговый датчик”* which means *“analog sensor”* in Russian) – agreement of a noun and an adjective;
- N+Ngent (*“автор учебника”* which means *“textbook author”* in Russian) – agreement of a noun and a noun in the genitive case;
- A+A+N (*“новая информационная технология” – “new information technology”*);
- N+Agent+Ngent (*“обработка естественного языка” – “natural language processing”*);
- A+N+Ngent (“*локальная степень вершины*” – “*local degree of a vertex*”),
- N+Ngent+Ngent (“*компонента связности графа*” – “*connected component of a graph*”) etc.

There are also terms with a more complex structure, for example, with dependent prepositional groups:

- N+PREP+N (“резервуар с жидкостью” – “reservoir with liquid”, “рассуждение по умолчанию” – “*default reasoning*”);
- N+PREP+N+N (“*поиск в пространстве состояний*” – “*search in the state space*”);
- N+PREP+A+N (*“автомат с переменной структурой” – “variable-structure automata”*) etc.

The system has its own component of multiword term extraction of the Russian language, which, according to a given set of words and their grammatical characteristics, checks agreement in accordance with one of the syntactic models and synthesizes a normal form of a multiword term. The multiword vocabulary term is uniquely identified by a triple <normal form, rule, < lexical structure>>. Such term has a syntactic head (a single-word term) and grammatical features that are formed on the basis of the grammatical features of the head.

3.2 Features of Terms

Depending on the problem being worked out, terms in the dictionary can be supplied with features of various types: statistical (for solving classification problems), genre (for text genre analysis), semantic (for semantic analysis), formal (for identifying markers of certain structures), etc.

Statistical features keep frequency information. When text is processed all terms occurred in it have their statistics updated. To perform text classification, we need a training corpus, i.e. corpus annotated with predefined set of interrelated topics. In the dictionary for each term we know how much times it occurred in the training corpus (this is called absolute frequency), and a number of texts in which the term occurred (text frequency). We also know a list of topics where term was found, absolute frequencies and text frequencies for each topic from the list. Some parameters (relative frequency, tf*idf, weight) are calculated dynamically.

The set of features user needs to markup dictionary terms with, are defined by him or her and depends on the task being addressed, so it is completely user-defined and problem oriented. To encode various information about the term (semantic, genre, stylistic, etc.), the following facilities are provided.

- Class. The term could be of one of the classes. A class hierarchy allows user to assign a term to a certain level of hierarchy: more general or specific, inheriting properties from upper classes.
- Attribute. Attributes are used to represent the lexical meaning of a term. Combining word's semantic attribute values, we can, to a certain extent, model the component semantic structure of a word. The main components of the semantic structure of the term can be considered as thesaurus descriptors.
- Alternative feature sets allow the term ambiguity to be expressed.

4 Working with Text Corpora

The developed environment consists of vocabulary components and processors that, on the one hand, allow automatic creation, fill and edit dictionaries, and, on the other hand, use those dictionaries in lexical text analysis. One of the most important features is user supporting tools such as term sorting, term filtering, text coverage visualization, concordance constructor, etc.

4.1 Corpus-based Vocabulary Learning

The terminology extraction process consists of steps that follows: a) text tokenization b) lexical and morphological analysis (lemmatization, extraction of lexical and grammatical features, normalization), c) extraction of phrases that are "look like" terms (phrase term-likeness is based on predefined grammatical models), d) update the statistics of found terms.

Following are the modules that are used for dictionary construction.

The morphological analysis is carried out on the basis of the Dialing module (www.aot.ru), which contains a dictionary of general Russian terms. This module supports search for words, along with their grammatical features and normal forms on the dictionary. It also provides an additional feature called predictor, which for any word that is not in the dictionary can make assumptions about part of speech, normal form and other features. Predictor can make up to three assumptions for a single term.

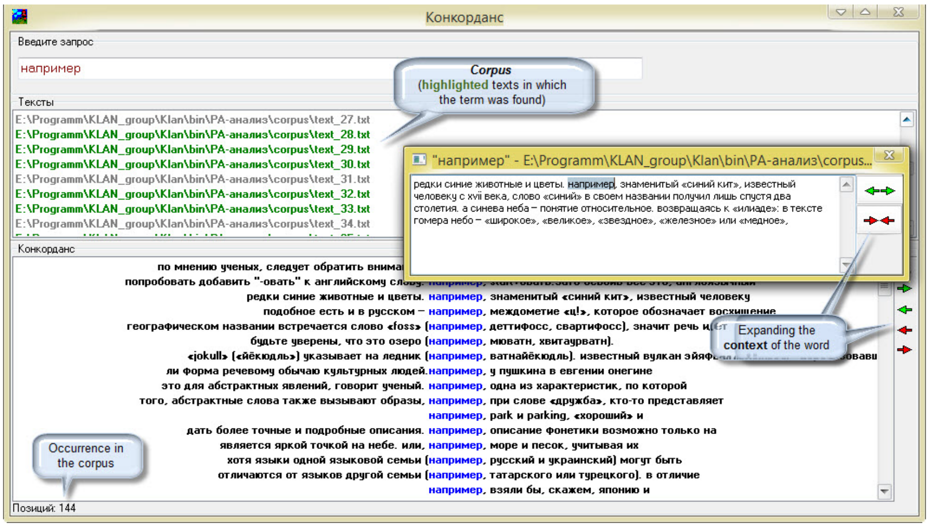
The multiword term extractor is applied to recognize phrases in accordance with a fixed set of grammatical rules. The main objective of the module is to identify the most important term-forming syntactic groups, most of which are nominal groups or are based on them.

Using aforementioned modules to process text corpus, we will end up with the resulting dictionary and statistics of frequencies of terms. If there were special features marked in the corpus, the corresponding terms are treated as having those features, and statistics are also kept with regard to the features.

The proposed environment hereby provides tools and features allowing automatic building a draft dictionary from scratch based on a text corpus of interest. On the basis of such a dictionary a further research could be carried out.

4.2 Concordance

A concordance is the traditional way of studying a corpus of texts. It contains a complete index of terms that share context with the selected one. The sizes of contexts may vary. Concordances allow the researcher to test his or her hypothesis about the functionality of a particular lexical unit. It could be said that a concordance connects dictionary terms with the text corpus, and serves as a linguistic markup at the morphological and shallow syntactic level.



Конкорданс
Введите запрос
Тексты
Конкорданс
Corpus
(highlighted texts in which the term was found)
Expanding the context of the word
Occurrence in the corpus

Введите запрос: например

Тексты:
E:\Programm\KLAN_group\Klan\bin\PA-анализ\corpus\text_27.txt
E:\Programm\KLAN_group\Klan\bin\PA-анализ\corpus\text_28.txt
E:\Programm\KLAN_group\Klan\bin\PA-анализ\corpus\text_29.txt
E:\Programm\KLAN_group\Klan\bin\PA-анализ\corpus\text_30.txt
E:\Programm\KLAN_group\Klan\bin\PA-анализ\corpus\text_31.txt
E:\Programm\KLAN_group\Klan\bin\PA-анализ\corpus\text_32.txt
E:\Programm\KLAN_group\Klan\bin\PA-анализ\corpus\text_33.txt
E:\Programm\KLAN_group\Klan\bin\PA-анализ\corpus\text_34.txt

"например" - E:\Programm\KLAN_group\Klan\bin\PA-анализ\corpus...

редки синие животные и цветы. например, знаменитый «синий кит», известный человеку с xvii века, слово «синий» в своем названии получил лишь спустя два столетия. а синева неба – понятие относительное. возвращаясь к «илиаде»: в тексте гомера небо – «широкое», «великое», «звездное», «железное» или «медное»,

Конкорданс:

```
                    по мнению ученых, следует обратить внима
 попробовать добавить "-овать" к английскому сл
                 редки синие животные и цветы. например, знаменитый «синий кит», известный человеку
                     подобное есть и в русском – например, междометие «ц!», которое обозначает восхищение
 географическом названии встречается слово «foss» (например, деттифосс, свартифосс), значит речь и
                  будьте уверены, что это озеро (например, мюватн, хвитаурватн).
          «jokull» («йёкюдль») указывает на ледник (например, ватнайёкюдль). известный вулкан эйяф
       ли форма речевому обычаю культурных людей.например, у пушкина в евгении онегине
      это для абстрактных явлений, говорит ученый. например, одна из характеристик, по которой
  того, абстрактные слова также вызывают образы, например, при слове «дружба», кто-то представляет
                                                 например, park и parking, «хороший» и
           дать более точные и подробные описания. например, описание фонетики возможно только на
                  является яркой точкой на небе. или, например, море и песок, учитывая их
                 хотя языки одной языковой семьи (например, русский и украинский) могут быть
              отличаются от языков другой семьи (например, татарского или турецкого). в отличие
                                                 например, взяли бы, скажем, японию и
```

Позиций: 144

Fig. 2. The concordance construction tool

The implemented in the environment concordance construction tool works with text files. The user can customize the size of a text fragment being viewed in a context of a term entry (Fig. 2.). An example of a concordance given in Fig. 2 for a word "*например*" (which means "*for example*" in Russian) includes 144 occurrences from the text corpus, and shows how context could be expanded word-wise, or how one or more paragraphs could be summoned to view by providing the selected term entry. In the example the research purpose was to test the hypothesis about the use of this term in the *argument from expert opinion*.

In general, this kind of research allows user to identify more complex language constructs that ensure the precision and recall of the information extraction process, and to identify additional features based on them. To describe constructions to be extracted, user can use search patterns which based on regular expressions, supported by a user-friendly language [7, 8].

4.3 Search Patterns

In our studies, we have been using different types of patterns and tools that support automatic text processing. In each case the toolkit was chosen based on a problem area and methods used to solve the target problem.

For example, in the project targeting the problem of filtering out prohibited content [9], in addition to being marked by thematic, genre and lexico-semantic features got from the vocabulary texts was processed with special patterns each of which described constructions specific to a particular Internet genre [10]. Those patterns have significantly improved the accuracy of the genre classification.

Taking a closer look, a pattern allowing detecting a block containing personal information on a website can be represented as follows:

_profile:[“личный кабинет”][“профиль”][“аккаунт”][“о себе"][“личный профиль”]

//_profile: [“personal account”] [“profile”] [“account”] [“about me”] [“personal profile”]

Profile Description / Contacts: [<_ profile, all_h>]

In this case, the *_profile* pattern is defined by a set of alternative terms. If any of these terms appears as a part of a header at any level (as indicated in the second pattern) we can classify a text block as a block containing user profile information. Patterns defined like one from the example belongs to logical combinatorial lexical patterns.

In another project our goal was to extract information from technical documentation texts. We built a glossary of terms with semantic subject-oriented markup, and applied search patterns to extract parametric information, which is often represented by numerical and symbolic notations and abbreviations. The patterns used are defined as follows.

class: ‘Object ACS’, template: ‘ACS TP’, type: ‘base’

[АСУ] = АСУ{ТП}; автоматизированн{…} систем{…} управления

[ACS] = ACS{TP}; automatic control {...} system

Patterns of this type are called lexical-semantic patterns [11].

Finally, there is yet one project where we target a philosophical problem of argument analysis. We build a dictionary of markers of argumentative structures on the basis of an annotated text corpus. Applying patterns allows us to represent area specific constructions, which could be consisting of more than one part, separated with gaps.

DSC = [begin: DS, w / <speech> <Verb, past | present>, Expert <N, им>, end: ES]

quote_l = [“|«]

quote_r = [”|»]

DS = [begin: quote_l, end: quote_r] // direct speech

Thus, in the experiment on the extraction of arguments from expert opinion, the search accuracy using patterns was 86.5%. Based on the above we can conclude that using our search patterns allows us to develop lexicographic resources containing not only the traditional vocabularies and stable inseparable lexical phrases, but also language constructs that have a more complex structure.

5 Conclusion

This paper is devoted to describe approaches and methods for development of lexicographic resources, conducting studies on text corpora in order to ensure the completeness and reliability of models being developed. The created toolkit is focused on lexical units acting as markers and indicators of higher level objects (semantic, pragmatic, structural-genre, logical-argumentative, etc.).

The considered software environment integrates basic tools required to conduct research on lexical characteristics of the text, which ensures a full cycle of the expert’s work. The environment has wide possibilities for tuning of parameters, ranging from grammatical categories, lexico-semantic characteristics, classification parameters, and ending with specific search patterns that ensure the search for contexts and the construction of concordances. Practical use of this software in various research projects showed usability, the relevance of functionality and adaptability for different tasks.

Consequently, distinctive features of the system are:

- possibility of multipurpose use in solving various text analysis tasks, such as text classification, information extraction, lexicographic research of a text corpus, genre analysis, etc.;
- integration of various tools within the same environment for conducting language researches and providing customization of vocabularies to a problem area: concordance, statistical study based on a corpus of texts, support for semantic markup of lexical units, a rich set of search tools and filtering.

The environment supports a rich lexical model that integrates various models of representation of lexical units and language constructs.

1. Expandable and customizable morphological model (in contrast to the well-known morphological analyzers aot, pymorphy, mystem, etc.);
2. Grammar models for Russian phrases extraction and the possibility of selectively use them;
3. Search patterns integrate semantic, grammatical, lexical and symbolic representations based on logical operations.

Further improvements of the system may lie in developing of corpus-based research tools, such as constructing concordances for the joint occurrence of terms, using conditions for the presence / absence of feature sets in search queries, etc. It is also planned to enhance the reusability of results of the research by storing the data in standard formats based on XML (TEI, OWL).

Acknowledgment

The work was carried out with the financial support of the Russian Foundation for Basic Research (grants 19-07-00762 and 18-00-01376 (18-00-00889)).

References

1. Loukachevitch, N.V.: Thesauri in information retrieval tasks. MSU Publ., Moscow (2011).
2. Sinclair, J. Corpus, Concordance, Collocation. Oxford University Press, Oxford (1991).
3. Zakharov, V.P. and Khokhlova M.V.: Automatic extracting of terminological phrases. Structural and Applied linguistics **10**, 182–200 (2014)
4. Bolshakova, E., Loukachevitch, N., and Nokel, M.: Topic models can improve domain term extraction. In: International conference on Information Retrieval ECIR-2013, pp. 684–687. Springer Verlag, (2013)
5. Mitrofanova, O.A. and Zaharov V.P.: Automatic extracting terminological phrases. In: Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialog-2009”, pp. 321–328. Moscow (2009).
6. Sokirko, A.V.: Morphological modules on the site www.aot.ru. In: Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialog-2004”, pp. 559–564. Nauka Publ, Moscow (2004).
7. Bol'shakova, E.I., Baeva, N.V., Bordachenkova, E.A., Vasil'eva, N.E., and Morozov S.S.: Lexicosyntactic patterns for automatic text processing. In: Proc. Int. Conf. Dialogue 2007, pp. 70–75. Moscow (2007).
8. Rabchevsky, E.A., Bulatova, G.I., and Sharafutdinov, I.M.: Application of lexical-syntactic patterns to the automation of ontology building process. In: Proc. 10th All-Rus. Conf. RCDL’2008 Electronic Libraries: Perspective Methods, Technologies, Electronic Collections, pp. 103–106. Dubna (2008).
9. Sidorova, E.A., Kononenko, I.S., and Zagorulko, Yu.A.: An approach to filtering prohibited content on the web. In: CEUR Workshop Proceedings, 2022. pp. 64–71. CEURWS.org (2017).
10. Sidorova, E.A. and Kononenko, I.S.: Genre aspects of websites classification. Software Engineering **8**, 32–40 (2015).
11. Sidorova, E.A. and Timofeev, P.S.: A lexico-semantic templates as a tool for declarative description language constructs linguistic text analysis. System Informatics **13**, 35–48 (2018) DOI: 10.31144/si.2307–6410.
12. Bol'shakova, E.I. and Ivanov, K.M.: Term extraction for constructing subject index of educational scientific text. In: Sixteenth Russian Conference on Artificial Intelligence RCAI-2018. T1, pp. 253–261. Moscow (2018).

Towards a Tatar Wordnet: a Methodology of Using Tatar Thesaurus

Alfiya Galieva[1], Alexander Kirillovich[2],
Natalia Loukachevich[3] and Olga Nevzorova[1,2]

[1] Tatarstan Academy of Sciences, Kazan, Russia
[2] Kazan (Volga region) Federal University, Kazan, Russia
[3] Lomonosov Moscow State University, Moscow, Russia

amgalieva@gmail.com, alik.kirillovich@gmail.com,
louk_nat@mail.ru, onevzoro@gmail.com

Abstract. For wordnet developing for a new language, the key problem is to find original resources that contain enough lexical data of the language in an appropriate format. This article discusses the structure, methodology of compilation and the current state of the bilingual Russian-Tatar Social-Political Thesaurus, which can serve as an initial resource for building the Tatar Wordnet. This thesaurus reflects the logical-semantic organization of lexical elements (synonymous, generic, and some other relationships) at the conceptual and lexical levels. Mainly, we focus on building synsets for *nouns* (single *nouns and noun phrases*).

Keywords: Tatar language, WordNet, Thesaurus, Linguistic ontology, Sociopolitical terminology.

1 Introduction

A great hindrance to develop linguistic ontologies for a new language and conceptual modeling is the lack of original lexicographic resources containing full and relevant linguistic data description.

Success of Princeton WordNet has determined emergence of wordnets and wordnet-like projects for different languages and multilingual wordnets. In wordnet building developers often use the Expand Model (Vossen 2002: 52) when available wordnets that serve as mapped linguistic relations between the items and ready synsets of a source language are translated using bilingual dictionaries into equivalent synsets in the target language. Most of the wordnets existing for today are implemented by translating Princeton English WordNet. The alternative approach is very laborious, time-consuming and difficult to implement, based on compiling synsets and mapping semantic relations between word senses directly on the data of the language for which a wordnet is developed. In this case good dictionaries of synonyms and other semantic dictionaries are required.

So the possibility of developing wordnets is largely determined by the presence of bilingual dictionaries or fairly complete descriptions of the semantic system of the language.

The absence of large English-Tatar dictionaries (the available ones are of very limited volume and may be used only for education purposes) makes it impossible to use the Expand Model to Tatar wordnet development, as well as absence of Tatar semantic dictionaries makes it almost impossible to develop original Tatar wordnet.

The objective of this article is to describe the methodology for constructing a Tatar wordnet based on a lexical resource such as the Tatar social-political thesaurus. This approach allows you to directly use the data of the thesaurus, primarily a set of synsets and relationships between synsets.

The body of the paper is organized as follows. Section 2 outlines the basic theoretical background of the study, and the main attention is paid to wordnet projects developed for the Turkic languages. Section 3 presents the methodology of compiling the Russian-Tatar socio-political thesaurus and its current state. Section 4 describes the most important aspects of implementing a wordnet-like resource using Tatar thesaurus synsets for Tatar nouns. Section 5 discusses the conclusions and outlines the prospects of future work.

2 Related Works

At present time, there are various wordnets for some Turkic languages.

Two Turkish wordnet projects have been developed for the Turkish language. The first one (Çetinoğlu, et al, 2018; Bilgin, et al, 2004) has been created at Sabancı University as part of the BalkaNet project (Tufis, et al, 2004). The BalkaNet project was built on the basis of a combination of expand and merge approaches. All wordnets contain many synonyms for Balkan common topics, as well as synsets typical for each of the BalkaNet languages. The size of the Turkish Wordnet is about 15,000 synsets.

Another Turkish wordnet is the KeNet (Ehsani, 2018; Ehsani, et al, 2018). This wordnet was built on the basis of modern Turkish dictionaries. A bottom-up approach was used to build this resource. Based on dictionaries, words were selected and then they were manually grouped into synsets. The relationships between words have been automatically extracted from dictionary definitions and then these relationships have been created between synsets. The size of this resource is about 113,000 synsets.

Unfortunately, the lack of large Turkish-Tatar dictionaries (as well as English-Tatar ones) makes it impossible to translate Turkish resources into the Tatar language. In this respect the Tatar language can be attributed to low-resource languages.

The Extended Open Multilingual Wordnet (Bond, 2013) resource is built from Open Multilingual Wordnet by replenishing the WordNet data automatically extracted from the Wiktionary and Unicode Common Locale Data Repository (CLDR). The resource contains wordnets for 150 languages, including several Turkic: Azerbaijani, Kazakh, Kirghiz, Tatar, Turkmen, Turkish, Uzbek. The Tatar wordnet contains a total of 550 concepts, which cover 5% of the PWN core concepts.

The BabelNet (Navigli and Ponzetto, 2012) resource contains a common network of concepts that have text inputs in many languages. The BabelNet contains 90,821 Tatar text entries that refer to 63,989 concepts. However, due to the fact that this resource was built automatically, it has quality problems.

Thus, the development of a qualitative Tatar wordnet with an emphasis on the specific features of the Tatar language based on the existing lexical resources is very relevant.

3 Tatar Socio-Political Thesaurus: Methodological Issues of Compiling and Its Current State

The conceptual model of the Tatar socio-political thesaurus (hereinafter referred to as TatThes), the general principles of displaying linguistic data are taken from the RuThes project (http://www.labinform.ru/pub/ruthes/) (Loukachevitch and Dobrov, 2014; Loukachevitch, Dobrov and Chetviorkin, 2014). The RuThes thesaurus build as is a hierarchical network of concepts with attributed lexical entries for automatic text processing.

In the RuThes each concept is linked with a set of language expressions (nouns, adjectives, verbs or multiword expressions of different structures – noun phrases and verb phrases) which refer to the concept in texts (lexical entries). The RuThes concepts have no internal structure as attributes (frame elements), so concept properties are described only by means of relations with other concepts.

Each of the RuThes concept is represented as a set of synonyms or near-synonyms (plesionyms). The RuThes developers use a weaker term, ontological synonyms, to designate words belonging to different parts of speech (like stabilization, to stabilize), the items may be related to different styles and genres. Ontological synonyms are the most appropriate means to represent cross-linguistic equivalents (correspondences), because such approach allows us to fix units of the same meaning disregarding surface grammatical differences between them. For example, Table 1 represents basic ways of translating Russian adjective + noun phrases into Tatar.

Table 2. Examples of Russian A*dj + Noun* phrases and ways of translating them into Tatar

Russian unit	Corresponding Tatar unit	The structure of Tatar unit	English translation
Пенсионный возраст	Пенсия яше	$N + N_{\mathrm{POSS\_3}}$	Retirement age
Рабочий класс	Эшчеләр сыйныфы	$N_{\mathrm{PL}} + N_{\mathrm{POSS\_3}}$	Working class
Консульская служба	Консуллык хезмәте	$N_{\mathrm{NMLZ}} + N_{\mathrm{POSS\_3}}$	Consular service
Сексуальное меньшинство	Сексуаль азчылык	ADJ + N	Sexual minority
Именная стипендия	Исемле стипендия	$N_{\mathrm{COMIT}} + N_{\mathrm{PL}}$	Nominal scholarship

The TatThes is based on the list of concepts of the RuThes, i.e. the Tatar component is based on the list of concepts of the RuThes thesaurus. The methodology of compiling the Tatar part of the thesaurus includes the following steps:

1. Search for equivalents (corresponding words and multiword expressions) which are actually used in Tatar as translations of Russian items.
2. Adding new concepts representing topics which are important for the sociopolitical and cultural life of the Tatar society and which are not presented in the original RuThes (for example, Islam-related concepts, designations of Tatar culture specific phenomena, etc.).
3. Revising relations between the concepts considering the place of each new concept in the hierarchy of the existing ones and, if necessary, adding the new concepts of the intermediate level. So an important step is to check up the parallelism of conceptual structures between the languages.

The TatThes is mainly being compiled by manual translation of terms from the RuThes into Tatar, besides the Tatar language specific concepts and their lexical entries are added (about 250 new concepts). Search for equivalents in the Tatar language in many cases became a time-consuming task, because available Russian-Tatar dictionaries of general purpose contain obsolete lexical data (Galieva, Kirillovich, et al., 2017). So when compiling the lists of concept names and lexical entries we manually browsed large arrays of official documents and media texts in Tatar. In the process of compiling the Thesaurus, data from the following available Tatar corpora is used:

1. Tatar National Corpus (http://tugantel.tatar/?lang=en);
2. Corpus of Written Tatar (http://www.corpus.tatar/en).

In the course of the project we found that distinguishing feature of the contemporary Tatar lexicon is a great deal of absolute synonyms of different origin in and structure, the main cause of the phenomenon is language contacts (Galieva, Nevzorova, et al., 2017; Galieva, 2018).

The TatThes is implemented as a web application and has a special site (http://tattez.turklang.tatar/). Additionally, it has been published in the Linguistic Linked Open Data cloud as part of RuThes Cloud project (Kirillovich, et al, 2017). Currently the TatThes contains 10,000 concepts, and 6,000 of them provided with lexical entries.

4 Tatar Thesaurus Data for Wordnet Implementation: Case of Nouns

Previously, the RuThes thesaurus has been semi-automatically converted to the WordNet-like structure, and Russian wordnet (RuWordNet) has been generated (Loukachevitch, et al, 2016; Loukachevitch, et al., 2018). The conversion included two main steps:

1. the automatic subdivision of the RuThes text entries into three nets of synsets according to parts of speech;
2. the semi-automatic conversion of RuThes relations to WordNet-like relations.

The current version of RuWordNet (http://ruwordnet.ru/eng) contains 110 thousand Russian unique words and expressions. The same approach can be used to transform TatThes to Tatar wordnet.

The TatThes data may be serve as an initial basis for wordnet building by the following reasons:

1. The sociopolitical sphere covers a broad area of modern social relations. This area comprises generally known terms of politics, international relations, economics and finance, technology, industrial production, warfare, art, religion, sports, etc.
2. Currently the TatThes, in addition to terminology, comprises some general lexicon branches representing lexical items which can be found in various domain specific texts.
3. Semantic relations in the TatThes are necessary and sufficient to arrange the Tatar nominal vocabulary (nouns and noun phrases) as a wordnet-like network of synsets.

Thesaurus concepts unite synonymous items, so we have ready sets of synonyms as building blocks for wordnet. The concepts are linked by semantic relations with each other. In the RuThes and in the TatThes there are four main types of relationships between concepts, see Table 2. Semantic relations, mapped in wordnet, are not all shared by all lexical categories, so thesaurus data converting into wordnet format require dissimilar ways for different parts of speech.

Table 2. Semantic relations between nouns in thesaurus and in wordnets

Semantic relations in Thesaurus	Semantic relations in wordnets
Hypernym — hyponyms	Hypernym — hyponyms
Holonym — meronym	Holonym — meronym
Symmetrical association (Asc)	
Asymmetric association (Asc1/Asc2)	

Asc and Asc1/Asc2 association relations need additional explanations. The Asc symmetrical association, distinguished in RuThes and inherited by Tatar Socio-Political Thesaurus, connects very similar concepts, which the developers did not dare to combine into the same concept (for example, cases of presynonymy of items).

The Asc1/Asc2 asymmetric association connects two concepts that cannot be described by the relations mentioned above, but neither of them could not exist without the existence of the other (for example, a concept SUMMIT MEETING needs existing the concept HEAD OF THE STATE). In studies of ontologies this relation may be mapped as the ontological dependence relation.

Nevertheless, basic semantic relations which we need to group nouns concepts into wordnet are presented in the TatThes.

The core of the TatThes is made up of nouns and noun phrases (see Table 3), so the bulk of thesaurus data may be used for Tatar wordnet building without significant changes (synonymous items are yet joined into synsets and the required relations between them are selected).

Table 3. Number of noun concepts and noun phrase concepts in TatThes (on data of the Russian part).

Structure of TatThes items	Number items
Noun	3387
Adj + Noun	3135
Noun + NounGEN	352
Other	3126
Total	10000

An important issue is reflecting Tatar language specific word usage features in the resource. Presence alone of the shared concepts in languages do not necessarily evidences the same ways of usage of individual words or of usage words of individual semantic classes. Consider this with an example. Specific feature of the Tatar language is using of hypernyms before a corresponding hyponym, and such using is not regarded as pleonasm in many cases (examples 1–3):

(1) *Париж шәһәрендә* ‘in the city of Paris’ (instead of ‘in Paris’);

(2) *кыз кеше* ‘girl human’ (instead of ‘a girl’);

(3) *май аенда* ‘in the month of May’ (instead of ‘in May’).

Table 4. Representing lexical entries of month names in Thesaurus.

Rus concept name	Russian lexical entries	Rus POS	Tatar concept name	Tatar lexical entries	Tat POS
ДЕКАБРЬ	Декабрь ‘December’ Декабрьский ‘of December’	N ADJ	Декабрь	Декабрь ‘December’ Декабрь ае ‘month of December’	N NP
ЯНВАРЬ	Январь ‘January’ Январский ‘of January’	N ADJ	Гыйнвар	Гыйнвар ‘January’ Гыйнвар ае ‘month of January’ Январь ‘January’ Январь ае ‘month of January’	N NP N NP
ФЕВРАЛЬ	Февраль ‘February’ Февральский ‘of February’	N ADJ	Февраль	Февраль ‘February’ Февраль ае ‘month of February’	N NP

In cases when such a usage is conventionalized and corpus data evidences that the usage has a high frequency, we include such hyponym-hypernym items into a list of lexical entries of a concept. Such manner of designating is a feature of using toponyms and some classes of general lexicon, so it should be considered in Tatar wordnet building. For example, lexical entries of month names include such conventionalized noun phrases, composed of the month name and the hyponym, designating month in general, see Table 4.

Because the RuThes concepts assemble ontological synonyms, the RuThes lexical entries bring together words of different part of speech. Therefore, in standard case a Russian synset joins a noun (often we use it as a concept name) and a relative adjective derived from a noun (Table 5; only core items of synsets are represented). In Tatar, like in other Turkic languages, there is no original relative adjectives (and existing ones are borrowed from European or Oriental languages), so in many cases the TatThes synsets are composed of items of the same part of speech, mainly of nouns. This circumstance greatly facilitates cleaning thesaurus synsets data for wordnet developing.

Table 5. Typical arrangement of Russian and Tatar Thesaurus synsets.

Basic lexical entries of a Russian concept	Part of speech of Russian words	Basic lexical entries of a Tatar concept	Part of speech of Tatar words
Река ‘river’ Речной ‘of river, fluvial’	N ADJ	Елга ‘river’	N
Факультет ‘faculty’ Факультетский ‘of faculty’	N ADJ	Факультет ‘faculty’	N
Преподаватель ‘teacher Преподавательский ‘of teacher’	N ADJ	Укытучы ‘teacher’	N
Больница ‘hospital’ Больничный ‘of hospital’	N ADJ	Хастаханә ‘hospital’ Сырхауханә ‘hospital’	N N

So the core of the TatThes is made up of nouns and noun phrases (69% of total number of concepts). At the moment semantic relations between nouns mapped in thesaurus, are necessary and sufficient to convert Tatar thesaurus data into the wordnet format.

4 Conclusion

When building a wordnet for a new language, in particular, for a low-resource one, a crucial issue is searching for appropriate sources. We are planning to use data of the TatThes as a base resource for developing Tatar wordnet.

The TatThes is being compiled by manual translation of terms from the RuThes into Tatar, with searching Tatar equivalents used in real texts, so the thesaurus contains relevant lexical data. In the TatThes each concept is linked with a set of language expressions (single words or multiword expressions) which refer to the concept in texts – lexical entries.

The analysis of thesaurus data shows that the bulk of the thesaurus synsets are formed around nouns or noun phrases. A mapping semantic relations of nouns in the thesaurus reproduces a mapping semantic relations in wordnets.

Future work includes adding material of verbs and other parts of speech. Also we are planning to develop some automatic approaches to mining terms and to asses the Tatar terminology coverage in Thesaurus on Tatar socio-political texts data.

Acknowledgements

This work was partially funded by the subsidy allocated to Kazan Federal University for the state assignment in the sphere of scientific activities, grant agreement no. 1.2368.2017 and partially by a subsidy assigned to the Institute of Applied Semiotics of the Tatarstan Academy of Sciences for the state assignment.

References

1. Orhan Bilgin, Özlem Çetinoğlu, and Kemal Oflazer: Building a Wordnet for Turkish. Romanian Journal of Information Science and Technology **7** (1–2), 163–172 (2004).
2. Francis Bond and Ryan Foster: Linking and Extending an Open Multilingual Wordnet. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013), 1352–1362. ACL (2013).
3. Özlem Çetinoğlu, Orhan Bilgin, and Kemal Oflazer: Turkish Wordnet. In: K. Oflazer and M. Saraçlar (eds). Turkish Natural Language Processing. Springer (2018). doi:10.1007/978-3-319-90165-7_15
4. Razieh Ehsani. KeNet: A Comprehensive Turkish Wordnet and Using It in Text Clustering. PhD Thesis. Işık University (2018).
5. Razieh Ehsani, Ercan Solak, and Olcay Taner Yildiz: Constructing a WordNet for Turkish Using Manual and Automatic Annotation. ACM Transactions on Asian and Low-Resource Language Information Processing **17** (3), Article No. 24 (2018). doi:10.1145/3185664
6. Alfiya Galieva, Alexander Kirillovich, Bulat Khakimov, Natalia Loukachevitch, Olga Nevzorova, and Dzhavdet Suleymanov: Toward Domain-Specific Russian-Tatar Thesaurus Construction. In: R. Bolgov, N. Borisov, et al. (eds.) Proceedings of the International Conference on Internet and Modern Society (IMS-2017), pp. 120–124. ACM Press, New York (2017). doi:10.1145/3143699.3143716
7. Alfiya Galieva, Olga Nevzorova, and Dilyara Yakubova: Russian-Tatar Socio-Political Thesaurus: Methodology, Challenges, the Status of the Project. In: R. Mitkov and G. Angelova (eds.) Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP 2017), pp. 245–252. INCOMA Ltd., Varn (2017). doi:10.26615/978-954-452-049-6_034
8. Alfiia Galieva: Synonymy in Modern Tatar Reflected by the Tatar-Russian Socio-Political Thesaurus. In: J. Čibej, V. Gorjanc, I. Kosem and S. Krek (eds.) Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts (Euralex 2018), pp. 585-994. Ljubljana University Press (2018).
9. Alexander Kirillovich, Olga Nevzorova, Emil Gimadiev, and Natalia Loukachevitch: RuThes Cloud: Towards a Multilevel Linguistic Linked Open Data Resource for Russian. In: P. Różewski and C. Lange (eds.) Proceedings of the 8th International Conference on Knowledge Engineering and Semantic Web (KESW 2017). Communications in Computer and Information Science **786**, 38–52. Springer (2017). doi:10.1007/978-3-319-69548-8_4

10. Natalia Loukachevitch and Boris Dobrov: RuThes Linguistic Ontology vs. Russian Wordnets. In: H. Orav, C. Fellbaum and P. Vossen (eds.) Proceedings of the 7th Conference on Global WordNet (GWC 2014), pp. 154–162. University of Tartu Press (2014).
11. Loukachevitch, N.V., Dobrov, B.V., and Chetviorkin, I. I.: RuThes-Lite, a Publicly Available Version of Thesauru of Russian Language RuThes. In: Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference “Dialogue”, pp. 340–349. RGGU (2014)
12. Loukachevitch, N.V., Lashevich, G., Gerasimova, A.A., Ivanov, V.V., and Dobrov, B.V.: Creating Russian wordnet by conversion. In: Computational Linguistics and Intellectual Technologies: papers from the Annual Conference “Dialogue”, pp. 405-415. RGGU (2016).
13. Natalia Loukachevitch, German Lashevich, and Boris Dobrov: Comparing Two Thesaurus Representations for Russian. In: F. Bond, T. Kuribayashi, C. Fellbaum and P. Vossen (eds.) Proceedings of the 9th Global WordNet Conference (GWC 2018), pp. 35–44. Global Wordnet Association (2018).
14. Roberto Navigli and Simone Paolo Ponzetto: BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. Artificial Intelligence **193**, December 2012, 217–250 (2012).
15. Tufis, D., Cristea, D., and Stamou, S.: BalkaNet: Aims, Methods, Results and Perspectives. A General Overview. Romanian Journal of Information Science and Technology **7** (1–2), 9–43 (2004).
16. Piek Vossen (ed). EuroWordNet: General Document (2002).

Interactive Knowledge Graphs, Inductive Reasoning on Graphs

Victor Telnov[0000-0003-0176-5016] and Yuri Korovin[0000-0002-8399-4439]

National Research Nuclear University “MEPhI”, 249040 Obninsk, Russia
telnov@bk.ru

Abstract. The technologies of knowledge representation and inference in an artificial intelligence system focused on the domain of nuclear physics and nuclear power engineering are considered. The possibilities of description logics and graph databases of nuclear knowledge for the generation of cognitive hypotheses, using in addition to deduction and other ways of reasoning, such as inductive inference and reasoning based on analogies, are discussed. Interactive visual navigation and reasoning on the knowledge graphs are performed by means of special retrieval widgets and the smart RDF browser. Operations with semantic repositories are implemented on cloud platforms using SPARQL queries and RESTful services. The proposed software solutions are based on cloud computing using DBaaS and PaaS service models to ensure scalability of data warehouses and network services. Example of use of the offered technologies and software is given.

Keywords: Nuclear Education, Semantic Web, Knowledge Graph, Cloud Computing

1 Introduction and Motivation

Since the 1960s, in the framework of research on artificial intelligence, various formalisms for knowledge representation [1] (semantic networks, frame systems, etc.) have been developed. In 2019, the ontology description languages RDF, OWL, knowledge graphs and description logics [2] provide a modern theoretical basis for the creation of systems and methods of acquisition, presentation, processing and integration of problem–oriented knowledge in computer systems, which, in particular, is confirmed by the current standards W3C in the field of semantic web.

The reports of International Conference on Semantic Systems, International Workshops on Description Logic noted the growing interest of giants of the IT industry (Google, Facebook, Wikimedia) to graph models of knowledge representation and description logics. As of 2019 educational web–portals of universities, national centers for the exchange of scientific information, world nuclear data centers underused semantic web technologies. As for the inductive inference rules in graphs, the following considerations make them useful. First, inductive inference rules based on consideration of possible alternatives (precedents) allow to generate cognitive hypotheses (fuzzy knowledge) that cannot be obtained directly by deductive reasoning on the graph. Secondly, inductive inference is one of the basic technologies of semantic annotation of network content, when it is necessary to redesign, expand and update existing graphs with new knowledge. With the help of inductive inference the problems of classification and clustering of new concepts and individuals in the semantic base of nuclear knowledge are solved.

The aim of the work presented in the paper is to create a semantic web portal of knowledge in the domain of nuclear physics and nuclear power engineering based on ontology and using graph databases deployed on cloud platforms. The task of the study was to create the following graphs of nuclear knowledge:

- World nuclear data centers;
- Nuclear research centers;
- Events and publications from CERN;
- IAEA databases and network services;
- Nuclear physics at MSU and MEPhI;
- Nuclear physics journals;
- Joint nuclear knowledge graph.

The potential beneficiaries of information solutions and technologies that are proposed in the paper are students, teachers, experts, engineers, managers and specialists in the domain of nuclear physics and nuclear power engineering (target audience).

2 Problem Description and State of the Art

Today standard ontology markup languages are supported by mature semantics of DL along with a number of available reasoning algorithms [2]. However, some tasks in the ontology life cycle, such as their construction and/or integration, still largely delegated to knowledge specialists. For the successful development of semantic technologies it is desirable that the construction of the knowledge databases should be supported by automated inductive inference procedures, including entity classification and clustering tasks. The induction of structural knowledge like the taxonomies is not new in machine learning, especially for the task where clusters of similar objects are aggregated in hierarchies according to heuristic criteria or similarity measures. In the Inductive Logic Programming attempts have been made to extend relational learning techniques towards representations based on both clausal and description logics. These methods mostly are based on an empirical search and generally implement bottom–up algorithms that tend to induce overly specific concept definitions and narrowly specialized ontologies.

Generally, the problem of the induction of structural knowledge turns out to be a uneasy task in first-order logic or equivalent representations. In order to overcome the existing difficulties, the last decades have seen the development of research related to the calculation of similarity measures for concepts and individuals in ontologies. Similarity measure plays an important role in information retrieval and information integration as a means for comparing concepts and/or concept instances that can be retrieved or integrated across heterogeneous knowledge databases. It seems that the quite significant from a practical point of view results were obtained in [4]. To determine the similarity measure a set of similarity values has to be define, usually a set of the real numbers is used for this. Then it is required to determine a function for a pair of objects that will calculate the measure of their similarity. Formal definitions of similarity and dissimilarity measures were given in [5] and [6].

Naive semantic similarity can be defined as a path distance between entities in the hierarchical structure of the ontology. More meaningful methods to assess semantic similarity within a single ontology are feature matching and information content. There are measures have been developed to compute similarity values among classes belonging to different ontologies. For instance, a similarity function can detect similar entity classes by using a matching process, making use of special dictionaries, semantic neighborhood, and discriminating features. Of particular interest is the approach, aimed at finding commonalities among concepts or among individuals, employs the Most Specific Concept (MSC) method, that turns the instance checking task (that is deciding whether an individual is an instance of a concept) into a TBox reasoning problem [5].

Let there be a knowledge database $KB=\langle T, A\rangle$, contains two components: a TBox T and an ABox A. Let C and D be two concept descriptions in a T. Given a concept C in T, it is possible to consider its extension C^I, where I is the interpretation function. Further the canonical interpretation of the ABox is considered, when constants in the ABox are interpreted as themselves and different names for individuals stand for different domain objects. The semantic similarity measure [5] is defined as in the following:

Definition 1 (Semantic Similarity Measure). *Let L be the set of all concepts in DL and let A be an ABox with canonical interpretation I. The Semantic Similarity Measure s is a function*

$$s\colon L\times L\to[0,1]$$

which is defined as follows:

$$s(C,D)=\frac{|X^I|}{|C^I|+|D^I|-|X^I|}\max\left(\frac{|X^I|}{|C^I|},\frac{|X^I|}{|D^I|}\right),$$

where $X=C\sqcap D$ and $(\cdot)^I$ computes the concept extension w.r.t. the interpretation I.

The measure can be explained as follows. In case of semantic equivalence of the concepts C and D, the maximum value of the similarity will be calculated. In case of disjunction, the minimum value of similarity will be assigned because the two concepts are totally different: their extensions do not overlap. Finally, in the case of overlapping concepts, a value in the range]0, 1[will be computed [5].

Definition 2 (Most Specific Concept). *Let there be a knowledge database $KB=\langle T, A\rangle$. Given an ABox A and an individual a, the Most Specific Concept of a w.r.t. A is the concept C, denoted $MSC_A(a)$, such that $A|=C(a)$ and $\forall D$ such that $A|=D(a)$, it holds: $C\sqsubseteq D$. Here |= stands for the standard semantic deduction.*

Once the most specific $MSC_A(a)$ of an individual a is known, to decide if $KB|=D(a)$ holds for an arbitrary concept D, it suffices to test if $T|=MSC_A(a)\sqsubseteq D$. This method, unfortunately, loses its simplicity and efficiency when applied to large and complex ontologies, as it tends to generate very large MSCs that could lead to intractable reasoning.

Let c and d two individuals in a given ABox. Then it is possible to calculate C=MSC$_A(c)$ and D=MSC$_A(d)$. According to [5], now the semantic similarity measure s can be applied to these concept descriptions, thus yielding the similarity value of two instances:

$$\forall c,d\text{: } s(c,d)=s(C,D)=s(\text{MSC}_A(c),\ \text{MSC}_A(d)).$$

The similarity value between a concept C and an individual a can be computed by determining the MSC of the individual and then applying the similarity measure:

$$\forall a\text{: } s(a,C)=s(\text{MSC}_A(a),C).$$

The complexity of s calculation depends on the complexity of the instance checking task for the adopted DL language, denote it as $C(\mathit{InstanceChecking})$. Similarity between concepts: s is a numerical measure, all calculus have constant complexity, instance checking is repeated three times: for concepts C, D and their intersection [5], so:

$$C(s)=3\cdot C(\mathit{InstanceChecking}).$$

Similarity between an individual and a concept: in this case, besides of the instance checking operations required by the previous case, the MSC of the considered individual is to be computed. Thus, denoted by C(MSC) the complexity of the MSC computing, get the complexity estimate:

$$C(s)=C(\text{MSC})+3\ C(\mathit{InstanceChecking}).$$

Similarity between individuals: this case is analogous to the previous one, the only difference is that now two MSC is to be computed for the arguments. So the complexity in this case is:

$$C(s)=2\cdot C(\text{MSC})+3\cdot C(\mathit{InstanceChecking}).$$

From the previous formulas it is clear that the computational complexity of the similarity measure sensitive to the choice of the DL. For the ALC logic, $C(\mathit{InstanceChecking})$ has polynomial complexity. Computation of the MSC also implies instance checking and depends on algorithm properties.

3 Methodological Approach and Technologies

3.1 Semantic Repositories, Search Widgets, Intelligent RDF Browser

From a practical point of view, knowledge graphs are placed in the data warehouse, which are called RDF-repositories or triple repositories. The project [3] largely uses the Google Cloud Platform and Apache Jena framework on the free quota with each of the repositories serviced by a dedicated virtual machine. Remote asynchronous work with the Google Cloud Platform is performed using the standard SPARQL 1.1 query language through application programming interfaces in Java and JavaScript. Common operations are creating, reading, updating and deleting data in knowledge graphs. For the practical implementation of network requests to repositories, HTTP protocol methods GET and POST are used.

Each of the knowledge graphs contains thousands of triplets. Search widgets, shown in [7], allow users to get to the right place of a specific knowledge graph, where the desired information objects will be detected and visualized. The principle of operation of search widgets is similar to the way information samples from the web using popular search engines (Google, Yandex, etc.). As the user types the characters of the keywords in the search widget's input line, the system rolls out an adequate list of entities from the corresponding knowledge graph. The user is prompted to select a suitable concept or individual and dive directly into the desired area of knowledge graph. Thereafter, a more accurate interactive visual navigation through the graph and inductive reasoning on graph becomes possible, which is implemented in an intuitive way using the intelligent RDF browser, as described below.

3.2 Interactive Reasoning on the Graph of Knowledge (Example)

The RDF browser is an essential attribute of the project [3], which distinguishes it from other known solutions in the field of semantic web. Once on the desired location of the desired knowledge graph using the search widget, then the user through the RDF browser can perform visual navigation on the graph, visiting its nodes in the correct order and extracting metadata, hypertext links, full-text and media content associated with the node. In this case, the neighborhood (environment, closure) of each node of the graph becomes visible and navigable. This neighborhood includes the nodes of the graph, through which the user initially entered the semantic web, as well as adjacent nodes of other graphs that are supported by the knowledge database.

The visual way of specifying the inference rules on the graph makes it stand out from the more traditional known reasoner's interfaces, where inference rules are specified using SWRL language, logical predicates or a SPARQL-like syntax. It seems, that the intuitively clear interactive visual way of specifying inference rules is more friendly for unsophisticated users of knowledge graphs.

As an example, consider the following situation. Some student is preparing to pass the exam in nuclear physics at the Physics Faculty of Moscow State University. Let the student know only the name of the training course: “Physics of the atomic nucleus and particles” and the name of the professor: “I.M. Kapitonov”. Let us formulate the task: using the semantic educational web portal [3], it is necessary to find and study all the video lectures of this professor on this training course. Also suppose, that the student discovered in YouTube a video lecture titled “Lecture 1. Physics of the atomic nucleus and particles”. He suggests, that this video lecture may be relevant to the training course being studied. Let us formulate the hypothesis: “Lecture 1. Physics of the atomic nucleus and particles” is taught by the professor “I.M. Kapitonov” at the

Physics Department of Moscow State University and it is included in the training course “Physics of the Atomic Nucleus and Particles”.

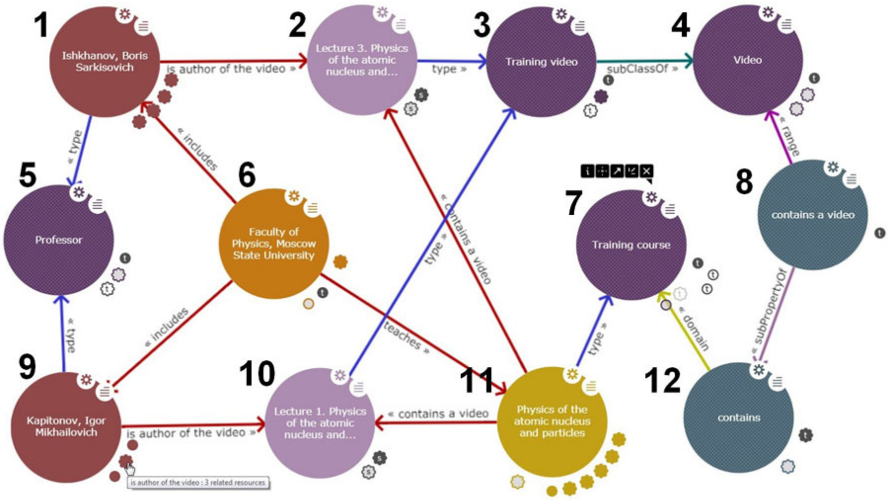

1
Ishkhanov, Boris Sarkisovich
is author of the video »
2
Lecture 3. Physics of the atomic nucleus and...
type »
3
Training video
subClassOf »
4
Video
« type
« includes
5
Professor
6
Faculty of Physics, Moscow State University
« contains a video
type »
7
Training course
8
contains a video
« range
« type
« includes
teaches »
« domain
« subPropertyOf
9
Kapitonov, Igor Mikhailovich
10
Lecture 1. Physics of the atomic nucleus and...
11
Physics of the atomic nucleus and particles
12
contains
is author of the video »
« contains a video
type »
is author of the video : 3 related resources

Fig. 1. Fragment of the knowledge graph titled “Nuclear Physics at MSU, MEPHI” as an example of the implementation of inductive reasoning on graph

To solve the task and test the validity of the hypothesis, it is necessary to perform the following obvious inductive reasoning on the knowledge graph step by step.

Step 1. Go to the educational web portal [3] and select the knowledge graph “Nuclear Physics at MSU, MEPHI”. It is possible to start the reasoning either with the classes “Training course”, “Training video”, “Professor”, etc. or with the specific entities “Physics of the atomic nucleus and particles”. “Lecture 1. Physics of the atomic nucleus and particles”, “Kapitonov”, etc.

Step 2. Let the student decided to begin the reasoning from the class “Training course”. The RDF browser workspace opens and the node of the graph with this name appears.

Step 3. The appeared node of the graph in Fig. 1 is shown under number 7. We are interested in objects belonging to the class “Training course”. There are three such objects and they are associated with our node by the “type” property. With a mouse click, we will open the object that is taught at the Physics Faculty of Moscow State University (see nodes with numbers 11 and 6 in Fig. 1).

Step 4. Continuing to similarly disclose neighboring nodes for the object “Physical Faculty of Moscow State University” (node number 6 in Fig. 1) by the “includes” property, for the “Kapitonov” object (node number 9 in Fig. 1) by the “is author of the video” and / or for the object “Physics of the atomic nucleus and particles” (node number 11 in Fig. 1) by the property “contains a video”, the student finally will make sure of the validity of the hypothesis and get the solution to the task, see node number 10 in Fig. 1.

Step 5. The result obtained in Step 4 could also be achieved in the course of deductive reasoning, without considering possible alternatives. However, the use of inductive inference allows one to naturally extract from the graph additional knowledge that will not be easy to obtain with a simple deductive inference. Acting as described above, it is easy to find that some video lectures on the training course “Physics of the atomic nucleus and particles” at the Physics Faculty of Moscow State University are also taught by professor B.S. Ishkhanov, see node 1 in Fig. 1. All video lectures and other learning materials of both professors for this training course became available. Through the knowledge graph, the full content of any training course is visually revealed and all the relationships are graphically shown.

As was shown in the above example, the process of inductive inference on knowledge graphs resembles a computer adventure game, does not require special skills, and is accessible to the inexperienced user. Knowledge graphs, similar to the above, are used in the educational process at the NRNU MEPHI. Practice shows that university students master the techniques of interactive work with knowledge graphs within a few minutes.

4 Related Works and Conclusion

University of Manchester, Stanford University, University of Bari and a number of other universities are focused on the issues of theory development and technology's implementation for semantic web, description logics and incarnations of the ontologies description language OWL. Special mention should be made on the project [4], where for the first time an attempt was made to put into practice the methods of inductive reasoning for the purpose of semantic annotation of content from the web. To date, such network services are offered by some software companies.

As for the issues of visualization linked data [8], here one of the first successful projects was Lodlive [9], which provided a tool for easier surfing through the DBpedia knowledge base. It is important continue to develop and improve tools for intuitive perception of linked data for non-professionals. VOWL [10] is one of the modern project for the user-oriented representation of ontologies, it proposes the visual language, which is based on a set of graphical primitives and an abstract color scheme. LinkDaViz [11] propose a web-based implementation of workflow which guides users through the process of creating visualizations by automatically categorizing and binding data to visualization parameters. The approach is based on a heuristic analysis of the structure of the input data and a visualization model facilitating the binding between data and visualization options. The resulting assignments are ranked and presented to the user. SynopsViz [8] is a tool for scalable multi-level charting and visual exploration of very large RDF & Linked Data datasets. The adopted hierarchical model provides effective information abstraction and summarization. Also, it allows efficient –on the fly– statistic computations, using aggregations over the hierarchy levels.

In contrast to the above solutions, the project [3] is mainly focused on the implementation in educational activities of universities and is not limited to visualization of knowledge graphs and interactive navigation, but is aimed at the introduction of the latest semantic web technologies to the learning process, taking into account the achievements in the field of uncertainty reasoning.

Acknowledgments

The reported study was funded by Russian Foundation for Basic Research according to the research projects Nos. 18–07–00583, 19–47–400002, by Government of the Kaluga Region according to the research project No. 19–47–400002 and was funded by Vladimir Potanin Foundation according to the project No. GC190001383.

References

1. Harmelen, F., Lifschitz, V., and Porter, B.: Handbook of Knowledge Representation, England, Oxford: Elsevier Science Oxford, 2008.
2. Baader, F. Calvanese, D., McGuinness, D. Nardi, D. and Patel-Schneider, P.: The Description Logic Handbook: Theory, Implementation and Applications, 2nd Ed.; Publisher: Cambridge University Press, New York, USA, 2010.
3. Semantic educational web portal. Graphs of nuclear knowledge. Intelligent search agents (2018), http://vt.obninsk.ru/x/, last accessed 2019/04/15
4. d'Amato, C., Fanizzi, N., Fazzinga, B., Gottlob, G., and Lukasiewicz, T.: Combining Semantic Web Search with the Power of Inductive Reasoning (2013), http://ceur-ws.org/Vol-527/paper2.pdf, last accessed 2019/04/15
5. d'Amato, C., Fanizzi, and Esposito, F.: A Semantic Similarity Measure for Expressive Description Logics (2009), http://arxiv.org/pdf/0911.5043v1.pdf, last accessed 2019/04/15
6. Minervini, P., d'Amato, C., Fanizzi, and Tresp, V.: Discovering Similarity and Dissimilarity Relations for Knowledge Propagation in Web Ontologies, Journal on Data Semantics **5** (4), 229–248 (2016).
7. Telnov, V.: Semantic Educational Web Portal, Selected Papers of the XIX International Conference on Data Analytics and Management in Data Intensive Domains (DAMDID/RCDL 2017), Moscow, October 10–13, 2017, 50–56, online: http://ceur-ws.org/Vol-2022/paper11.pdf, last accessed 2019/04/15
8. Bikakis, N. and Sellis, T.: Exploration and Visualization in the Web of Big Linked Data: A Survey of the State of the Art (2016), http://arxiv.org/pdf/1601.08059.pdf, last accessed 2019/04/15
9. Camarda, D.V., Mazzini, S., and Antonuccio, A.: Lodlive, exploring the web of data, In Proceedings of the 8th International Conference on Semantic Systems, I-SEMANTICS, ACM, 5–7 September, 2012, Graz, Austria, pp. 197–200.
10. Schlobach, S. and Janowicz, K.: Visualizing ontologies with VOWL, Semantic Web **7** (4), 399–419 (2016).
11. Thellmann, K., Galkin, M., Orlandi, F., and Auer, S.: LinkDaViz – Automatic Binding of Linked Data to Visualizations, Proceedings of the 15th International Semantic Web Conference, Bethlehem, PA, USA, October 11–15, 2015, pp. 147–162.

DACOMSIN WORKSHOP

DATA FORMATS, METADATA AND ONTOLOGIES IN SUPPORT OF MATERIALS RESEARCH

Data Curation Approach to Management of Research Data. Use Cases for a Upgrade of the Thermophysical Database THERMAL

A.V. Kosinov[1], A.O. Erkimbaev[1], G.A. Kobzev[1], and V.Y. Zitserman[1]

[1] Joint Institute for High Temperatures, Russian Academy of Sciences, Russia
vz1941@mail.ru

Abstract. Procedures are considered to support extensive archives of digital data called “data storage”. Particular attention is paid to the support of scientific data. It is shown that the activities aimed at updating the thermophysical THERMAL database correspond to the approaches provided by the “data curation”. A communication system for metadata with external ontology is proposed. The new version of metadata provides the possibility of multilateral assessment of the origin, quality and status of scientific data. It is shown that the use of new metadata provides a significant increase in the value of these studies.

Keywords: Research Data, Data Curation, Data Quality, Thermophysical Database

1 Introduction

Digital data uploaded to repositories or databases require permanent procedures that guarantee safety, quality top-level and enduring access to data. The set of such procedures is the subject of particular activity of managing the use of data called **data curation.** This term is rarely used in the Russian literature, although all the necessary actions for data integrity and management are of course performed to support digital repositories or databases. The meaning and content of the concept of curation of data can be revealed by referring to the history of its appearance. This concept originates from the Museum's practice, which is traditionally based on the curator's work on preservation, renovation and description of exhibits.

As for the term “data curation”, apparently, it first appeared in the article by Diana Zorich [1], which pointed to the common problems facing libraries, museums and research centers involved in supporting digital collections. According to [1], digital archives, as well as their supporting tools (vocabularies, thesauruses, metadata) should be regularly monitored and updated for data consistency, maintaining its quality, availability, etc., and activities in this direction is the essence of curating process.

Oddly enough, digital data is subject to erosion, as are physical artifacts, manuscripts or museum exhibits. It can be related to the use of outdated metadata, terminology, dictionaries, formats, software, as well as the absence of references to more relevant documents or external resources. By analogy with engineering, it can be considered as technological obsolescence (or deterioration) of the data structure, file formats, software, etc. An unrecoverable failure of storage media (bit rot in IT slang) during data storage is possible in parallel with content obsolescence.

History of the “Digital Curation” concept and its gradual adoption in data manager’s community is considered by [2]. In particular, the concept of “data curation” here is clearly separated from the narrower and service-oriented concepts: “archiving” and “preserving”. Unlike the latter, curation involves not only the preservation and maintenance of digital storage, but its indispensable enrichment by expanding the functions and content. For example, a common and effective way of enriching content is to place it in a wider context by linking the data set to thematically related resources, so called **Contextualizing**.

Among the main objectives of data curation, as a rule, such are mentioned as their storage, description, safety measures, the so-called cleaning, that is, monitoring and restoring quality, as well as a number of other measures. The expanded definition of the Digital Curation Center [3] covers all activities related to data management, starting with data creation, digitization, documentation, and accessibility and future reuse. The detailing of these processes, carried out in [4], allowed us to identify about 50 curation practices, most of which also fall into such categories as data preservation, data cleaning, and finally, description in terms of the complex structure of metadata.

1 Curation of Research Data

The purpose of this report is to discuss the specific recipes foreseen in the framework of digital curation in the implementation of the project for updating the Thermophysical Database THERMAL. The project considered in the report at the previous conference [5] includes: a significant expansion of the database volume due to the rejection of the restrictions adopted at the stage of creation in the 70s of the last century; creation of tools for flexible variation of the data structure, reflecting the uniqueness of objects and their characteristics; transition to a new platform that allows you to store and process data of various structures and formats. A significant part of the activities performed during the project, in fact, refers specifically to the data curation, as it comes down to checking and correcting old documents in accordance with the newly adopted format, vocabulary and requirements for data completeness and quality.

In general, curation refers to digital objects of arbitrary origin and kind. Therefore, numerous measures for data preservation such as regular back up, defect detection at the bit level, overcoming technological obsolescence of hardware or file formats are applicable in all cases regardless of the content. In solving scientific problems, the curation process provides not just conservation, but confirmation and reliable expansion of previous data of the experiment or simulation. On the contrary, the absence or poor quality of the curation process inevitably leads to the loss, distortion or misinterpretation of data.

A brief list of features and capabilities of “data curation” as applied to e-Science (or Data-intensive science) was given by the Digital Curation Centre [6]. In this list, the specificity of scientific data, to a certain extent, is taken into account at all stages of the curation process. For example, long-term data storage may require replacement of obsolescence of storage devices, which has already been encountered in astronomy. At the stage of data cleaning (that is, data correction and updating), it is important to establish links between different versions of evolving datasets or between primary and secondary data. However, the most noticeable specificity of data curation is manifested in their description, that is, in the composition and structure of metadata.

2 Metadata (Update and Extension)

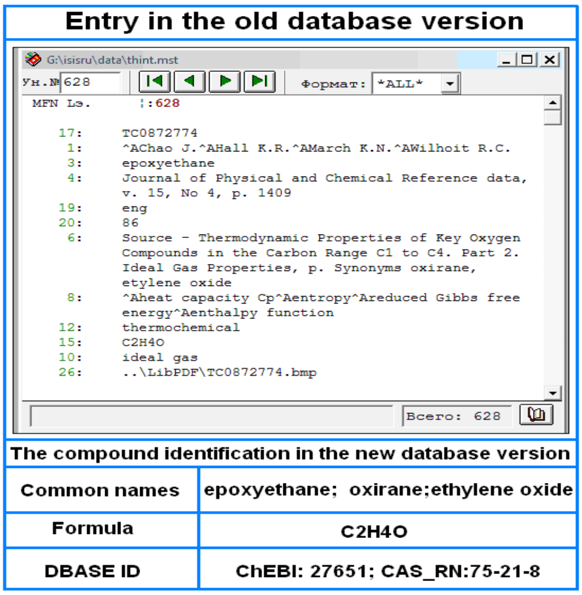
In general terms, metadata document the context and record information about how the data was obtained and what processing and verification procedures were performed during the retention period. There is an extensive literature on scientific metadata and their use in various disciplines [7–9]. Metadata, accompanying subject information, allow you to: identify a dataset with its position in the repository; define access rules; describe the logical structure and data formats; to ensure the operation of various data analysis tools. Metadata standards for different disciplines and types of documents are collected in the catalog (rd-alliance.github.io/metadata-directory/) and the "Disciplinary metadata" section of the Digital Curation Center (www.dcc.ac.uk/resources/metadata-standards). Both mentioned sources contain also references to domain-agnostic standards for formal description of digital resources (e.g. the **Dublin Core metadata set**), or for the identification and citation of digital resources (e.g. **DataCite Metadata Store**). Metadata for thermophysical properties (ThermoML), characteristics of ordinary materials (MatML) and nanomaterials are described in detail in [7, 10–12].

Regardless of the subject area, scientific metadata must satisfy a number of requirements that guarantee sufficient completeness and accuracy in the presentation of each data set. Relevant elements should provide: unambiguous identification of the object of study; presentation of information about the source and data acquisition method (research method, equipment, program code etc); uncertainty and data quality information; linking with controlled vocabularies or ontologies; the possibility of flexible adjustment to the features of the object and its characteristics. The expansion of metadata carried out in the updating the database THERMAL, provides for the implementation of each of these requirements, see Table 1.

First of all, in the progress of curation, the possibilities of identifying objects are expanded, which can now include, along with inorganic substances, complex organics, natural and industrial materials, and so on. The "**Identification**'' metadata element provides, along with a stoichiometric formula, the use of several common names (synonyms), as well as links to publicly available databases. The pointer to the database and the corresponding identifier uniquely identify the object, providing, in addition, access to information that complements the information stored in THERMAL. As an example in Fig. 1 the identification of the compound called **epoxyethane** (oxirane, ethylene oxide) in the old and new versions of the database is shown.

Table 1. A comparison between old and updated metadata versions

Old metadata set	**New metadata structure**	
Unique record ID	Unique record ID	
	Data type [bibl, full-text, factual]	
	Data status [experiment or simulation, predicted, critical evaluated, recommended, stale]	
	Research type [experimental, theory, simulation, review]	
Source	Provenance	Source [bibl, database, external agency]
		Data origin [method, equipment, software]
Abstract		
Stoichiometric formula	Identification [common names, stoichiometric formula, public database ID]	
Substance class	Linking to ontology classes [Linking to sub- classes of the **Chemical_entity**]	
Properties	Linking to ontology classes [Linking to sub- classes of the **Quality**]	
Properties type		
Phase	Linking to ontology classes [Linking to sub-classes of the **State_of_matter**]	
Phase transition	Linking to ontology classes [Linking to sub-classes of the **Transition**]	
	Data Quality	Uncertainty [type, value]
		Data quality attributes (timeliness, reliability, currency, completeness etc)
	Data Features [SubstanceFeatures, Sample, Influence Factors]	
	References	Full-text
		Tables or equations
		External documents (from Web or Server)



Entry in the old database version

```
MFN Lэ.        ¦:628

   17:     TC0872774
    1:     ^AChao J.^AHall K.R.^AMarch K.N.^AWilhoit R.C.
    3:     epoxyethane
    4:     Journal of Physical and Chemical Reference data,
           v. 15, No 4, p. 1409
   19:     eng
   20:     86
    6:     Source - Thermodynamic Properties of Key Oxygen
           Compounds in the Carbon Range C1 to C4. Part 2.
           Ideal Gas Properties, p. Synonyms oxirane,
           etylene oxide
    8:     ^Aheat capacity Cp^Aentropy^Areduced Gibbs free
           energy^Aenthalpy function
   12:     thermochemical
   15:     C2H4O
   10:     ideal gas
   26:     ..\LibPDF\TC0872774.bmp
```

The compound identification in the new database version

Common names	epoxyethane; oxirane;ethylene oxide
Formula	C2H4O
DBASE ID	ChEBI: 27651; CAS_RN:75-21-8

Fig. 1. Identification of the substance “ethylene oxide” in the old (top) and new (bottom) versions of the database.

Another example is the natural mineral **mullite**, where variations of the elemental composition allow the use of several stoichiometric formulas (for instance, Al6Si2O13 и Al4SiO8), and the exact identification is provided by the record (URL: www.webmineral.com/data/Mullite.shtml#.WaMcG_hJaUk) in the mineralogical database WEBMINERAL.

A more complete identification is provided by linking metadata with ontology classes, which includes entities that reflect the types of objects, their states and properties, Fig. 2.

In particular, **chemical_entity** identifies types of substances or materials, focusing on the systematics adopted in chemistry (elements, oxides, acids ...), as well as on categories determined by properties or by application (polymer, solution, mixture, refrigerant, fuel ...). A pointer to subclasses in relation to the class **mixture** allows to identify binary and multicomponent alloys and solutions, for example, such relevant objects in thermophysics as air, humid air, combustion products, etc. The **State_of_matter** class allows you to detail the phase and type of the crystal lattice based on an extensive hierarchy of child classes.

Similarly, linking to classes that inherit the **Quality** and **Transition** classes reflects the rich variety of physical properties inherent in an object and the phase transitions that occur in it. It is essential that linking of metadata to ontology during the curation provides unambiguous interpretation of terms and concept, and through editing of ontology, the possibility of flexible adjustment in connection with the emergence of new objects and concepts. For example, concepts such as **second critical point** [13],

topological insulator [14] or previously unknown allotropic forms of carbon (metallic carbon, T-carbon) have recently been included.

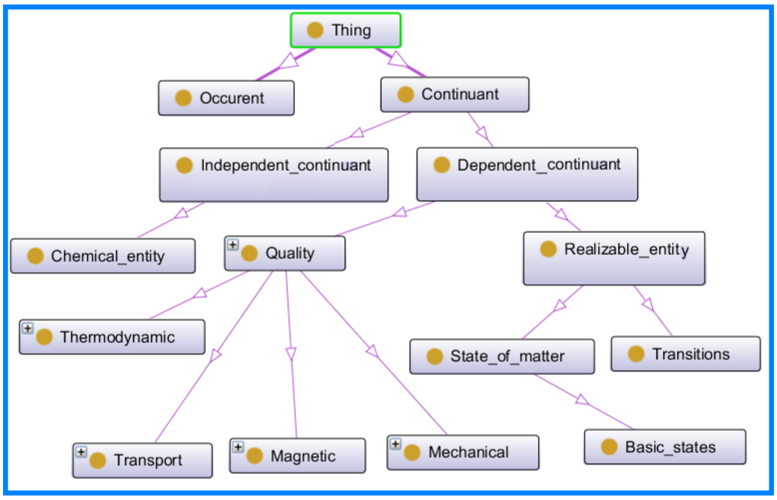


Fig. 2. Fragment of the ontology – top level classes

3 Data Quality and Features

New curatorial perspectives have emerged with concepts such as "**Data Quality**" and "**Data Features**" in the metadata set, see Table 1. Alternative choices for evaluation the scientific data quality are discussed in detail in article [12]. It was shown that the best way to assess the research data quality was to combine the traditional uncertainty assessment with a certification procedure based on several quality attributes, every of them represents some aspect of quality. It is necessary to select a specific metric for the multidimensional certification of the dataset and the data compliance indicator for each of the quality attributes. In many cases, it is useful to use a domain-agnostic metrics reflecting quality factors that are important to data consumers, for example accuracy, timeliness, reliability, completeness, relevancy, interpretability etc. Such an approach to quality certification is most justified in interdisciplinary projects, for example, when integrating thermal data with the performance characteristics of structural materials.

However, when you update the thermophysical database THERMAL, you can reduce the number of quality attributes. Data certification proposed by the authors [15] in relation to physico-chemical properties is based on the use of three attributes: completeness of information about the state and preparing the sample in the experiment; completeness of the method and measuring instruments description or codes and data processing in the case of simulation; the consistency of the numerical data with ground rules and regularities as well as with previous fairly reliable measurements. Combined with uncertainty assessment, certification identifies three main aspects of data quality: accuracy, completeness, and consistency. The technique provides a generalized evaluation of the data set, by assigning each of the attributes of the quality level (high, medium and low), focusing on compliance with the requirements of completeness and consistency. The expert gets the opportunity, based on this data curation technique [15] select the data set of top-quality by assigning them the special **Recommended** status (for what to use the **Data Stutus** element). The information needed to assess the data quality allows it to be carried out only for two types of data (Full-text and Factual) specified in the **Data Type** element, Table 1. At the same time, for unstructured data in the form of the text of an article (full-text data) it is justified to conduct only certification with indication of quality attributes, but without its own assessment of uncertainty.

Table 2. Examples of the non-standard data sets

Data Set Title	Data Feature
Amorphous polymeric nitrogen-toward an equation of state	SubstanceFeatures [amorphous, polymeric]
Melting point of high-purity germanium stable isotopes	SubstanceFeatures [stable isotopes]
Relationship between changes in the crystal lattice strain and thermal conductivity of high burnup UO2 pellets	Sample [pellet] Influence Factors [high burnup]
Study of near-critical states of liquid - vapor phase transition of metals by isentropic expansion method of shock-compressed porous sample	Sample [porous]
Thermophysical properties of liquid Co measured by electromagnetic levitation technique in a static magnetic field	Influence factors [field]
Phase diagram of water under an applied electric field	Influence factors [field]
Shock compression of preheated molybdenum	Influence factors [prehistory]

The concept of **Data Features** in a metadata set is based on other data evaluation criterion. It allows you to select those data sets where there is any deviation from the standard (i.e. an anomaly) in the characterization of the object or its properties. The well-defined specificity (features) that distinguishes one dataset from another allows you to overcome the inevitable contradiction between structured data and poorly formalized information hidden in context. As can be seen from the Table 1, “**Data Features**’’ element includes three groups of the features: SubstanceFeatures, Sample, Influence Factor. The first allows you to extend the traditional substance identification, indicating the isomeric form, nonstoichiometry, isotopic composition, etc. The indication **Sample** includes pointers to the features of the sample: shape, size, surface condition, prehistory, etc. Finally, the **Influence factor** sign includes pointers to external factors that determine the experiment and properties of the substance: external field, mechanical load, environment, radiation, etc. Some examples of non-standard data sets from Table 2 illustrate the signs defining the specifics of a substance and its properties. In so doing the specificity can be attributed to any data set of the three types indicated in the **Data Type** element (Table 1), in contrast to the quality assessment, depending on the type of data.

4 Data Cleaning and Preservation

The update procedure of the THERMAL database, includes (along with the volume expansion) revision of old records based on the new metadata system. As a result, the data curator needs to check the completed conversion. This activity, called “**Data Cleaning**” (or cleansing), involves the detection and correction of “dirty”, that is distorted or incomplete data [16]. Data pollution occurs for various reasons, among which may be distortions of old records (input errors and duplication, incorrect data distribution by fields) and errors when using new metadata to determine the object and properties, see Table 1.

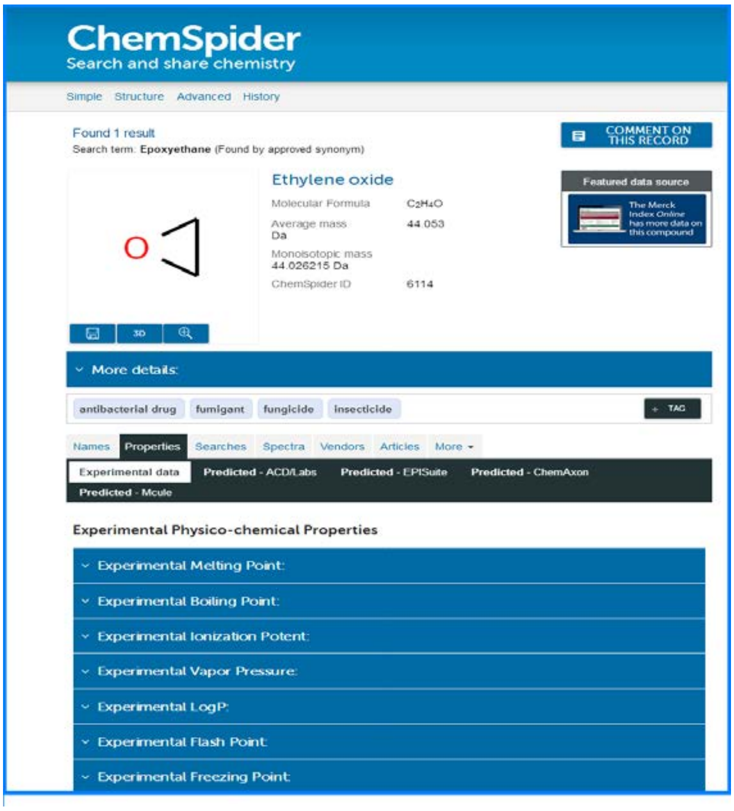
Previous data were entered without the use of controlled dictionaries, so the most important task of cleaning is to eliminate ambiguity in terms and concepts, subjecting them to ontology classes. For example, earlier a whole set of lexical elements (including English and Russian terms) was associated with the concept of dynamic compressibility, namely: **Hugoniot**, **Hugoniot data**, **Hugoniot adiabat**, **shock Hugoniot**, **shock adiabat**, **shock compression**, **release isentrope** etc. Linking this whole set of terms with a single ontology class (**Dynamic Compressibility**) eliminates synonymy and dramatically facilitates semantic search. The same procedure requires that the names of substances used in different records now appear in each record as a set of common names, which eliminates search losses.

In addition to linking and unifying terms, data cleaning also provides for the correction of content, first of all, the correction (or fixation) of obviously obsolete numerical data. At the same time, old records with correct filling of fields have historical value, even with outdated data. Therefore, measures are proposed to clean the data, excluding the physical removal of the record. One of them is to add to the old record an indication of the unreliability of the data by linking it with later (or network resources), including reliable data. Another measure is to assign the obsolete data to the sign “**low quality**” using the attribute “**consistency**” (see above “**Data Quality**”). Finally, you can assign the status to “**Stale**” or “**Recommended**”, which allows you to immediately separate high-quality data from clearly obsolete data during the search.

Regular data cleaning inevitably requires entering them into a new context, explaining concepts, offering an introduction to the available handbooks, databases, manuals, and so on. In the list of data curation practices such activity was named as

Contextualizing, i.e. “Use metadata to link the data set to related publications, dissertations, and/or projects that provide added context for how the data were generated and why”. A new set of metadata (Table 1) allows linking with external resources through two elements, “**Identification**” and “**References**”. The first uses a link to Public Databases to accurately identify a substance with access to additional data.

For example, by including the reference CSID: 6114 (ID from Database ChemSpider) in the “**Identification**” field, we can select “**Ethylene oxide**” from the group of substances with the same formula C2H4O, gaining access to the reference data, Fig. 3.



Found 1 result
Search term: **Epoxyethane** (Found by approved synonym)

Ethylene oxide

Molecular Formula C_2H_4O
Average mass 44.053 Da
Monoisotopic mass 44.026215 Da
ChemSpider ID 6114

Experimental Physico-chemical Properties

- Experimental Melting Point:
- Experimental Boiling Point:
- Experimental Ionization Potent:
- Experimental Vapor Pressure:
- Experimental LogP:
- Experimental Flash Point:
- Experimental Freezing Point:

Fig. 3. Typical entry for ethylene oxide from ChemSpider (CSID 6114)

The **References** element (Table 1) provides a link to thematically pertaining resources, but without requiring exact identification of the object. An example is the linking with the Sacada database **[Samara Carbon Allotrope Database, http://sacada.sctms.ru/]**, which expands the set of information on carbon allotropes presented in the THERMAL database**.**

Obviously, contextualization as an data curation practice requires the participation of human experts, but not programmers. Therefore, the term “clearing” within the framework of data curation means not only the rejection of dirty data, but also data analysis and decision making. Thus, contextualization as an element of data curation completely corresponds to the expression “added value to digital research data throughout its lifecycle”, in response to the question “What is digital curation?”

4.1 Preservation

Long-term storage, to a large extent, a purely technological problem related to service life of memory devices (a maximum of 100 years), the solution of which requires significant funding. The required measures include regular backup and failure detection at the bit level. Particular attention should be paid to the physical protection of the data storage, as the frequency of bit rot in data significantly increases due to pollution, thermal and radiation exposure and other external agents. Some of the protections activities are most adequate for research data. Among them are format migration (i.e. consistent change in line with technological changes) and emulation recreating outdated hardware and software on a modern platform.

When transferring the THERMAL database to the Big Data platform, the obsolete ISO-2709 format, adopted in the 60s of the last century [17] as ISO standard for bibliographic description, is discarded. In the new version, documents in ISO format are converted to structured text in JSON format, one of the most convenient for exchanging data and metadata [5]. The advantage of a text document is the possibility of simple reading and editing, accessibility for human perception, convenient form of storage and exchange of arbitrary structured information. The JSON-format (unlike ISO) is convenient for storing factual information in the form of tables and nested structures, as well as numerous links to files of different formats (images, presentation files, Web-pages, etc.), which is especially important when expanding the functions of the THERMAL database. It is also important that the JSON format is a working object for some platforms, in particular for **Apache Spark**, allowing for the exchange, storage and queries for distributed data. There is already an experience of using structured text as a means of thermal properties data interchange [12].

Along with the obsolescence of formats, outdated software that is incompatible with more modern platforms affects long-term storage. This fully applies to the database THERMAL, built on the basis of the documentary Database Management System (DBMS) CDS/ISIS [17] with a fairly limited scope, mainly for the storage of catalogue cards. The transition to the Apache Spark platform (http://spark.apache.org/docs/) in combination with ontology-based data management opens up much greater opportunities in storing and integrating data of arbitrary format, converted into JSON-format. In turn, ontology supports a single vocabulary of all concepts, expanded by logical connections and axioms. An ontology encoded as an OWL file becomes a control superstructure capable of semantic integration of heterogeneous data.

The arsenal of tools proposed in the project (JSON, Apache Spark, SPARQL) meets modern standards for storing and processing heterogeneous data and is capable of supporting interaction with many types of storage. Thereby, long-term data storage will be provided with the possibility of painless migration to subsequent versions of the software.

References

1. Zorich, D.M.: Data management: Managing electronic information: Data curation in museums. Museum Management and Curatorship **14** (4), 431 (1995).
2. Beagrie, N.: Digital curation for science, digital libraries, and individuals. The International Journal of Digital Curation **1** (1), 3–16 (2006).
3. Abbott, D.: "What is Digital Curation?". DCC Briefing Papers: Introduction to Curation. Edinburgh: Digital Curation Centre. Handle: 1842/3362 (2008). Available online: http://www.dcc.ac.uk/resources/briefing-papers/introduction-curation
4. Johnson, L.R., et al.: How important are data curation activities to researchers? Gaps and opportunities for academic libraries. Journal of Librarianship and Scholarly Communication, 6(General Issue), eP2198 (2018). Available online: https://doi.org/10.7710/2162-3309.2198
5. Kosinov, A.V., Erkimbaev, A.O., Zitserman, V.Yu., Kobzev G.A.: Ontology-based methods of thermophysical data integration. In: XV Russian Conference (with international participation) on Thermophysical Properties of Substances (RCTP-15), 103–104. Book of Abstracts. Moscow, Russia, (2018).
6. Pennock, M.: Curating e-Science Data. DCC Briefing Papers: Introduction to Curation. Edinburgh: Digital Curation Centre. Handle: 1842/3330 (2006). Available online: http://www.dcc.ac.uk/resources/briefing-papers/introduction-curation
7. Yerkimbaev, A.O., Zitserman, V.Yu., and Kobzev, G.A.: The role of metadata in the creation and application of information resources on the properties of substances and materials. Sci. Tech. Information Process **35** (6), 47–255 (2008).
8. Davenhall, C.: “Scientific Metadata”, DCC Digital Curation Manual, J. Davidson, S. Ross, M. Day (eds), (2011). Available online: http://www.dcc.ac.uk/resources/curation-reference-manual/scientific-metadata
9. Willis, C., Greenberg, J., and White, H.: Analysis and synthesis of metadata goals for scientific data. J. American Soc. for Information Science and Technology **63** (8), 1505–1520 (2012).
10. Erkimbaev, A.O., Zitserman, V.Yu., Kobzev, G.A., and Fokin L.R.: The logical structure of physicochemical data: problems of numerical data standardization and exchange. Russian Journal of Physical Chemistry A. **82** (1), 15–25 (2008).
11. Erkimbaev, A.O., Zitserman, V.Yu., Kobzev, G.A., and Trakhtenhers, M.S.: A universal metadata system for the characterization of nanomaterials. Sci. Tech. Inf. Process **42** (4), 211–222 (2015).
12. Erkimbaev, A.O., Zitserman, V.Yu., and Kobzev, G.A.: The intensive use of digital data in modern natural science. Automatic Documentation and Mathematical Linguistics **51** (5), 201–213 (2017).
13. Water Structure and Science. P7. Supercooled water has two phases and a second critical point. Available online: http://www1.lsbu.ac.uk/water/phase_anomalies.html
14. Kane, C. and Moore, J.: Topological insulators. Physics World 32–36 (2011).
15. Eletskii, A.V., Erkimbaev, A.O., Kobzev, G.A., Trachtengerts, M.S., and Zitserman V.Y.: Properties of nanostructures: data acquisition, categorization, and evaluation. Data Science Journal 11, 126–139 (2012). Available online: https://www.jstage.jst.go.jp/browse/dsj
16. Rahm, E. and Do, H.H.: Data cleaning: problems and current approaches. IEEE Data Eng. Bull. **23** (4), 3–13 (2000).
17. CDS/ISIS for Windows: Reference Manual (Version 1.31). Paris: UNESCO (1998).

Data Analysis Environment for Materials Science and Engineering Integrating Heterogeneous Data Resources

Toshihiro Ashino [1], Nobutaka Nishikawa [2], and Takuya Kadohira[3]

[1] Toyo University, 5-28-20 Hakusan, Bunkyo-ku, Tokyo 112-8606, Japan
ashino@acm.org
[2] Mizuho Information & Research Institute, Inc. 2-3 Kanda-Nishikicho, Chiyoda-ku, Tokyo 101-8443, Japan
[3] National Institute for Materials Science, 1-2-1 Sengen, Tsukuba, Ibaraki, 305-0047, Japan

Abstract. Materials performance analysis requires to integrate many heterogeneous data and information resources, experimental data, empirical/theoretical models and computational simulations. It means data analysis platform for materials science and engineering should provide many functionalities, e.g., data retrieval, processing, statistical analysis, symbolic mathematics, visualization and scripting capabilities to store the typical data analysis process and also, these heterogeneous data resources should be accessed unified way. Scripting language Python provides many of these capabilities with additional software modules and widely applied to interactive/non-interactive data processing environment. In this paper, a prototype design and implementation of data analysis environment for materials science and engineering is presented.

Keywords: virtual research environment, materials integration, materials ontology, semantic web, heterogeneous data integration

1 Introduction

In many research area, data intensive research, so called the Fourth Paradigm [1], have been increasing its importance. In materials science and engineering, there is a long tradition developing computerized materials property databases [2, 3]. But materials experiment requires huge cost and high skill, materials represent wide variation of properties, there are various measurement methods and substances, data intensive approach is delayed to be introduced into materials design process.

But advancement of computer simulation technology and new measurement method presented a possibility to obtain huge amount of data in this field. It enables to evaluate materials properties such as physical properties and long term performance with minimum experiment, relatively low cost and short period, furthermore, enables to predict materials performance without real experiment [4–6].

One of the important application area is to develop software platform for high throughput computational approach for materials design focused on functional materials which performances are directly reflect micro-scale physical properties [7, 8]. However, in case of structural materials performances prediction, e.g. creep rupture property, different scales and complexed interactions of physical phenomenon affect the total performance, it requires to integrate heterogeneous data and models.

This approach is called ICME (Integrated Computational Materials Engineering) [9]. In Japan, SIP-MI (Strategic Innovation Promotion Program: Materials Integration) is a project to implement ICME concept. Information platform for MI is required to handle and integrate many kind of information resources, such as experimental data, simulation modules and mathematical equations. Semantic description of data, relationships among data and attributes of data are essential in order to integrate these heterogeneous information.

We applied the Semantic Web framework to this application. It provides several machine readable semantic description standards, XML Schema [10], RDF (Resource Description Framework)/SPARQL (SPARQL Protocol and RDF Query Language) [11, 12], OWL (Web Ontology Language) [13] and OpenMath [14]. MI prototype data platform which can handle these data formats and enables to describe workflows of materials data processing has been developed.

2 Design and Implementation of the Prototype

The prototype system is based on a mathematical system, SageMath [15], which is an open source project integrates many open source mathematical systems, SciPy, R and others. It is based on Python programming environment and this means, various software modules developed for Python can be used in this system and it is easy to develop original data processing modules for this data processing environment.

Fig. 1 shows the design of the prototype system. In order to achieve flexible data management, since it should manage continuously evolving materials measurement and new materials data, metadata, which describes the structure of database is stored in Apache/Jena Fuseki SPARQL endpoint as RDF files. RDF provides conceptual description on the data resources and it is retrieved by using SPARAL query language.

Metadata which describes experimental data and mathematical equations, target materials, equation names, target property, application conditions and link to data and equation body, are written in RDF for retrieval by SPARQL. Sample experimental databases is stored as XML (Extensible Markup Language) documents, they can be accessed by their URI’s listed in RDF files. Equation bodies are also stored as XML documents which written in OpenMath semantic representation of mathematics, which provides rich vocabularies contain many operators and mathematical functions [16].

Python modules XML, RDFlib, SPARQLWrapper and py-openmath are incorporated into SageMath symbolic-math environment and original OpenMath parser have been developed for this prototype. Metadata which describes experimental data and mathematical equations, target materials, equation names, target property, application conditions and link to data and equation body, are written in RDF for retrieval by

SPARQL. Materials Ontology written in OWL is managed by the same SPARQL endpoint.

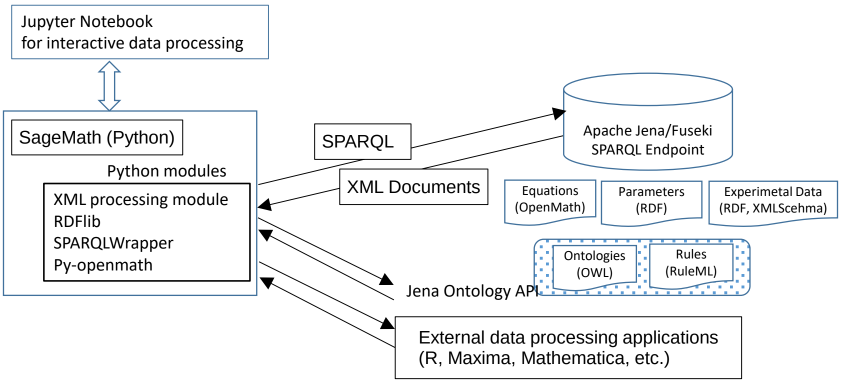
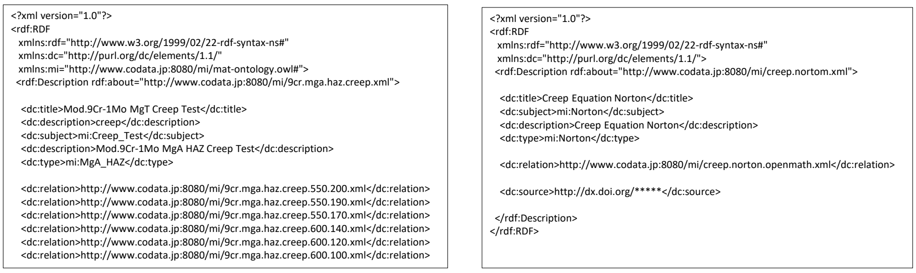


Fig. 1. Concept of the prototype system for materials data processing

Fig. 2 shows examples of metadata description for experimental data (a) and equation (b) in RDF. In current sample database, tags are selected from Dublin Core tag set defined in order to describe metadata [17], but there are many tag sets which defined to represent data meanings and any of them can be added into these RDF data anytime.

Experimental datasets and equation bodies are divided from RDF metadata file. RDF file contains URI's (Uniform Resource Identifier) indicate datasets and equations, since such files may have written in different data formats like XML Schema and OpenMath. Data retrieval requires two-steps, at first, find a RDF description by SPARQL and second, traverse the URI which is indicated by <dc:relation> tags.

Vocabularies used in database, property names, material names, units for measured values and other keywords are selected from extended Materials Ontology [18]. It intended to realize uniform data retrieval on heterogeneous data resources, in this case, experimental data and equation library stored in different RDF documents. In current prototype, words are selected manually from the ontology as a common vocabulary.



```
<?xml version="1.0"?>
<rdf:RDF
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:dc="http://purl.org/dc/elements/1.1/"
  xmlns:mi="http://www.codata.jp:8080/mi/mat-ontology.owl#">
 <rdf:Description rdf:about="http://www.codata.jp:8080/mi/9cr.mga.haz.creep.xml">

   <dc:title>Mod.9Cr-1Mo MgT Creep Test</dc:title>
   <dc:description>creep</dc:description>
   <dc:subject>mi:Creep_Test</dc:subject>
   <dc:description>Mod.9Cr-1Mo MgA HAZ Creep Test</dc:description>
   <dc:type>mi:MgA_HAZ</dc:type>

   <dc:relation>http://www.codata.jp:8080/mi/9cr.mga.haz.creep.550.200.xml</dc:relation>
   <dc:relation>http://www.codata.jp:8080/mi/9cr.mga.haz.creep.550.190.xml</dc:relation>
   <dc:relation>http://www.codata.jp:8080/mi/9cr.mga.haz.creep.550.170.xml</dc:relation>
   <dc:relation>http://www.codata.jp:8080/mi/9cr.mga.haz.creep.600.140.xml</dc:relation>
   <dc:relation>http://www.codata.jp:8080/mi/9cr.mga.haz.creep.600.120.xml</dc:relation>
   <dc:relation>http://www.codata.jp:8080/mi/9cr.mga.haz.creep.600.100.xml</dc:relation>
```

```
<?xml version="1.0"?>
<rdf:RDF
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:dc="http://purl.org/dc/elements/1.1/">
 <rdf:Description rdf:about="http://www.codata.jp:8080/mi/creep.nortom.xml">

   <dc:title>Creep Equation Norton</dc:title>
   <dc:subject>mi:Norton</dc:subject>
   <dc:description>Creep Equation Norton</dc:description>
   <dc:type>mi:Norton</dc:type>

   <dc:relation>http://www.codata.jp:8080/mi/creep.norton.openmath.xml</dc:relation>

   <dc:source>http://dx.doi.org/*****</dc:source>

 </rdf:Description>
</rdf:RDF>
```

Fig. 2. Metadata description in RDF for (a) experimental data and (b) constitution equation. Data and equations are stored in XML files pointed by URI's

3 An Example Materials Performance Analysis Workflow in Python

One of the typical materials data processing workflow, creep data analysis is displayed in Fig. 3. Workflows can be written in Python scripting language in the prototype, it provides quite flexible and extensible description. 1st, relevant creep experimental data is retrieved from database with SPARQL. Results are obtained in XML documents and they are transformed into appropriate format for further processing by the XPath functions of Python XML processing module. XML data format stored in database is defined in this project locally, but it should be standardized for test method or property in XML Schema.

2nd, appropriate equation, in this case Norton equation, constitution equation for creep behavior is selected by its metadata written in RDF. The metadata contains a URI which points semantic representation of the equation in OpenMath. It can be parsed and converted into the corresponding input format required by specified data processing package, e.g. R, SciPy and other packages which is integrated to SageMath.

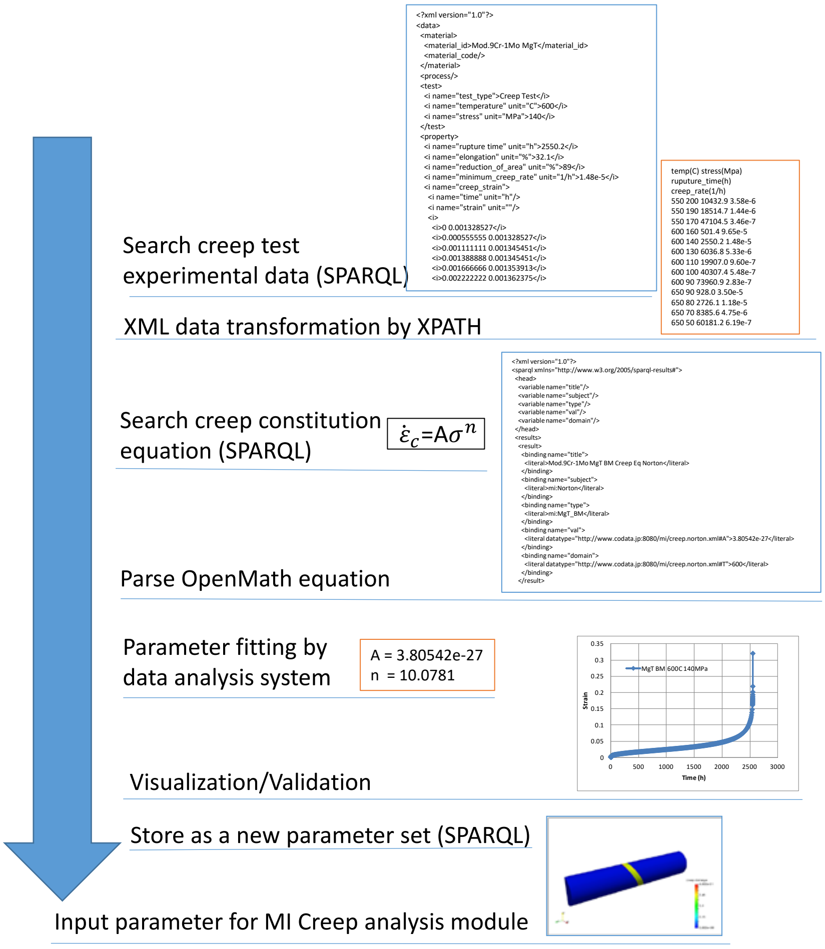
In the package, non-linear least square method is applied to the equation with the retrieved experimental data set. Obtained parameter values, in this case A and n, are written into RDF format, added appropriate metadata, e.g. link to corresponding experimental data, equation and version of software package, and stored into the database for further utilization in MI software modules.

This workflow can be stored as Python script and also, all functions can be used in interactive programming environment Jupyter notebook. This script has properly worked and proved the extensibility and flexibility of this system.

4 Discussions

There are many trials to develop ontology and integrate data with ontology [19–22]. Ontology can be used a fundamental dictionary for data integration. But in order to integrate heterogeneous information resources, all description of these resources should be based on common ontology or be mapped to the correspondence of ontology. This work is done manually, it requires continuous efforts to standardize and disseminate ontology, and also support system to select vocabulary with ontology reasoner.

Materials ontology has been extended to contain some concepts which relate to creep performance evaluation. In this prototype, ontology written in OWL can be accessed via Apache/Jena API, we are now testing utilization of reasoner in data retrieval and rule based data analysis with this functionality.



```
<?xml version="1.0"?>
<data>
 <material>
  <material_id>Mod.9Cr-1Mo MgT</material_id>
  <material_code/>
 </material>
 <process/>
 <test>
  <i name="test_type">Creep Test</i>
  <i name="temperature" unit="C">600</i>
  <i name="stress" unit="MPa">140</i>
 </test>
 <property>
  <i name="rupture time" unit="h">2550.2</i>
  <i name="elongation" unit="%">32.1</i>
  <i name="reduction_of_area" unit="%">89</i>
  <i name="minimum_creep_rate" unit="1/h">1.48e-5</i>
  <i name="creep_strain">
   <i name="time" unit="h"/>
   <i name="strain" unit=""/>
   <i>
    <i>0 0.001328527</i>
    <i>0.000555555 0.001328527</i>
    <i>0.001111111 0.001345451</i>
    <i>0.001388888 0.001345451</i>
    <i>0.001666666 0.001353913</i>
    <i>0.002222222 0.001362375</i>
```

```
temp(C) stress(Mpa)
ruputure_time(h)
creep_rate(1/h)
550 200 10432.9 3.58e-6
550 190 18514.7 1.44e-6
550 170 47104.5 3.46e-7
600 160 501.4 9.65e-5
600 140 2550.2 1.48e-5
600 130 6036.8 5.33e-6
600 110 19907.0 9.60e-7
600 100 40307.4 5.48e-7
600 90 73960.9 2.83e-7
650 90 928.0 3.50e-5
650 80 2726.1 1.18e-5
650 70 8385.6 4.75e-6
650 50 60181.2 6.19e-7
```

```
<?xml version="1.0"?>
<sparql xmlns="http://www.w3.org/2005/sparql-results#">
 <head>
  <variable name="title"/>
  <variable name="subject"/>
  <variable name="type"/>
  <variable name="val"/>
  <variable name="domain"/>
 </head>
 <results>
  <result>
   <binding name="title">
    <literal>Mod.9Cr-1Mo MgT BM Creep Eq Norton</literal>
   </binding>
   <binding name="subject">
    <literal>mi:Norton</literal>
   </binding>
   <binding name="type">
    <literal>mi:MgT_BM</literal>
   </binding>
   <binding name="val">
    <literal datatype="http://www.codata.jp:8080/mi/creep.norton.xml#A">3.80542e-27</literal>
   </binding>
   <binding name="domain">
    <literal datatype="http://www.codata.jp:8080/mi/creep.norton.xml#T">600</literal>
   </binding>
  </result>
```

Fig. 3. A creep data processing workflow and corresponding operation on the system

5 Conclusion

Prototype of data analysis environment which has capability integrating heterogeneous materials information resources have been developed based on Python programming language and the design have been verified by sample database and script. RDF metadata representation for materials experimental data and mathematical equations is defined and tested for further development of MI system.

Acknowledgments

This work was supported by Council for Science, Technology and Innovation (CSTI), Cross-ministerial Strategic Innovation Promotion Program (SIP), Structural Materials for Innovation" (Funding by JST).

References

1. Hey, T., Tansley, S., and Tolle, K.M.: The fourth paradigm: data-intensive scientific discovery (Microsoft Research, Redmond, 1969).
2. Rumble Jr., J.R.: Integr. Mater. Manuf. Innov. (6), 172–186 (2017).
3. Austin, T.: Mater. Discov. **3**, 1–12 (2016).
4. Curtarolo, S., Hart, G.L.W., Nardelli, M.B., Mingo, N., Sanvito, S., and Levy, O.: Nature Mater. **20**, 191–201 (2013).
5. Broderick, S.R., Santhanam, G.R., and Rajan, K.: JOM **68**, 2109–2115 (2016).
6. Editorial: Scripta Mater. **70**, 1–2 (2014).
7. Ong, S.P., Richards, W.D., Jain, A., Hautier, G., Kocher, M., Cholia, S., Gunter, D., Chevrier, V.L., Persson, K.A., and Ceder, G.: Comp. Mater. Sci. **68**, 314–319 (2013).
8. Kalidindi, S.R., Niezgoda, S., Landi, G., and Fast, T.: Comp., Mater. and Cont. **17,** 103–125 (2010).
9. National Research Council: Integrated Computational Materials Engineering: A Transformational Discipline for Improved Competitiveness and National Security (The National Academies Press, Washington, DC. 2008).
10. W3C: https://www.w3.org/standards/xml/schema, last accessed 2019/5/5
11. W3C: https://www.w3.org/TR/2014/NOTE-rdf11-primer-20140624/, last accessed 2019/5/5
12. W3C: https://www.w3.org/TR/2013/REC-sparql11-overview-20130321/, last accessed 2019/5/5
13. W3C: https://www.w3.org/TR/2012/REC-owl2-overview-20121211/, last accessed 2019/5/5
14. OpenMath Society: https://www.openmath.org/standard/om20-2017-07-22/, last accessed 2019/5/5
15. SageMath, the Sage Mathematics Software System (Version 8.0), The Sage Developers, 2017, https://www.sagemath.org. last accessed 2019/5/5
16. Ashino, T. and Yamashita, Y.: Data Sci. J. **11**, ASMD17-ASMD21 (2012).
17. Dublin Core Initiative: http://dublincore.org/, last accessed 2019/5/5
18. Ashino, T.: Data Sci. J. **9**, 54–61 (2010).
19. Zhao, S. and Qian, Q.: AIP Advances **7**, 105325 (2017).
20. LeBlanc, E., Balduccini, M., and Regli, W.C.: AAAI-14 Workshop (AAAI, Quebec, 2014) 39–42.
21. Madalli, D., Sulochana, A., and Singh, A.K.: Data Technol and Appl. **50**, 103–117 (2016).
22. Remolona, M.F.M., Conway, M.F., Balasubramanian, S., Fan, L., Feng, Z., Gu, T., Kim, H., Nirantar, P.M., Panda, S., Ranabothu, N.R., Rastogi, N., and Venkatasubramanian, V.: Comp. and Chem. Eng. **107**, 49–60 (2017).

Graph Representation of Materials Research on Diamond Light Source

Vasily Bunakov[1]

[1] Science and Technology Facilities Council, United Kingdom

Abstract. The demonstration presents published outcomes of materials research performed on Diamond Light Source over more than ten years of its operation since 2008, in the form of a labelled graph. Examples of queries and visualizations are given that can support knowledge discovery and research impact studies.

Keywords: synchrotron radiation source, materials research, Open Science, graph databases, query languages, visualizations

Diamond Light Source [1] is sometimes called the most expensive research instrument in the United Kingdom. Visitor scientists supported by the Diamond staff scientists and technicians conduct multiple investigations that result in data collections and numerous research publications. Amount and quality of research performed on Diamond are substantial; handy means of this knowledge discovery and representation are important to ensure that this costly knowledge is reused to its full potential. Proper representation of research outcomes can also support measuring impact of research on such large-scale facilities and prepare compelling cases for next rounds of investment in their infrastructure and staff.

The demonstration presents published Diamond outcomes related to materials research as a graph in the Neo4j graph database [2] that integrates a few data sources including the Diamond own bibliographic database and popular Open Science data services such as Unpaywall [3] and CORE [4]. The resulted graph allows performing queries and visualizations in support of both aforementioned generic use cases: knowledge discovery and impact studies. In addition, it can be used to measure trends in Open Access to research information in the materials domain.

This work is one of the outcomes of the FREYA project [5] that develops requirements and services for the research information infrastructure based on extensive use of persistent identifiers. The FREYA project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No. 777523.

References

1. Diamond Light Source. https://www.diamond.ac.uk/ Last accessed 2019/08/19
2. Neo4j graph database. https://neo4j.com/ Last accessed 2019/08/19
3. Unpaywall service. https://unpaywall.org/ Last accessed 2019/08/19
4. CORE aggregator of Open Access publications. https://core.ac.uk/ Last accessed 2019/08/19
5. FREYA project. https://www.project-freya.eu/ Last accessed 2019/08/19

MATERIALS DATABASES, MATERIALS DATA INFRASTRUCTURES AND DATA SERVICES

Thermodynamic Database for Pure Substances IVTANTHERMO-Online

Igor Morozov [0000-0002-0122-9400]

Joint Institute for High Temperatures of RAS, Izhorskaya 13, bld 2, Moscow 125412, Russia
morozov@ihed.ras.ru

Abstract. Development of the IVTANTHERMO-Online information system is discussed including the corresponding thermodynamic database for pure substances, bibliographic database, user interface and supplementary software. In contrast to numerous thermodynamic databases for engineers, the IVTANTHERMO-Online is meant as a database for researchers, so that scientists from different institutions can contribute to the database extension and regular updates.

Keywords: thermodynamic properties, database, pure substances

1 Introduction

Thermodynamic databases play an essential role in a wide range of applications such as rocket engine engineering, nuclear power, chemical technology, metallurgy, resource usage, waste recycling, etc. The IVTANTHERMO information system [1] based on the reference book [2] has made a significant contribution to the accumulation of thermodynamic data. It has been developed since 1966 in the Institute of High Temperatures of the Academy of Sciences of the USSR. Initially it was focused at thermodynamics of the rocket fuel combustion products but later the area of its applicability was extended substantially. Nowadays the development is continued in the Laboratory for Thermophysical Databases of JIHT RAS.

The thermodynamic databases are being developed in many research centers [3–7]. Some of them are united into international organizations such as SGTE [8]. Typically, the information from these databases is not open and provided for a fee. The most well-known open web resource is the NIST WebBook [3, 9] which provides access to a part of information about thermodynamical properties, thermochemical data, energy spectra of ions, vibrational and rotational spectra of molecules, etc.

In the course of the IVTANTHERMO system development a set of requirements was formulated that any fundamental thermodynamic database or a reference book should fulfill [1]. Fitting to these requirements is used to distinguish between the “critical reference books” or “expert level databases” and the databases compiled from different sources and providing information “as is” without references, clear identification of the calculation procedures and evaluation of its reliability and accuracy. The goal of further IVTANTHERMO development is to keep adherence to the requirements [1] using at the same time contemporary technologies to make the system more functional, user friendly and accessible by different kinds of researchers.

2 IVTANTHERMO Database Development

2.1 IVTANTHERMO-Online Project

The first version of the IVTANTHERMO ran on the HP3000 hardware system whereas the subsequent versions were implemented for PC as “INVATHERMO for DOS” and “INVATHERMO for Windows” software packages. The later was distributed since the end of 1990s [10]. Nowadays it includes information about more than 3400 substances, formed of 96 chemical elements, as well as supplementary software for analysis of experimental results, data fitting, calculation and estimation of thermodynamical functions and thermochemistry quantities.

Recently a next generation of the IVTANTHERMO database and related software is proposed called “IVTANTHERMO-Online” [11] (Fig. 1). It has the following main features:

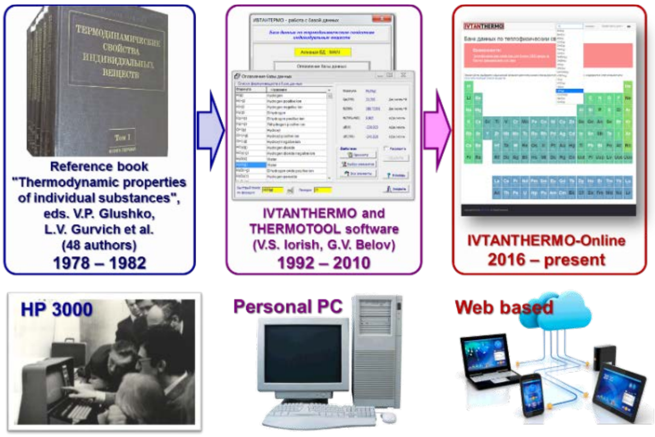


Fig. 1. Development of the IVTANTHERMO information system

- Modern user-friendly web interface with client-server architecture for remote access of users and contributors.
- Regular data updates with a full history of all changes.
- Two-level scheme of new data submission and assessment by an expert committee.
- Storing original tables of data (experimental or simulation results) with the information about data sources, methods of processing and fitting, fundamental constants.
- Detailed bibliography information for all data sources.
- Integration with the existing codes for calculation of the thermodynamical functions.
- Substance search by names, chemical elements, chemical formula, CAS and InChi numbers and substance categories.

- Extensible database design.
- Online calculation services (chemical composition, etc).

2.2 Revision Control System

One of the key features that can make the IVTANTHERMO-Online a collaboration platform is the ability to store multiple revisions of the thermodynamic data in the similar way that it is done in the contemporary version control systems such as Git, Subversion, etc. The data are slit into blocks, each block is an elementary unit for tracking revisions. The block can include, for instance, the thermodynamic properties of a particular substance in a given phase or basic properties of a molecule.

A user that can contribute to the database is called as an “expert”. The experts have additional rights to add or edit the data for a group of substances assigned by the site administrator. After adding or editing, a new block of data is created and marked as unchecked or preliminary (see Fig. 2). Then the expert can continue to edit the data creating, if needed, additional versions of the data block. Finally, the new data are submitted to the expert committee that checks the data and possibly accepts it as a new recommended version that becomes visible at the web site for all users.

As seen from Fig. 3 all the versions that have been recommended earlier are stored in the database and available for users. It allows users to find what has been changed, when the changes have been accepted, who was an expert and what was the reason for the changes (appearance of new experimental data, better approximation, etc).

Moreover, the database is designed to store additional metadata for each data block which may include comments, original experimental data, bibliography, full papers and preprints. Storage of these metadata increases usability and reproducibility of the main data and allows new experts to get full information of previous studies when the data is to be updated.

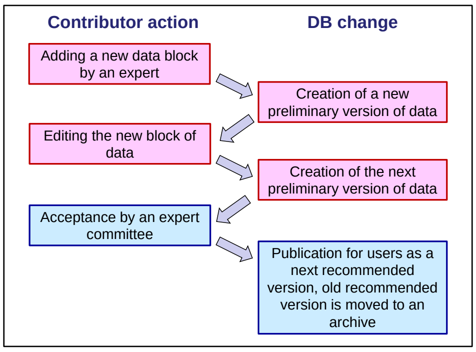

Contributor action
DB change
Adding a new data block by an expert
Creation of a new preliminary version of data
Editing the new block of data
Creation of the next preliminary version of data
Acceptance by an expert committee
Publication for users as a next recommended version, old recommended version is moved to an archive

Fig. 2. Life cycle of a data block in the IVTANTHERMO-Online system

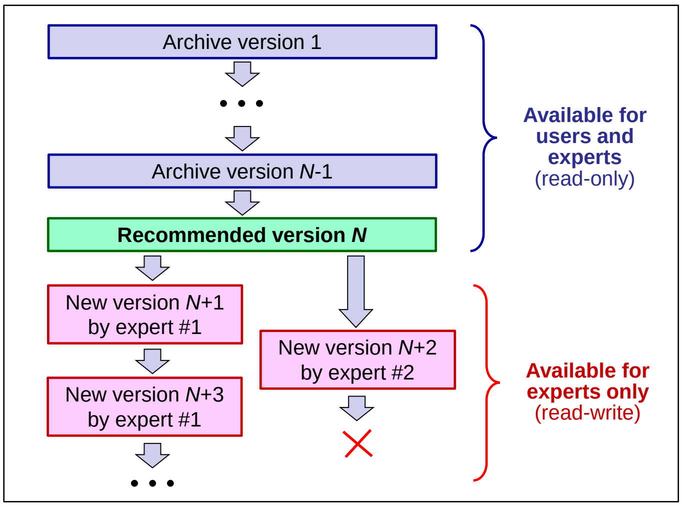

Archive version 1
• • •
Archive version N-1
Recommended version N
New version N+1 by expert #1
New version N+3 by expert #1
• • •
New version N+2 by expert #2
Available for users and experts (read-only)
Available for experts only (read-write)

Fig. 3. Storing of multiple revisions of the same data block in the IVTANTHERMO-Online database

2.3 Supplementary Software for Experts

Traditionally the thermodynamic data for substances in the condensed and gaseous states are treated separately. The partition functions of a molecular gas can be obtained from either molecular constants or direct solution of the Schrödinger equation based on the interatomic potential calculated by quantum chemistry or other approaches [12, 13]. On the contrary, ab initio calculations of the thermodynamic properties of solid and liquid substances which provide reasonable accuracy are rather challenging and there are few of them in the literature. Therefore, assessment of the thermodynamic properties of a condensed matter is based typically on the statistical analysis of direct measurements of enthalpy increments and/or heat capacity including estimation of initial data error, robust statistics, examining phase transition areas and approximation of the measured values with appropriate analytical dependencies [14, 15].

Figs. 4 and 5 present the corresponding supplementary software packages that is being developed in JIHT RAS. These packages are to be used by experts for assessment of experimental or ab initio data and updating the database. They include comparison of different experimental measurements, weighted approximation, estimation of errors, etc. It is planed that the IVTANTHERMO-Online system will include the web version of these codes which will allow experts to perform the whole research cycle online storing the intermediate results and all the metadata in the same database and sharing them with each other.

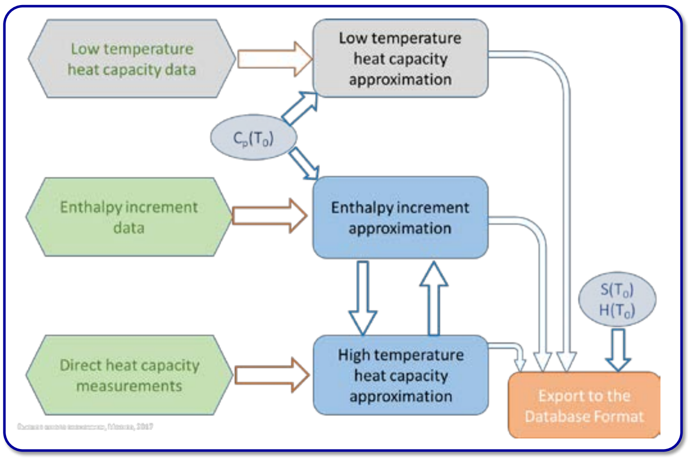

Low temperature heat capacity data
Low temperature heat capacity approximation
$C_p(T_0)$
Enthalpy increment data
Enthalpy increment approximation
$S(T_0)$
$H(T_0)$
Direct heat capacity measurements
High temperature heat capacity approximation
Export to the Database Format

Fig. 4. CondensedThermoFit package [15] for assessment of the thermodynamic properties of substances in the condensed state

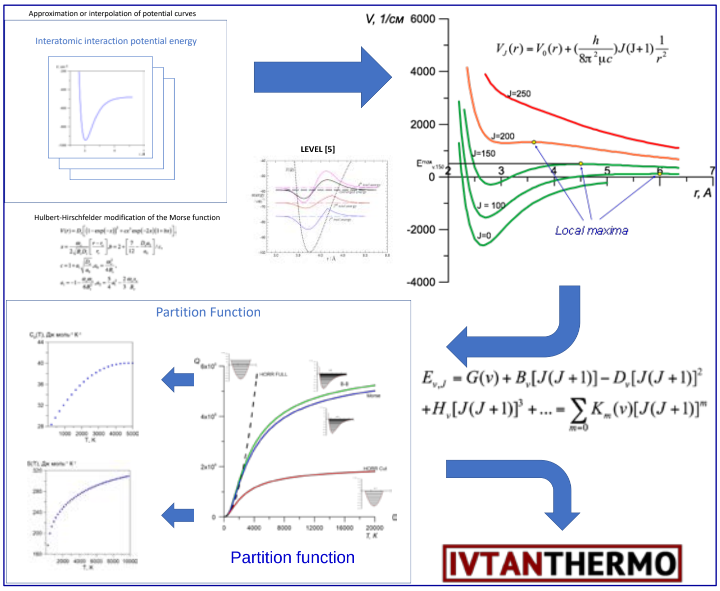

Approximation or interpolation of potential curves
Interatomic interaction potential energy
Hulbert-Hirschfelder modification of the Morse function
LEVEL [5]
V, 1/см
6000
4000
2000
0
-2000
-4000
$V_J(r) = V_0(r) + (\frac{h}{8\pi^2 \mu c})J(J+1)\frac{1}{r^2}$
J=250
J=200
J=150
J = 100
J=0
Emax v.150
2
3
4
5
6
7
r, A
Local maxima
Partition Function
Partition function
HORR FULL
Morse
HORR Cut
$E_{v,J} = G(v) + B_v[J(J+1)] - D_v[J(J+1)]^2$
$+H_v[J(J+1)]^3 + ... = \sum_{m=0} K_m(v)[J(J+1)]^m$
IVTANTHERMO

Fig. 5. Calculation of the thermodynamic properties of diatomic molecules in the gaseous state based on the interatomic potential

3 Conclusions

The IVTANTHERMO-Online thermodynamic database is discussed that continues the series of IVTANTHERMO information systems. It meets both the requirements to critical reference books and the standards of modern web-based systems. Opposed to other thermodynamic databases it keeps track of the old versions of data, contains full information on data sources and methods of data processing. Supplementary thermodynamic data processing software should allow experts to add or update information in the database without using external codes.

References

1. Gurvich, L.V.: Reference books and data banks on the thermodynamic properties of individual substances. Pure Appl. Chem. **61**, 1027–1031 (1989).
2. Gurvich, L.V., Bergman, G.A., Veyts, I.V., et al.: Thermodynamic Properties of Individual Substances. Glushko V.P. (ed). New York, Hemisphere Publishing (1989).
3. Chemistry WebBook Homepage (NIST), http://webbook.nist.gov/chemistry/, last accessed 2019/08/26
4. Facility for the Analysis of Chemical Thermodynamics (FACT) Homepage, http://www.crct.polymtl.ca/fact/, last accessed 2019/08/26
5. Thermodata, France Homepage, http://thermodata.online.fr/, last accessed 2019/08/26
6. Thermo-Calc Software Homepage, http://www.thermocalc.com/, last accessed 2019/08/26.
7. Materials-oriented Little Thermodynamic Database (MALT) Homepage, http:// kagaku.com/malt/, last accessed 2019/08/26
8. Scientific Group Thermodata Europe (SGTE) Homepage, http://www.crct.polymtl.ca/sgte/, last accessed 2019/08/26
9. Linstrom, P.J. and Mallard, W.G.: The NIST Chemistry WebBook: A chemical data resource on the internet. J. Chem. Eng. Data 46, 1059–1063 (2001).
10. Belov, G.V., Iorish, V.S., and Yungman, V.S.: IVTANTHERMO for Windows – database on thermodynamic properties and related software. CALPHAD **23** (2), 173–180 (1999).
11. Belov, G.V., Dyachkov, S.A., Levashov, P.R., Lomonosov, I.V., Minakov, D.V., Morozov, I.V., Sineva, M.A., and Smirnov, V.N.: The IVTANTHERMO-Online database for thermodynamic properties of individual substances with web interface. J. Phys. Conf. Ser. **946**, 012120 (2018).
12. Maltsev, M.A., Kulikov, A.N., and Morozov, I.V.: Thermodynamic properties of vanadium and cobalt argide ions, VAr^+ and $CoAr^+$. J. Phys. Conf. Ser. **774,** 012023 (2016).
13. Maltsev, M.A., Morozov, I.V., and Osina, E.L.: Thermodynamic Properties of Ar_2^+ and Ar_2 Argon Dimers. High Temp. **57** (1), 37–40 (2019).
14. Aristova, N.M., Belov, G.V., Morozov, I.V., and Sineva, M.A.: Thermodynamic properties of condensed uranium dioxide. High Temp. **56** (5), 652–661 (2018).
15. Sineva, M.A., Morozov, I.V., Belov, G.V., Aristova, N.M., and Lavrinenko, Ya.: Simultaneous analysis of the enthalpy increment and heat capacity data measurements for updating the IVTANTHERMO database. J. Phys. Conf. Ser. (2019), in press.

The AiiDA Ecosystemfor Computational Materials Science

Leopold Talirz[1,2,3], Aliaksandr V. Yakutovich[1,2,3], Sebastiaan P. Huber[1,2],
Martin Uhrin[2], Spyros Zoupanos[1,2], Leonid Kahle[1,2], Conrad Johnston[1,2],
Nicolas Mounet[2], Rico Häuselmann[2], Dominik Gresch[6], Tiziano Müller[1,7],
Andrea Cepellotti[2], Fernando Gargiulo[2], Snehal Kumbhar[1,2], Elsa Passaro[1,2],
Marco Borelli[2], Andrius Merkys[2], Ole Schütt[4], Berend Smit[1,3], Daniele Passerone[1,4],
Carlo A. Pignedoli[1,4], Boris Kozinsky[8], Joost VandeVondele[1,5,,6],
Thomas Schulthess[1,5,6], Nicola Marzari[1,2], Giovanni Pizzi[1,2]

[1]National Centre for Computational Design and Discovery of Novel Materials (MARVEL),
École Polytechnique Fédérale de Lausanne, CH-1015 Lausanne, Switzerland
[2]Theory and Simulation of Materials (THEOS), Faculté des Sciences et Techniques
de l'Ingénieur, École Polytechnique Fédérale de Lausanne, CH-1015 Lausanne, Switzerland
[3]Laboratory of Molecular Simulation (LSMO), Institut des Sciences et Ingénierie Chimiques,
École Polytechnique Fédérale de Lausanne, CH-1951 Sion, Switzerland
[4]nanotech@surfaces laboratory, Swiss Federal Laboratories for Materials
Science and Technology (Empa), CH-8600 Dübendorf, Switzerland
[5]Swiss National Supercomputing Centre, CH-6900 Lugano, Switzerland
[6]ETH Zürich, Switzerland
[7]University of Zurich, Switzerland
[8]Research and Technology Center, Robert Bosch LLC, Cambridge, MA 02139, USA

leopold.talirz@epfl.ch, aliaksandr.yakutovich@epfl.ch

Abstract. AiiDA (aiida.net) is a workflow manager for computational science with a strong focus on provenance, performance and extensibility. When executing a workflow, AiiDA records the provenance − calculations performed, codes used and data generated − in a directed acyclic graph tailored to provide full reproducibility of any given result. The AiiDA engine relies on a message queue in order to support high-throughput use cases of up to 50k calculations per hour, and the relational database backend enables performant queries on graphs of millions of nodes. AiiDA *plugins* can extend the core python framework in numerous ways, adding not only new workflows and connections to new simulation codes but also support for new types of job schedulers, transport protocols and extensions of the AiiDA command line interface. While domain experts can install AiiDA on their own hardware, the AiiDA lab web platform gives novice users access to their personal AiiDA environment in the cloud, where they can run and manage workflows through tailored and lightweight web applications in the browser. The ecosystem is completed by the Materials Cloud dissemination portal, where researchers can publish their AiiDA graphs, thus providing access not only to the results of calculations, but to every step along the way. Peers can browse the database interactively, download individual files or the whole database, and start their research right from where the original author left off.

Keywords: Computational Materials Science, Provenance, Workflows, SaaS

Today, many open questions in computational science call for more than individual computations using a single code. As the demand for integration and throughput increases, the design of robust and reproducible workflows is becoming ever more important. In this context, the move towards open science [1–3] raises the level of scrutiny and demands that workflows and data be recorded in a way that can be inspected and reused by scientific peers.

AiiDA. AiiDA is a python framework designed around the four pillars of computational science: Automation, Data, Environment and Sharing (ADES) [4]. In the default usage model, AiiDA is installed on the workstation of a researcher and connects to remote compute resources through the secure shell protocol (SSH). In order to support high-throughput use cases of 50k calculations per hour, the AiiDA daemon relies on the RabbitMQ message broker, while the PostgreSQL database backend enables performant queries on data sets of millions of nodes.

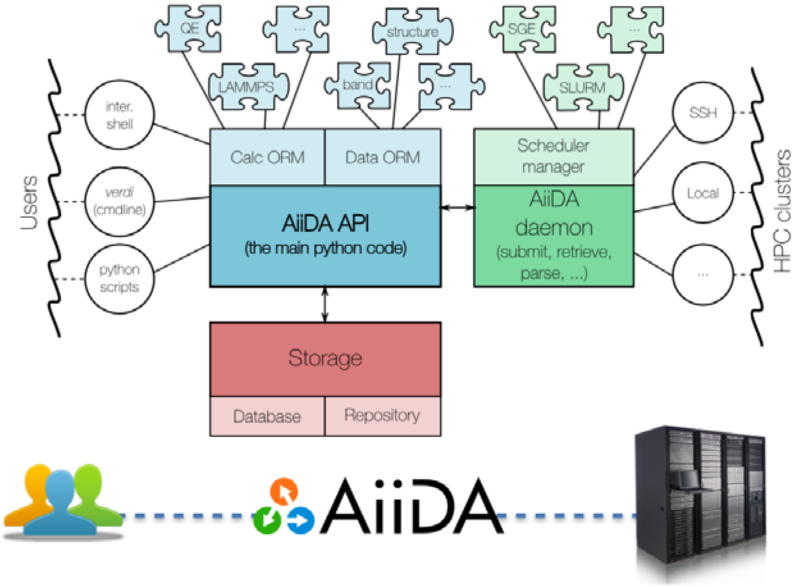


Fig. 1. AiiDA infrastructure. Users interact with AiiDA through the `verdi` command line, through interactive shell or via python scripts. AiiDA records the provenance of calculations in a database and file repository, while the AiiDA daemon automates workflows and interacts with remote compute resources. Figure reproduced with permission [4]

The focus on provenance and extensibility is a design choice that differentiates AiiDA from other workflow managers in the field of computational materials science, such as Aflow [5], atomate [6], MAST [7] or OQMD [8]. AiiDA plugins leverage python entry points to extend both the AiiDA command line interface and the python API – for example, AiiDA plugins can provide new workflows, connect to new simulation codes, provide support for new types of schedulers, transport protocols and seamlessly extend the existing command line interface. A template helps getting started with plugin development [9], and a plugin registry [10] provides a central point for registering the plugin. For example, a 2019 survey on the AiiDA mailing list points to more than 30 AiiDA-powered research projects using >25 different AiiDA plugins.

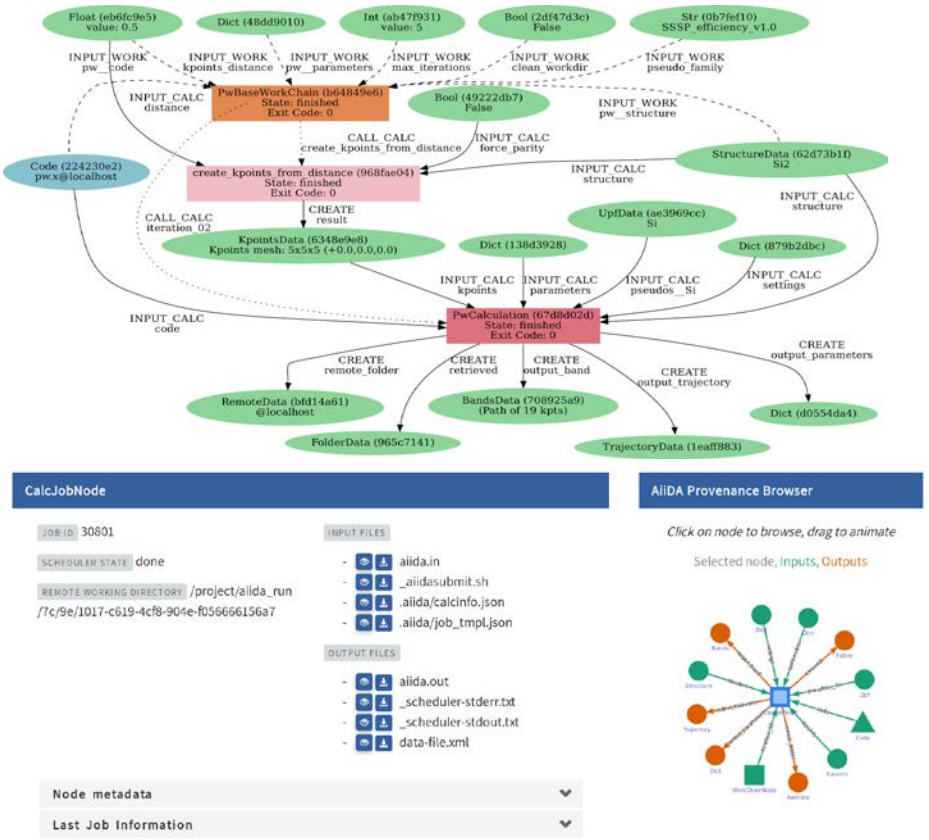
Instead of defining a new workflow markup language based on XML derivatives (Karajan [11], Askalon [12]) or JSON/yaml (Fireworks [13], Common Workflow Language [14]), AiiDA aims to make it easy for users to write workflows directly in python, providing full access to the AiiDA API, including queries of the entire provenance of previous calculations.

Since the release of the first paper [4], the provenance model of AiiDA has been extended to include a representation of workflows. While the basic building blocks of data and calculation nodes are sufficient for recording “data provenance” in a directed acyclic graph, workflow nodes provide logical abstraction by bundling several calculations (Fig. 2).

Further updates include switching the workflow engine from a polling mechanism to a message queue, which reduces overhead for quick calculations by orders of magnitude and makes it possible to run 50k calculations per hour. Reusing one workflow in another has become easier, and workflows now are auto documenting, telling users what inputs they expect and what outputs they produce without the need to read code. AiiDA now includes measures to deal with stability issues when connecting to remote clusters (network issues, cluster down). The command line interface has been overhauled, providing a uniform feel across all commands, dramatically increasing code reuse as well as test coverage. Writing AiiDA plugins requires significantly less boilerplate code, and AiiDA 1.0 is python3 compatible.

AiiDA was developed with the computational scientist in mind – a demographic familiar with UNIX operating systems, the terminal and python, interested in designing and tweaking complex workflows. The availability of robust materials science workflows, however, makes AiiDA interesting for a new user base: non-specialists, such as experimentalists or researchers at companies, who would like to run well-defined turnkey solutions using an intuitive graphical user interface.

AiiDA lab. The AiiDA lab leverages state-of-the-art technologies (JupyterHub, Jupyter widgets and kubernetes) to provide AiiDA-powered “apps” that run in the web browser. After logging in to the platform, an AiiDA lab user has access to a personal Docker container through a Jupyter-based graphical user interface. The container comes preinstalled with AiiDA as well as a selection of apps for common tasks, such as connecting AiiDA to a remote compute resource or performing a geometry optimization or band structure calculation using the Quantum ESPRESSO [15] and CP2K [16] density functional theory (DFT) codes.



JOB ID 30801

SCHEDULER STATE done

REMOTE WORKING DIRECTORY /project/aiida_run /7c/9e/1017-c619-4cf8-904e-f056666156a7

INPUT FILES
- aiida.in
- _aiidasubmit.sh
- .aiida/calcinfo.json
- .aiida/job_tmpl.json

OUTPUT FILES
- aiida.out
- _scheduler-stderr.txt
- _scheduler-stdout.txt
- data-file.xml

Fig. 2. AiiDA provenance graph. AiiDA tracks the provenance of data, calculations and workflows, allowing share it with other AiiDA users, to query and visualize it on the fly. (top) Auto-generated provenance graph of a calculation (red box) that takes four input data and produces five output data (green ellipses). In addition to the "data provenance layer" (solid lines), the graph includes the "logical provenance layer" with the workflow (orange box, dashed lines). While this workflow simply wraps the calculation, this might just represent the first of many workflow steps (not shown) (bottom) Same graph viewed through the interactive provenance browser driven by the AiiDA REST API. The selected calculation node is shown in the center, with arrows to/from connected nodes

AiiDA lab apps are nothing but Jupyter notebooks rendered in "app mode" [17]. Developers can therefore write powerful apps directly in python (no JavaScript required), minimizing the entry barrier for existing AiiDA users to writing such apps. A library of AiiDA-specific, reusable widgets further simplifies the task of creating apps, making e.g. the upload of a structure just one line of code. One context, in which this model has already been taken up, are mixed experimental/theoretical groups, where it frees computational scientists from repetitive tasks by letting the experimentalists run the corresponding workflows themselves.

The AiiDA ecosystem is completed by the Materials Cloud Archive, a moderated research data repository for computational materials science registered on re3data [18], FAIRsharing [19] and recommended by Nature Scientific Data [20]. Besides welcoming relevant data from computational materials science in general, the Materials Cloud Archive accepts AiiDA databases. By uploading an AiiDA database, researchers provide access to the full provenance of their calculations, enabling peers to browse the database interactively, download individual files or the whole database, and start their research right from where the original author left off.

AiiDA is free and open source (MIT license), and deployment scripts for the AiiDA lab are scheduled to be released under the same license later this year.

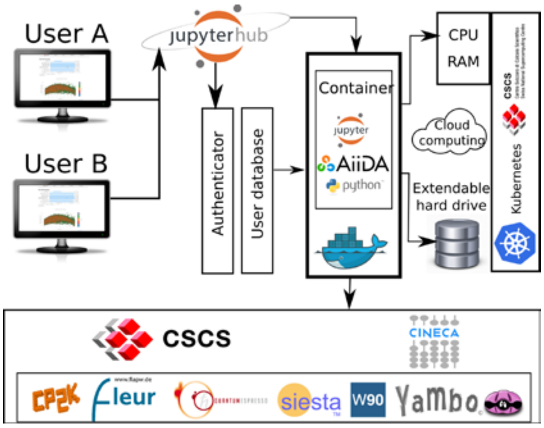
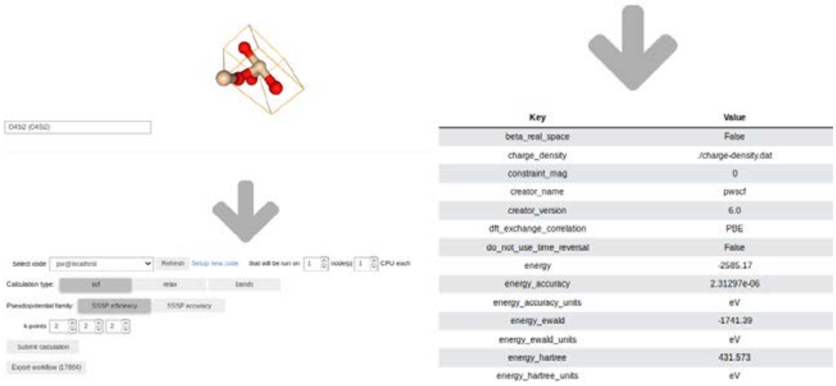


Fig. 3. AiiDA lab infrastructure. The AiiDA lab login page is provided by the JupyterHub that manages user authentication. Once user is logged in, JupyterHub will launch a Docker container and will expose access to Jupyter notebooks running inside the container. To balance the server load we deployed AiiDA lab on Kubernetes platform provided by CSCS. Every AiiDA lab container comes with AiiDA pre-installed and pre-configured

Demo. Jupyter notebooks will be used to demonstrate how to solve a range of common tasks using the AiiDA python & command line interfaces (not shown). The demo will also include an AiiDA lab application that allows to perform electronic structure calculations with Quantum ESPRESSO [15] and CP2K [16]. Fig. 4 provides a glimpse of the interface for preparing the inputs of a Quantum ESPRESSO calculation and displaying its results. The use of Jupyter notebooks enables a smooth transition from regular use to development, with Jupyter widgets providing interactive JavaScript components while programming in python.

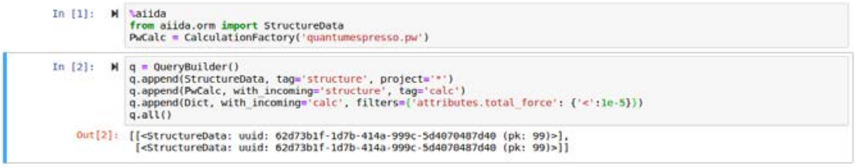
Since the calculation is managed by AiiDA all the data are stored in the AiiDA graph. To access it one can employ the QueryBuilder a tool that allows to query the AiiDA database. An example of a query is provided below.



O4Si2 (O4Si2)

Key	Value
beta_real_space	False
charge_density	./charge-density.dat
constraint_mag	0
creator_name	pwscf
creator_version	6.0
dft_exchange_correlation	PBE
do_not_use_time_reversal	False
energy	-2585.17
energy_accuracy	2.31297e-06
energy_accuracy_units	eV
energy_ewald	-1741.39
energy_ewald_units	eV
energy_hartree	431.573
energy_hartree_units	eV

Fig. 4. Electronic structure application. After providing minimal inputs via a Jupyter notebook in AppMode, a dedicated AiiDA WorkChain generates the full inputs required by the DFT code and submits the calculation. From then on AiiDA takes over, managing the preparation of the input files, sending them to the supercomputer, waiting until the calculation is finished, retrieving the results back and parsing the output



```
In [1]:   %aiida
          from aiida.orm import StructureData
          PwCalc = CalculationFactory('quantumespresso.pw')

In [2]:   q = QueryBuilder()
          q.append(StructureData, tag='structure', project='*')
          q.append(PwCalc, with_incoming='structure', tag='calc')
          q.append(Dict, with_incoming='calc', filters={'attributes.total_force': {'<':1e-5}})
          q.all()

Out[2]:  [[<StructureData: uuid: 62d73b1f-1d7b-414a-999c-5d4070487d40 (pk: 99)>],
          [<StructureData: uuid: 62d73b1f-1d7b-414a-999c-5d4070487d40 (pk: 99)>]]
```

Fig. 5. Querying via the AiiDA python API. This example query searches for atomic structures that were used as inputs to a Quantum ESPRESSO PwCalculation, filtering only those structures for which the total computed force has converged to less than 1e-5 eV/Å

Acknowledgements

The AiiDA ecosystem is supported by the MARVEL National Centre of Competence in Research, funded by the Swiss National Science Foundation, the European Centre of Excellence MaX (grant no. 824143), the INTERSECT project (grant no. 814487), the swissuniversities P-5 “Materials Cloud” project (ID: 182-008), and the OSSCAR project of the EPFL Open Science Fund.

The Swiss National Supercomputing Centre is acknowledged for virtual hardware, storage and high-performance computing resources.

References

1. Concordat on Open Research Data. https://www.ukri.org/files/legacy/documents/concordatonopenresearchdata-pdf/
2. Research Data Alliance (2014) The Data Harvest Report – sharing data for knowledge, jobs and growth. https://rd-alliance.org/data-harvest-report-sharing-data-knowledge-jobs-and-growth.html. Accessed 28 Oct 2018
3. Ministerie van Onderwijs C en W (2016) Amsterdam Call for Action on Open Science – Report.

https://www.government.nl/documents/reports/2016/04/04/amsterdam-call-for-action-on-open-science. Accessed 19 Aug. 2019

4. Pizzi, G., Cepellotti, A, Sabatini, R., et al.: AiiDA: automated interactive infrastructure and database for computational science. Comput. Mater Sci. **111**, 218–230 (2016). https://doi.org/10.1016/j.commatsci.2015.09.013
5. Curtarolo, S., Setyawan, W., Hart, GLW, et al.: AFLOW: An automatic framework for high-throughput materials discovery. Comput. Mater. Sci. **58,** 218–226 (2012). https://doi.org/10.1016/j.commatsci.2012.02.005
6. Mathew, K., Montoya, JH., Faghaninia, A., et al.: Atomate: A high-level interface to generate, execute, and analyze computational materials science workflows. Comput. Mater. Sci. **139**, 140–152 (2017). https://doi.org/10.1016/j.commatsci.2017.07.030
7. Mayeshiba, T, Wu, H., Angsten, T., et al.: The MAterials Simulation Toolkit (MAST) for atomistic modeling of defects and diffusion. Comput. Mater. Sci. **126**, 90–102 (2017). https://doi.org/10.1016/j.commatsci.2016.09.018
8. Saal, J.E., Kirklin, S., Aykol, M., et al.: Materials Design and Discovery with High-Throughput Density Functional Theory: The Open Quantum Materials Database (OQMD). JOM **65**, 1501–1509 (2013). https://doi.org/10.1007/s11837-013-0755-4
9. Cookie cutter recipe for AiiDA plugins. AiiDA team. https://github.com/aiidateam/aiida-plugin-cutter. Accessed 28 May 2019
10. AiiDA Team AiiDA registry of plugins. https://aiidateam.github.io/aiida-registry/. Accessed 28 May 2019
11. von Laszewski, G., Hategan, M., and Kodeboyina, D.: Java CoG Kit Workflow. In: Taylor IJ, Deelman E, Gannon DB, Shields M (eds) Workflows for e-Science: Scientific Workflows for Grids. Springer London, London, 2007. Pp. 340–356.
12. Fahringer, T., Prodan, R., Rubing, Duan, et al.: ASKALON: a Grid application development and computing environment. In: The 6th IEEE/ACM International Workshop on Grid Computing, 2005. 10 p.
13. Jain, A., Ong, S.P., Chen, W., et al.: FireWorks: a dynamic workflow system designed for high-throughput applications. Concurr. Comput. Pract. Exp. **27**, 5037–5059 (2015). https://doi.org/10.1002/cpe.3505
14. Amstutz, P., Crusoe, M.R., Tijanić, N., et al.: Common Workflow Language, v1.0 (206).
15. Giannozzi, P., Baroni, S., Bonini, N., et al.: QUANTUM ESPRESSO: a modular and open-source software project for quantum simulations of materials. J. Phys. Condens. Matter. **21**, 395502 (2009). https://doi.org/10.1088/0953-8984/21/39/395502
16. Hutter, J., Iannuzzi, M., Schiffmann, F., and Vandevondele, J.: Cp2k: Atomistic simulations of condensed matter systems. Wiley Interdiscip. Rev. Comput. Mol. Sci. **4**, 15–25 (2014). https://doi.org/10.1002/wcms.1159
17. Schütt, O.: (2019) Appmode: a Jupyter extension that turns notebooks into web applications. https://github.com/oschuett/appmode. Accessed 28 May 2019
18. Re3data.Org (2018) Materials Cloud Archive. https://doi.org/10.17616/r3zj5w
19. FAIRsharing Team (2018) Materials Cloud
20. Recommended Data Repositories Scientific Data. https://www.nature.com/sdata/policies/repositories. Accessed 29 Nov 2018

New Thermodynamic Database and Software for the Applied Chemistry and Materials Science

A.L. Voskov[1], I.A. Uspenskaya[1], N.A. Kovalenko[1], I.B. Kutsenok[1], A.V. Gorbachev[1], N.M. Konstantinova[1], I. Bazhenova[2], A.V. Khvan[2]

[1] Lomonosov MSU, Leninskye Gory, 1/3, Moscow 119991, Russia
[2] NUST MISIS, Leninsky Prospekt, 119049 Moscow, Russia
ira@td.chem.msu.ru, alvoskov@gmail.com

Abstract. The approaches applicable for development of a new thermodynamic database and software for the applied chemistry and materials science are discussed. The framework (platform) for the database was developed. It includes tools for database of thermodynamic properties of substances, thermodynamic models parameters of both individual substances and multicomponent solutions. It also supports Redlich–Kister polynomials, Local Composition (NRTL, UNIQUAC, GLCM) and Pitzer models “from the box”. Polynomials and Einstein functions are supported for description of thermodynamic functions of individual substances. The framework includes GUI for visualization of thermodynamic functions and phase diagrams.

Keywords: Thermodynamic Database, Self-Consistent Data, Equilibria Modeling.

1 Introduction

Nowadays thermodynamic calculations are widely used for modeling of chemical processes to provide optimal conditions of desirable phase’s synthesis, to predict its stability, to calculate phase equilibria, material and heat balances in the development of new technologies and so on.

New programs and databases for such calculations are presented on the world software market, but similar large-scale developments in Russia are absent, with the exception of the JIHT RAS (IVTANTHERMO) database. The main advantage of this database is the self-consistent set of the thermodynamic properties for a large number of individual substances, while the main drawback is the inability to simulate the properties of condensed solutions and limited software capabilities (in fact, only chemical equilibria can be calculated). Keeping it in mind, one of the actual tasks for the Russian thermodynamic community is the creation and development of an innovative software platform for calculating phase and chemical equilibria for the needs of modern chemistry and materials science.

An important issue that must be taken into account during the platform development is appearing of so-called third generation CALPHAD database that are designed for description of thermodynamic properties of substances in both low-(0-298.15 K) and high-temperature (up to compound melting or decomposition temperature) ranges using one self-consistent set of parameters and functions. Such kind of databases for elements began to develop several years ago [1], there are also examples of third generation CALPHAD databases for simple oxides [2] and complex oxides such as zeolites [3]. They usually use either combination of Einstein or Debye functions with polynomials [1] or just weighted sum of several Einstein functions [2, 3]. The similar approach that includes usage of exponent-based parameters in thermodynamic models of multicomponent solution for extension of its temperature range is described in [4].

The aim of this work is development of framework, i.e. software platform that is suitable for construction of thermodynamic databases using a wide range of thermodynamic models. It will not just reproduce existing solutions, but include more flexible tools for implementation and customization of thermodynamic models to allow easier construction of so-called third generation CALPHAD databases. It also includes additional features for modeling of systems that include aqueous solutions of electrolytes.

2 Thermodynamic Database Structure

There are five key functions that are required for construction of thermodynamic database that is applicable for solving practical tasks in applied chemistry and materials science. The first one is tools for systematic collection of existing experimental data with the corresponding bibliographical information. The second one is tools for keeping model parameters for both individual substances and multicomponent solutions. The third one is software for calculation of phase and chemical equilibria using the supplied thermodynamic models. The fourth one is subroutines for implementation of thermodynamic models and some pre-programmed models, e.g. polynomials, Einstein functions, Pitzer model etc. And the fifth one is module for solution of so called “inverse problems”, i.e. optimization of model parameters in a consistent manner using existing experimental data and by means of nonlinear regression. The interconnection between the modules is shown in Fig. 1. GUI is graphical user interface used for control over the framework.

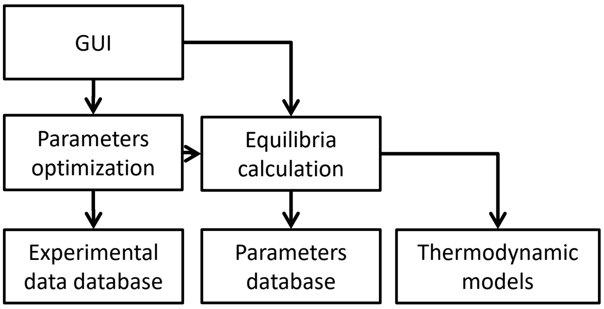


Fig. 1. The structure of the developed framework

The developed framework for implementation of thermodynamic databases includes all of them and is suitable for solution of practical tasks of materials science and applied chemistry. It consists from five key modules (parts) that are described below. All of them are accessible by means of GUI (graphic user interface).

2.1 Experimental Thermodynamic Data Database

This database contains the raw experimental data on thermodynamic properties of phases, both individual substances and multicomponent solutions. It supports the next important features:

- Raw data repository data that keeps experimental data points values, assigned statistical weights and bibliographical references.
- Bibliographical links including full bibliographical info, DOI, copies of original papers etc.

2.2 Parameters of Thermodynamic Database

This block is database of parameters sets for thermodynamic models. It supports the next three kinds of parameters:

- Thermodynamic functions of individual substances, i.e. heat capacity, heat content, enthalpy of formation, entropy etc. Both polynomials and weighted Einstein functions sum are supported “from the box”.
- Parameters of thermodynamic models of multicomponent solutions, e.g. solid solutions, melts, aqueous solutions etc.
- Parameters of incremental models for estimation of the thermodynamic properties of uninvestigated substances with established compositions and structures.

Its unique feature is support of arbitrary algebraic user-defined forms of parameters based on Lua programming language. However, it’s possible to use some pre-defined functions such as polynomials, Einstein functions, Debye function etc. Usage of this module provides self-consistent values of different thermodynamic functions of substances such as heat capacity, enthalpy, entropy etc. Flexibility of the used format allows to easily import SGTE database for elements.

2.3 Thermodynamic Models

Although users can implement required thermodynamic models directly using either Lua or C++ programming languages, it is a complicated approach that requires computer programming skills. To make modeling easier, the developed framework contains some subroutines for implementation of thermodynamic models (such as functions for conversion between different concentration scales, molecular masses calculation etc.). Also there are some basic pre-defined models in the framework:

- For individual substances: polynomials, Einstein–Planck functions, Debye function.

- For multicomponent solutions: Redlich–Kister polynomials, Compound energy formalism, Pitzer model in molality scale.
- Different Methods for extrapolation of binary and ternary systems models to multicomponent systems.

2.4 Equilbria Calculation Module

Equilibria calculation module is designed for calculation of phase and chemical equilibria using thermodynamic models of individual substances and multicomponent solid, liquid and gaseous non-ideal solutions. It allows evaluation of both equilibrium composition and thermodynamic functions for a given composition and construct phase diagrams. It supports usage of both traditional approaches implemented in modern software platforms for thermodynamic calculations (THERMO-CALC, MTDATA, PyCalphad, Open Calphad, FACTSAGE, HYSIS, PANDAT, OLI, THEREDA, PHREEQC, etc.) and original methods (the convex hull method, description of temperature dependences using the combination of Planck–Einstein functions et al. [2, 3, 5–8]).

The convex hull method advantage on traditional approach based on Gibbs energy global minimization is more reliable search of the global minima, i.e. true equilibrium, without any requirements for initial approximations. The developed module uses robust and widely applied qhull library [8] for construction of convex hulls. The results obtained by the convex hull method can be used as an initial approximation for traditional methods. The traditional methods may be required to improve accuracy of the convex hull method output.

Einstein–Planck functions allow to approximate experimental data (their non-anomalous part) in a whole temperature range, i.e. from 0 K to the melting points.

2.5 Parameters Optimizer

The optimizer includes tools for solution of so called “inverse task”, i.e. finding thermodynamic models parameters using non-linear regression based on the least squares method. It uses experimental data database and equilibria calculator as its basis and existing well-tested implementation of the least squares method (Ceres Solver [9]). Support of both individual substances and multicomponent solutions is included. The important features of the software:

- Estimation of confidence intervals for both evaluated parameters and obtained thermodynamic functions. It can be used as a protection from model overfitting and for comparison of different fitting functions.
- Tools for comfortable optimization of data sets from different sources (automatic recalculation of units, improved statistical weight assignment etc.).

3 Examples

The developed framework includes not only software for construction of thermodynamic database but also small database for demonstration of its functionality. It includes thermodynamic functions based on weighted Einstein–Planck functions sums for the next substances:

- Several solid elements: Sn and Pb (from 0 K to melting points)
- Two oxides: uranium oxide UO_2 and thorium oxide ThO_2 (from 0 K to about 3000 K, i.e. up to their melting points). Thermodynamic models are taken from [2].
- 46 zeolites: both functions of individual substances and additive scheme for prediction of thermodynamic properties of zeolites using its compositions. Thermodynamic models are taken from [3].

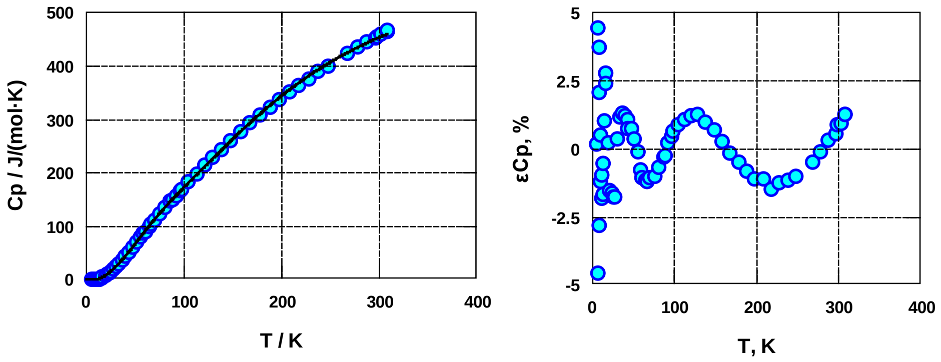


Fig. 2. Approximation of leonhardite $Na_{0.33}K_{0.43}Ca_{0.61}Al_{1.98}Si_{4.02}O_{12}$*3.27$H_2O$ heat capacity by the Einstein–Planck functions sum. Thermodynamic model is taken from [3]

There are also thermodynamic models of several binary and ternary systems that are based either on Redlich–Kister polynomials (metallic systems, oxide systems etc.) or on Pitzer molality scale models (binary and ternary aqueous solutions of electrolytes).

4 Further Development

The developed framework for construction of thermodynamic databases and calculation of phase and chemical equilibria could be extended for the better practical applicability. The most prospective ways of extension are:

- Addition of the sublattices based models implementation. It will be required for thermodynamic modeling of some intermetallic and mixed oxide compounds and phase equilibria with their participation.
- Addition of the most popular models for electrolytes solutions, such as eNRTL, eUNIQUAC, ePC-SAFT, Pitzer model in molar fraction scale etc.
- Extension of the database: third generation CALPHAD database for the elements; database for electrolytes and some metallic and non-metallic systems.

References

1. Bigdeli, S., Mao, H., and Selleby, M.: On the third generation Calphad databased: an updated description of Mn. Phys. Status Solidi B. **252** (10), 2199–2209 (2015).
2. Voskov, A.L., Kutsenok, I.B., and Voronin, G.F.: CpFit program for approximation of heat capacities and enthalpies by Einstein–Planck functions sum. Calphad. **61**, 50–61 (2018).
3. Voskov, A.L., Voronin, G.F., Kutsenok, I.B., and Kozin, N.Yu.: Thermodynamic database of zeolites and new method of their thermodynamic properties evaluation for a wide temperature. Calphad. **66**, 101623 (2019).
4. Kaptay, G.: A new equation for the temperature dependence of the excess Gibbs energy of solution phases. Calphad**. 28** (2), 115–124 (2004).
5. Voronin, G.F.: Convex functions in the thermodynamics of heterogeneous substances. Russ. J. Phys. Chem. A, **79** (12), 1890–1902 (2005).
6. Voronin, G.F. and Kutsenok, I.B.: Universal method for approximating the standard thermodynamic functions of solids. J. Chem. Eng. Data **58** (7), 2083–2094 (2013).
7. Voskov, A.L., Dzuban, A.V., and Maksimov, A.I.: TernAPI program for the calculation of ternary phase diagrams with isolated miscibility gaps by the convex hull method. Fluid Phase Equilibria **388**, 50–59 (2015).
8. Barber, C.B., Dobkin, D.P., and Huhdanpaa, H.T.: The Quickhull algorithm for convex hulls. ACM Trans. On Mathematical Software **22** (4), 469–483 (1996).
9. Agarwal, S., Mierle, K. et al.: Ceres Solver. http://ceres-solver.org

Relevance Evaluation of Information Retrieval in the Integration of Information Systems on Inorganic Substances Properties

V.A. Dudarev[1][0000-0002-3583-0704], N.N. Kiselyova[1][0000-0001-7243-9096], and I.O. Temkin[2]

[1] A.A. Baikov Institute of Metallurgy and Materials Science of RAS (IMET RAS),
Moscow, 119334, Russia
[2] National University of Science and Technology MISIS (Moscow Institute of Steel and
Alloys), Moscow, 119049, Russia
vic@imet.ac.ru

Abstract. One of the main tasks in the integration of information systems is to provide relevant retrieval of information consolidated from heterogeneous sources. In the field of inorganic chemistry and materials science, set-theoretic methods of searching for relevant information are known. They ensure the construction of a sufficiently high-quality response to user requests. However, the problem of quantifying evaluation of information search relevance in this subject area remains open. This paper proposes an approach to quantifying evaluation of the relevance of information retrieval in integrated systems on inorganic substances and materials properties.

Keywords: relevance evaluation, database integration, inorganic substances

1 Introduction

The development and use of integrated information systems on substances and materials properties that consolidate information from heterogeneous information sources is worldwide common trend. These systems ensure that specialists are capable to quickly find the required information. When developing such systems, the fundamental thing is data representation method that describes corresponding chemical objects and their properties. Furthermore, chemical objects data representation method, in its turn, determines the class of methods for ensuring the search for relevant information and their functionality. The purpose of this paper is to present a new approach for quantifying evaluation of the relevance of information retrieval for integrated information systems (IS) on inorganic substances and materials properties (ISMP) based on information structures describing the qualitative and/or quantitative substance composition.

2 The Current State of the Problem

2.1 Heterogeneous Information Systems

The information technologies development and the emergence of powerful hardware and software tools for storing and processing information stimulated works on information systems development in the field of inorganic materials science. As a result, a large number of highly specialized information systems have been developed that are focused on solving problems with due regard for specificity, conditioned by a specific subject domain and research areas of a specific organization developing IS. An example is a number of information systems based on databases developed and maintained by IMET RAS. The IMET RAS information systems core consists of a number of databases which store data on a variety of properties of substances:

- “Diagram” – database (DB) on the phase diagrams of semiconductor systems;
- “Crystal” – DB on the properties of acoustooptical, electro-optical and nonlinear-optical substances;
- “Phases” – DB on the general properties of ternary and quaternary compounds;
- “Bandgap” – DB on the band gap of inorganic substances [1];
- “Elements” – DB on the properties of chemical elements.

These databases are heterogeneous not only by data structures, but also by software and hardware tools ensuring their operation [2]. It should be noted that above mentioned DBs contain extensive information, but in a fairly narrow area. The situation when none of the developed information systems contains a complete set of data on properties of an object (substance or material) and the specialist needs to use several information resources at once to search for the necessary information is typical not only for inorganic materials science, but also for other subject domains.

Obviously, to ensure a high-quality information service for materials scientists, information systems integration in this subject domain is necessary. In Russia, the first successful attempts in this direction were undertaken at the beginning of the century at the IMET RAS for the integration of information systems mostly used by Russian users [3]. The integration allowed a consolidation of information resources for end users and a significant reduction of the time spent by specialists to find the necessary information. The applied consolidation approach was based on the Enterprise Application Integration (EAI) method and showed its efficiency and good scalability when connecting resources developed in different organizations to the integrated information system [8]. For example, «TCS» (Thermal Constants of Substances - reference book on substances thermal constants, developed by the Joint Institute for High Temperatures of Russian Academy of Sciences (JIHT RAS) together with the Moscow State University (MSU)) and “AtomWork” (information system on inorganic substances properties, developed by the National Institute for Materials Science (NIMS), Japan) are among successfully integrated systems [4].

One of the main difficulties in the heterogeneous information system (IS) integration is the diversity of the chemical objects described in them. So, for example, “Diagram" IS contains information at the level of the chemical system, i.e. a set of chemical elements that form a certain phase diagram of a semiconductor system. Other IS on inorganic substances and materials properties (ISMP) describe the properties at a specific quantitative composition level (with a specific ratio of elements in chemical system), taking into account crystalline modifications of substances, i.e. the quantitative composition of the substance and its crystal lattice are described at this level. Such chemical objects descriptions incompatibility in different IS ISMP dictates the need to use a different description of chemical objects in an integrated IS ISMP, at least it's required to distinguish between several types of chemical objects: chemical systems, substances and their crystal modifications.

2.2 Chemical Objects Hierarchy

To describe the basic chemical objects of the considered problem domain the set theory is used, taking into account that each subsequent level in the problem domain hierarchy complements the description of the chemical object. The notation is the following: *S* is the set of chemical systems; *C* is the set of chemical substances, i.e. chemical compounds, solid solutions, heterogeneous mixtures, etc.; *M* is the set of crystal modifications. Then the chemical system is denoted as *s* (where $s \in S$), the chemical substance is denoted by *c* (where $c \in C$), and the crystal modifications is *m* (where $m \in M$).

Having designated second level objects by the "substance" term, we get three-level chemical objects hierarchy: chemical system, chemical substance and chemical modification [5]. As far as information stored in DBs on inorganic substances properties can be considered at chemical system level, for simplicity we'll use this level from the top of the objects hierarchy. So, the chemical objects hierarchy and relationships between chemical objects can be described by means of chemical objects hierarchy in tree form (Fig. 1).

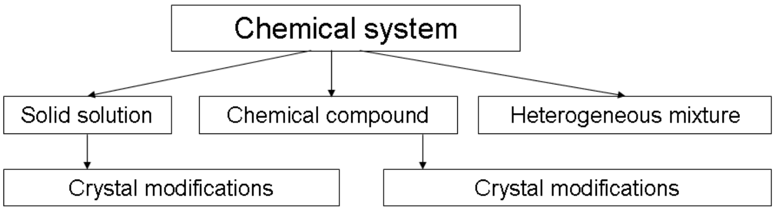


Fig. 1. Chemical objects hierarchy

Any chemical system *s* can be represented as a set of chemical elements e_i: $s = \{e_1, e_2, \dots, e_n\}$. Any chemical substance *c* is defined not only by the set of atoms (chemical elements), but also by their quantitative incorporation into the composition of the compound, solution or mixture. Therefore, any substance *c* can be represented by a tuple (*s*, *f*), where $s \in S$, and *f* is a mapping of the set of atoms (chemical elements) that make up the substance, in the set of $\boldsymbol{R}^* \times \boldsymbol{R}^*$ pairs that define the minimum and maximum incorporation of a given chemical element in a compound, solution or mixture *c*.

That is, f: $e_i \rightarrow (\boldsymbol{R}^*_{min}, \boldsymbol{R}^*_{max})$, where $\boldsymbol{R}^* = \boldsymbol{R}^+ \cup \{x\}$. $\boldsymbol{R}^+$ is the set of non-negative real numbers, and $\boldsymbol{R}$* is the set of $\boldsymbol{R}^+$ extended by the element x. The element x is used to denote an unknown number, since in the notation of mixtures where the incorporation of components may vary, it is customary to use x to denote an unknown, for example, $Fe_{1-x}Se_x$. $\boldsymbol{R}^*_{min}$ and $\boldsymbol{R}^*_{max}$ are, respectively, the minimum and maximum concentration of the chemical element e_i in the substance c.

In the case when the concentration of a particular chemical element e_i in the substance c is fixed, then $\boldsymbol{R}^*_{min} = \boldsymbol{R}^*_{max}$. Chemical modification m can be represented by a tuple (s, f, mod), where $s \in S$, f: $e_i \rightarrow (\boldsymbol{R}^*_{min}, \boldsymbol{R}^*_{max})$, and mod is the string notation for the crystal modification of a substance – common for integrated IS ISMP (one of the singony enumeration values: {*Triclinic*, *Monoclinic*, *Orthorhombic*, *Tetragonal*, *Trigonal*, *Hexagonal*, *Cubic*}).

2.3 Metabase Structure

Quite reasonably, when designing integrated IS ISMP, it's required to provide search facilities for relevant information contained in other IS ISMP of distributed system. Therefore, it's required to develop some active data store that should "know" what information is contained in every integrated IS ISMP. Considering chemical objects hierarchy, some database should exist that describes information contained in integrated resources in terms of chemical systems, substances and crystal modifications. Here we come to the metabase concept – a special database that contains metadata that describe integrated IS ISMP contents in terms of chemical objects hierarchy as well as some additional information on users and their permissions together with information required to integrate distributed IS ISMP (Fig. 2).

The metabase defines integrated IS capabilities. Its structure should be flexible enough to represent metadata on integrated ISs ISMP contents and at the same time the metabase structure should be simple and versatile to describe arbitrary data source on inorganic substances properties without exhaustive additional payload currently offered by numerous materials ontologies. Taking into consideration the fact that chemical objects and their corresponding properties description is given at different detail level in different ISs ISMP, it's important to develop metabase structure that would be suitable for description of information residing in different ISs ISMP. For example, some integrated DBs contain information on particular crystal modifications properties while others contain properties description at chemical system level. Thus, integrated ISs ISMP deal with different chemical objects situated at different chemical objects hierarchy levels. For simplicity in current paper we consider only a part of metabase structure that is devoted to chemical systems and their properties (Fig. 2). The amount of this metainformation should be enough to perform search for relevant information on systems and corresponding properties.

All tables (Fig. 2) can be logically separated into several groups according to their purpose:

- DBInfo – root table, that contains information on integrated database systems;

- DBExcludeCompatibility – table that stores exception list of ISs for relevant information search;
- UsersInfo, UsersAccess – tables that contain information on integrated system users and their access rights to integrated IS ISMP;
- SystemInfo, PropertiesInfo, DBContent – tables that describe contents of integrated IS ISMP;
- CompatibilityClasses, Compatibility, Systems2ConsiderInCompatibility – tables that contain information on accessible relevance classes and their contents (currently 3 relevance classes are used [4]);
- Meta_Systems, Meta_DBSystems, Meta_SystemsHierarchy, Meta_SystemsElement – tables to describe all chemical systems contained within integrated IS ISMP with respect to their relation to each other and chemical elements, they consist of;
- Versions – service table (not shown on diagram). It is used for database schema update and versioning.

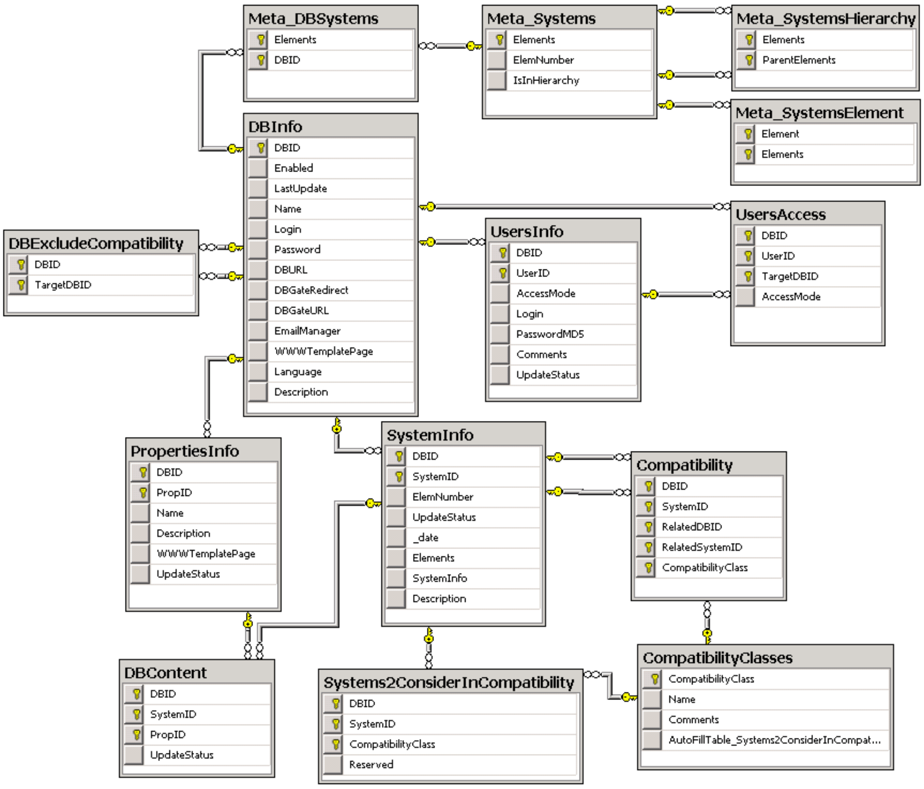


Fig. 2. A Part of the metabase logical structure on the chemical systems level only

Taking into account chemical objects hierarchy description, a special method was developed to search for relevant information in the context of an integrated information system, based on a set-theoretic approach [5].

3 A Set-theoretic Approach to Relevance Evaluation

Relevance itself and its notion to information search is a philosophic term, covered in numerous publications. A comprehensive review of relevance itself is given by Tefko Saracevic [6]. We consider information search relevance in application to integrated IS ISMP, that area is close to "chemical similarity" [9]. So, considering chemical objects hierarchy description, a special method was developed to search for relevant information in the context of an integrated IS ISMP, based on a set-theoretic approach [5]. The main essence of set-theoretic approach is in the use of abovementioned metabase structure, that is a special database that contains information on integrable IS ISMP (set D), chemical systems (set S) and their properties (set P). To describe the relationship between the elements of the sets D, S, and P, the ternary relation W was defined on the set U (universum), which is the Cartesian product: $U=D\times S\times P$. The element (d, s, p) belongs to the relation W, where $d\in D$, $s\in S$, $p\in P$ is interpreted as follows: "the integrable IS ISMP d contains information on the p property of the chemical system s".

Thus, according to accepted notation the search for relevant information on a particular chemical system s can be reduced to proper definition of an R relation, which is a subset of the $S\times S$ Cartesian product (in other words, $R\subset S^2$). Thus, for any pair $(s_1,s_2)\in R$, we can state that the s_2 system is relevant to the s_1 system. For the practical solution of the problems of searching for relevant information in integrable information systems, the following rules are often used to construct R [3]:

4. For any set $s_1\in S$, $s_2\in S$, which includes the notation of chemical elements e_{ij}, $s_1=\{e_{11}, e_{12}, ..., e_{1n}\}$, $s_2=\{e_{21}, e_{22}, ..., e_{2n}\}$, it's true, that if $s_1\subseteq s_2$ (that is, all chemical elements from s_1 system are contained in s_2 system), then $(s_1,s_2)\in R$.
5. The relation R is symmetric. In other words, for any $s_1\in S$, $s_2\in S$ it is true that, if $(s_1,s_2)\in R$, then $(s_2,s_1)\in R$.

It should be noted that abovementioned automatic variant of R relation generation is just one of the simplest and most obvious variants of such rules, and in fact more complex mechanisms can be used to get R relation. Other alternatives are used to build the R relation, called *relevance classes*. For example, browsing information on a particular property of a compound in one of integrated IS ISMP (in fact, it is information defined by (d_1, s_1, p_1) triplet), we consider (d_2, s_2, p_2) triplet to be relevant information. (d_2, s_2, p_2) triplet characterizes information on some other property of a chemical system from another integrated IS ISMP. This enables us to define relevant information more precisely, e.g. if we consider the R relations in the form: $R\subset(d_1, s_1, p_1)\times(d_2, s_2, p_2)$, where $d_1, d_2\in D$, $s_1, s_2\in S$, $p_1, p_2\in P$. Actually, it's possible to even define a set of several R relations $(R_1, R_2, \ldots, R_n)$ by applying different rules to enable users to perform search for relevant information based on wide variety of R interpretations. However complex interpretations of R $(R\subset(d_1, s_1, p_1)\times(d_2, s_2, p_2))$ are not being currently used in IMET RAS, since metabase structure would be more complex to store such relations however its reasonability is not so clear. In IMET RAS simple relevancy relations of $R \subset S^2$ are used. More rules to form *relevance classes* are given in [4].

Improvement of the search relevance can also be achieved by using the c_i level, i.e. taking into account the quantitative composition of a substance, or crystal modifications of a specific substance m_i instead of chemical system designations s_i in cases when a user requests relevant information, being at the level of inorganic substances or their modifications in the system-substance-modification hierarchy concepts [5].

When searching at the substance level, the quantitative compound composition is taken into account. The pair ($a_{i\min}$, $a_{i\max}$) denotes the quantitative inclusion of chemical element $e_i \in s$ into the composition, $a_{i\min}$, $a_{i\max}{\in}R^+$, $a_{i\min}{\leq}a_{i\max}$. If $a_{i\min}{=}a_{i\max}$, then the substance has a constant composition by the element $e_i{\in}s$. For each element of the chemical system $e_i{\in}s$, user during the search could specify a pair ($r_{i\min}$,$r_{i\max}$), where $r_{i\min}$, $r_{i\max}{\in}\boldsymbol{R}^+$, denoting the allowable interval of the *i*-th element in the substance (R^+ is the set of non-negative real numbers). Then all substances belonging to the same chemical system are considered relevant, if for each pair ($r_{i\min}$, $r_{i\max}$) the following is correct: $a_{i\min}{\in}[r_{i\min},r_{i\max}]$ or $a_{i\max}{\in}[r_{i\min},r_{i\max}]$. In other words, if the logical disjunction $[r_{i\min}{\leq}a_{i\min}\&a_{i\min}{\leq}r_{i\max}]+[r_{i\min}{\leq}a_{i\max}\&a_{i\max}{\leq}r_{i\max}]{=}\mathit{true}$ for all $e_i{\in}s$, then the data on the substance are considered relevant.

When searching for relevant information taking into account the crystal modifications of m_i, crystal systems are taken into account, since often information on crystal structures is shown in different ways. For example, for lithium niobate ($LiNbO_3$) a hexagonal or trigonal crystallographic system is indicated in different information sources of the IS ISMP, which, in fact, corresponds to the same crystal modification.

However, it should be noted that despite the fact that the described approach, in general, provides an acceptable level of search relevance for inorganic compounds, it suffers from the inability to obtain a quantitative assessment of the search relevance and, as a consequence, the fundamental inability of search results changes by adjusting some parameters or corresponding metrics. Note that such an adjustment is useful in some cases, in particular when preparing training data sets for machine learning tasks in computer-aided construction of inorganic compounds [7].

4 Graph Approach to Relevance Assessment

To search for relevant information and obtain a quantitative measure of relevance assessment within an integrated information system based on the properties of inorganic substances and materials, we propose to use a graph model based on the weighted graph $G{=}(V,E)$, built on chemical objects described as part of an integrated information system.

Let's define a set of vertices V for graph G. In accordance with the accepted three-level description of chemical objects in an integrated information system, the set of vertices consists of three disjoint subsets $V{=}\{S,C,M\}$, where S is the set of chemical systems s_i (qualitative compound composition), C is the set of chemical compounds c_i (the quantitative compound composition or the substance formula), M is the set of crystal modifications m_i of specific substances.

Define a set of edges E for graph G, as the union of non-intersecting subsets $E=Es\cup Ec\cup Em\cup Esc\cup\ cm$, where Es *is* the edges that are incidental only to the set of vertices S; Ec *is* the edges that are incidental only to the set of substances C; Em *is* the edges that are incidental only to modifications set M. The vertices connectivity for the classes of S, C, M is achieved by two sets of edges: Esc edges to connect vertices from S and C sets; and Ecm edges to connect vertices from C and M. Please note, that the edges connecting vertices from S and M sets, are absent.

To define the elements of the E subsets we need to introduce a couple of trivial functions: $Fs(c)$ and $Fc(m)$. The $Fs(c)$ function returns the chemical system for a given compound c, i.e. it allows to get qualitative composition from quantitative composition. The $Fc(m)$ function returns quantitative composition of a particular crystal modification of the substance, i.e. it allows to get quantitative composition from a particular crystal structure of the compound. Then, given that the chemical system is a set of chemical elements $s=\{e_1, e_2, ..., e_n\}$ we get the following set of edges:

$$Es=\{(s_i, s_j)\},\ s_i=\{e_{i1}, e_{i2}, ..., e_{in}\},\ s_j=\{e_{j1}, e_{j2}, ..., e_{jm}\},\ |s_i|=n,\ |s_j|=m,\ m–n=1,\ s_i\subset s_j;$$

$$Ec=\{(c_i,c_j)\},\text{ where } Fs(c_i)=s(c_j);$$

$$Em=\{(m_i, m_j)\},\text{ where } Fc(m_i)=Fc(m_j);$$

$$Esc=\{(s_i, c_j)\},\text{ where } Fs(c_j)=s_i;$$

$$Ecm=\{(c_i, m_j)\},\text{ where } Fc(m_j)=c_i.$$

When searching for relevant information for a chemical object, it is necessary that a path should exist in graph between the corresponding object and a relevant one, and it is easy to calculate the measure of relevance by adding the weights of the edges on the corresponding path. Thus, we come to the necessity of introducing a real-valued function W defined on the set of graph edges:

$$W(Es)=1000;\ W(Ec)=W((c_i,c_j))=\min\left(\sum_{k=0}^{n} 10^k|q_{ik}-q_{jk}|\right);$$

where $n=|Fs(c_i)|=|Fs(c_j)|$, q_{ik} and q_{jk} are quantitative occurrence of k-th element at c_i and c_j compositions, i.e. Q: $e_k\rightarrow \boldsymbol{R}^+$ (respectively $Q(el_{ik})=q_{ik}$, $Q(e_{jk})=q_{jk}$), and the order of elements in substances is selected so to ensure the minimum value of the $W(Ec)$ objective function.

$$W(Em)= 0.1;\ W(Esc)=100;\ W(Ecm)=1.$$

As an example, we give a fragment of the relevance graph for chemical systems Cu-In-S and In-S (Fig. 3). On this example we emphasize its properties and justify the role of edge weights for quantitative assessment of the chemical objects' relevance.

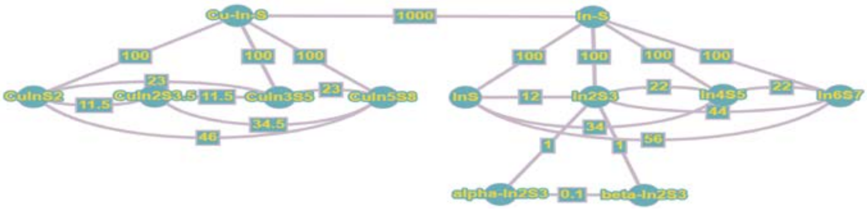


Fig. 3. Fragment of the relevance graph for Cu-In-S and In-S chemical systems

Based on the definition of the set of edges E, it can be seen that the relevance graph is partitioned into subgraphs based on the vertices of a set S (chemical systems). Moreover, there is no path in the graph between substances from different chemical systems, bypassing the vertices of chemical systems. The vertices of the systems themselves are connected by an edge only if the set of elements of one of the systems is an own subset of the other system and their powers (i.e. a number of chemical elements that built up a system) differ by one.

Consider a subgraph constructed on the basis of the In-S chemical system vertex and consisting of substances and their corresponding modifications related to this system. It should be noted that the subgraph composed of vertices of a *C* set (compounds, i.e. qualitative formula) is complete, as all the vertices (InS, In_2S_3, In_4S_5, In_6S_7) are connected to each other and form a clique. Note, however, that the weights of the ribs connecting the vertex substances are different. Edge weight is a quantity characterizing the degree of closeness of corresponding quantitative compositions: the smaller the difference, the lower the weight («cost») of transition along the edge, and the corresponding substance is considered more relevant than other with greater weight of transition.

Similarly, modifications subgraph constructed on the basis of the vertex designating a particular compound is complete, and the weights of all edges are equal to 0.1. In Fig. 3 such edges are connected to each other, e.g. α-In_2S_3 and β-In_2S_3 vertices. Note, that the transition from modification to the corresponding substance has a cost of 1, and the transition from substance to the system – 100, which makes more relevant data on other modifications (including crystal structure) than the transition to the level of substances to choose another qualitative composition.

5 Discussion and Further Model Development

The proposed graph model is an attempt to reflect the similarity degree of various chemical objects even at different representation level (system, compound, modification). In this sense, the path cost is a measure of the difference between the corresponding chemical objects, which are the vertices of the graph. The more similar the objects, the “closer” they are, meaning the path cost in the graph is less. It is worth noting that, in a broad sense, according to the definitions given in the paper, the overall relevance graph is disconnected due to the absence of a path between the vertices of chemical systems, that have no common chemical elements (i.e. $s_1 \in S$, $s_2 \in S$ such that $s_1 \cap s_2 = \emptyset$). For example, in the current model, there is no connectivity between In-S and Ga-As chemical systems, although In and Ga are similar in many ways, as far as In and Ga are elements from the same subgroup of the periodic system. In this sense, it is advisable to introduce rules for the formation of edges between similar substances and systems (in which an element from the same periodic system subgroup changes), although such an edge should have an appropriate (sufficiently large) weight comparing with analogues with common chemical elements.

As possible ways of further graph model development, one can offer the transformation of edges from the sets *Esc* and *Ecm* in pairs of arcs. In this case, the weight of the arc in the direction from the modification to the substance and from the substance to the system should be made much less than the weight of the original edge, and the reverse arc should preserve the original edge weight. This measure will allow to obtain relevant information, described one or two levels above, which is a common way of information search in chemistry.

6 Conclusion

In the paper by means of the graph model, the concept of relevant information search was extended regarding to integrated IS ISMP. The new model allows to obtain quantitative relevance assessment of information retrieval based on the path calculation in a weighted graph, which allows ranking of chemical information found in consolidated data sources. The proposed approach is applicable not only to improve information retrieval for end users – material chemists, but also for application to computer aided design of inorganic compounds at the stage of training samples formation based on the quantitative relevance assessment.

Acknowledgments

This work was partially supported by the Russian Foundation for Basic Research (project No. 18-07-00080) and the State task No. 075-00746-19-00.

References

1. Kiselyova, N.N., Dudarev, V.A., and Korzhuyev, M.A.: Database on the bandgap of inorganic substances and materials, Inorganic Materials: Applied Research **7** (1), 34–39 (2016).
2. Kiseleva, N.N., Prokoshev, I.V., Dudarev, V.A., Khorbenko, V.V., Belokurova, I.N., Podbelsky, V.V., and Zemskov, V.S.: Database system on materials for electronics on the Internet. Inorganic Materials **40** (3), 380–384 (2004).
3. Kornyshko, V.F. and Dudarev, V.A.: Software development for distributed electronics materials. In proceedings of the Third International Conference "Information Research, Applications and Education – i.Tech ", Sofia, FOI-Commerce, 27–33 (2005).
4. Dudarev, V.A., Kiselyova, N.N., Xu, Y., and Yamazaki, M.: Virtual integration of the Russian and Japanese databases on properties of inorganic substances and materials. MITS 2009. In Proceedings of Symposium on Materials Database, National Institute for Materials Science (NIMS), Materials Database Station (MDBS), 37–48 (2009).
5. Dudarev, V.A.: Integration of information systems in the field of inorganic chemistry and materials science. Moscow: KRASAND, 2016, 320 p.
6. Saracevic, T.: Relevance: A review of the literature and a framework for thinking on the notion in information science. Part II: nature and manifestations of relevance. Journal of the American Society for Information Science and Technology **58** (3), 1915–1933 (2007).
7. Sen'ko, O.V., Kiselyova, N.N, Dudarev, V.A., Dokukin, A.A., and Ryazanov, V.V.: Various machine learning methods efficiency comparison in application to inorganic compounds design. In Selected Papers of the Data Analytics and Management in Data Intensive Domains. Proceedings of the XX International Conference – DAMDID / RCDL'2018, October 9–12, 2018, Moscow **2277**, 152–158 (2018).
8. Serain, D. Middleware and Enterprise Application Integration. London: Springer-Verlag, 2002, 288 p.
9. Johnson, A.M. and Maggiora, G.M.: Concepts and Applications of Molecular Similarity. New York: John Willey & Sons, 1990. 393 p.

IT APPLICATIONS AND IT PLATFORMS FOR MATERIALS DESIGN AND SIMULATION

Predicting the Topology of Crystal Structures Using Angular Fingerprints

A.P. Shevchenko[1,2]

[1] Samara University, Samara, Russia

[2] Samara State Technical University, Samara, Russia

allex.shevchenko@gmail.com

Abstract. A method for determining the geometric shape of the coordination figure of structural building units in crystal structure has been proposed. The method that is based on a comparison of angular fingerprints of underlying net nodes was implemented in the ToposPro software package and tested on 6171 crystal structures of coordination compounds with oxo-ligands. The four-coordination figures and the correlation of their shape with the topological type of the underlying net were studied.

Keywords: Coordination figure, Angular fingerprints, Underlying net topology, Coordination compounds, Oxygen containing ligands, Data meaning

1 Introduction

Currently, the scientific community has a great interest in predicting the structure and properties of new crystalline compounds [1]. A variety of descriptors can be used for this, such as chemical composition, physical properties, and the electronic structure of the constituent atoms or atomic ions. This approach is based on the properties of chemical elements and works well for structures of simple inorganic compounds. COF, MOF, intermetallic compounds that are currently under investigation have a complex composition and crystal structure. The use of such a set of descriptors in such cases does not allow explaining and describing the structural features of the crystal and, especially, making any predictions based on them.

In the past 15–20 years, we have seen a great progress in the so-called *topological crystal chemistry* or *reticular chemistry* [2–4]. Such an approach practically has no limitations when analyzing structures of any complexity, since structural building blocks (SBUs), rather than atoms, are used in it. In this case, the original structure is represented in the form of a underlying net (Fig. 1), which is obtained as a result of the simplification of the first, and the SBUs themselves are replaced by their geometric centers of mass. Obviously, SBUs are characterized by a different set of descriptors than a single atom. This study is devoted to detailed exploration of just one of the topological descriptors, the so-called *coordination figure* (CF). Coordination figure is an object that is formed around a given point by all adjacent points [6]. Thus, it can be referred to both topological and geometrical descriptors as it characterizes both local topology of the point environment and its spatial embedding. It is known [2, 7] that the coordination figure significantly affects the topology of the crystal as a whole. However, to the best of our knowledge, there is no strict determination of this descriptor that could be used in computer algorithms to process the crystallographic information.

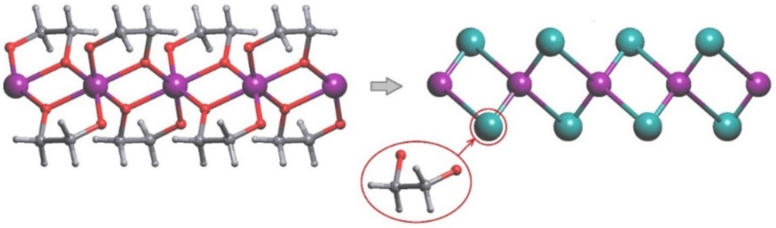


Fig. 1. A fragments of the original structure (left) and its underlying net (right) are shown for compound bis(μ_2-ethylene glycolato)-titanium (CSD Refcode KEPFEQ) [5]

2 Results

We developed one algorithm for determining the shape of a coordination figure [8] and test it on a large sample of crystal structures. The approach that we have used is based on the well-known method for comparing radial fingerprints [9]. The angular fingerprints were calculated according to the following algorithm:

(i) The angles at the central atom were calculated for all pairs of the neighboring atoms of the coordination figure.

(ii) The angle values were arranged over the intervals of width δ in the range 0–180°.

(iii) The occurrences for each interval were smoothed by a Gaussian function (formula (1)) with a standard deviation σ, and an average value μ equal to the mid-interval:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}. \qquad (1)$$

(iv) The total contribution F_i was calculated as a centuplicate sum of function values in each interval, which normalized to the sum of the values of these functions at all intervals (formula (2)).

$$F_i = 100\frac{\sum_{k=1}^{180/\delta} f_k(\mu_i)}{\sum_{m=1}^{180/\delta}\sum_{k=1}^{180/\delta} f_k(\mu_m)}. \qquad (2)$$

As a result, each coordination figure corresponds to a fingerprint-vector F, whose dimension is determined by δ and equal to the number of intervals; F is normalized to 100. The distance r between F and F' fingerprints of two coordination figures (formula (3)) is used to compare them:

$$r = \sqrt{\sum_{i=1}^{180/\delta}\left(\frac{F_i - F'_i}{2}\right)^2}. \qquad (3)$$

The coordinating figure is related to the closest type, to which it has the minimal distance. By definition, this fingerprint is sensitive only to the angular distribution of atoms of the coordination figure. To determine the shape of the coordination figures, we have calculated the fingerprints with δ=18° and σ=9° (Fig. 2), and then computed the distances *r* between the figures. We chose such parameter values δ and σ so that they divided the entire range from 0 to 180°, the double σ value was equal to the δ value, and the length of the fingerprint vector was small (the dimension of the space was 10 in the proposed model). Additional studies have shown that the synchronous change of δ and σ values in two or three times did not significantly affect the results of recognition of coordination figures. The maxima of the fingerprint distributions indicate most typical bond angles for the coordination figures. The Hartshorn abbreviation [10] was used for the names of coordination figures and we supplemented it with new ones. Hereinafter, we will use for the designation of a specific coordination figure sets of angles characterizing it (in degrees), reduced to a scale in steps of 15° (0, 15, 30, 45 ... 150, 165, 180) and ordered in ascending order. Such a transformation is necessary to reduce the variety of coordination figures in solving problems of classification and the search for patterns.

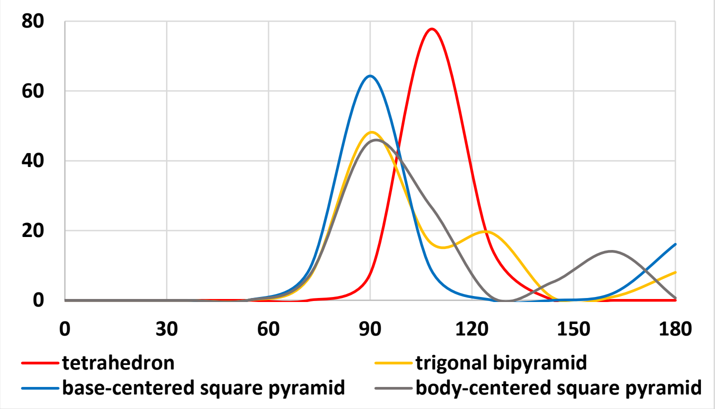


Fig. 2. Fingerprints corresponding to four regular coordination figures. The total area under each curve is equal to 100

So in the underlying net of structure bis(μ₂-ethylene glycolato)-titanium (CSD Reference Code KEPFEQ) [5] (Fig. 1), the ligand corresponds to an angular coordination figure with an angle of 69.13°, from which we get 75° when reduced to the scale with step equal to 15°. As a result, this coordination figure will receive the designation A-2{75}. Six different angles characterize the tetrahedron in the general case; therefore, we obtain of calculations a sequence of angles (88.8, 104.7, 104.7, 110.9, 110.9, 129.5) corresponding to three different values after bringing them to the scale with the step. If such a set is present in the list of reference figures, then the coordination figure of the atom of titanium will receive the designation T–4{90,105^4,135}. Otherwise, the comparison algorithm will select the coordination figure closest to it in terms of distance (*DRank*). In our case, such a figure is absent among the reference figures, so T–4{105^3,120^3} and *DRank* = 4.1 will be chosen. The T–4{90,105^2,120^3}

coordination figure is 5.7 units removed from the original and will be selected if there are no T–4{105^3,120^3} in the library. The list of reference figures was based on cluster analysis of the six-dimensional space of their angles using the Python software environment [11].

Information on compounds was selected from the Cambridge Structure Database (CSD, 2019) [12], and Inorganic Crystal Structures Database (ICSD, release 2018/2) [13]. The filtration of structures, the isolation of SBUs, and the calculation of their descriptors were performed by complex of structurally topological programs ToposPro [14]. The underlying net topology was calculated for the structures in standard representation of valence-bonded MOFs. The TTD collection [15] was used in its calculation. A collection of topological types of ligands (TTL-collection) was created by the ADS module [14]. Topologically and chemically different ligands, their occurrence, as well as geometric-topological descriptors, were saved in the collection. All the considered structures were a coordination compounds of only ligands $C_nH_mO_k$, where n≥0, m≥0, k>0, then the molecular particles H_2O, H_3O^+, OH^-, O^{2-} or O_2^{2-} were admitted in them. The coordination of all ligands to the metal atom only by a σ-bonded oxygen atom was another important selection condition. The above conditions were satisfied by 7690 structural studies, 6171 crystal structures, and 1279 ligands, differing in chemical composition or molecular graph.

4-c Nodes in the underlying nets are found for 2077 ligands and 2620 metal atoms. Four coordination figures are possible for coordination with CN=4 [10]: tetrahedron (T–4), a flat square (SP-4), a square pyramid (SPY-4) and a swing-shaped shape (SS-4). The angular space will be six-dimensional, which significantly complicates the visual representation of the distribution. To perform the analysis, we reduced the dimension by combining the values in pairs (α_1,α_2), (α_3,α_4), (α_5,α_6) and taking their average α_{min}, α_{mid} and α_{max}, respectively (Fig. 3).

The resulting set of coordination figures can be divided into four groups: tetrahedral T–4{α_{min}^2,α_{max}^4}, rectangular RAP–4{α_{min}^2,(180-α_{min})^2,180^2}, fanlike FAN–4{α_{min}^3,2α_{min}^2,3α_{min}}, and square pyramidal SPY–4{α_{min}^4,α_{max}^2}. If we take into account both idealized and distorted variants of coordination figures, then the RAP–4 (43.5%), FAN–4 (17.9%), SPY–4 (16.6%) for ligands (CF_{Lig}) and the T–4 (43.6%), SS–4 (26.2%), RAP–4 (13.4%) for metal atoms (CF_{Me}) are in the top three. The predominance of the tetrahedral coordination figure for metals indicates their uniform environment, compared with ligands.

The proposed shape descriptor of a coordination figure can be used to search for possible options for assembling an underlying net from one or several coordination figures of a given shape. In particular, Table 1 contains topological types that can be collected from only four-coordinated nodes. The most likely topological type of **pts** is obtained from CF_{Me} T–4 and CF_{Lig} RAP–4 (Fig. 4). In more rare cases, combinations of FAN–4, SS–4 or RAP–4 with RAP–4 may provide the same topology. If the coordination figures of RAP–4 correspond to CF_{Me} and CF_{Lig}, then the **cds** topology will be obtained. If SPY–4 corresponds to CF_{Lig}, then two-dimensional nets prefer to be realized (Table 1).

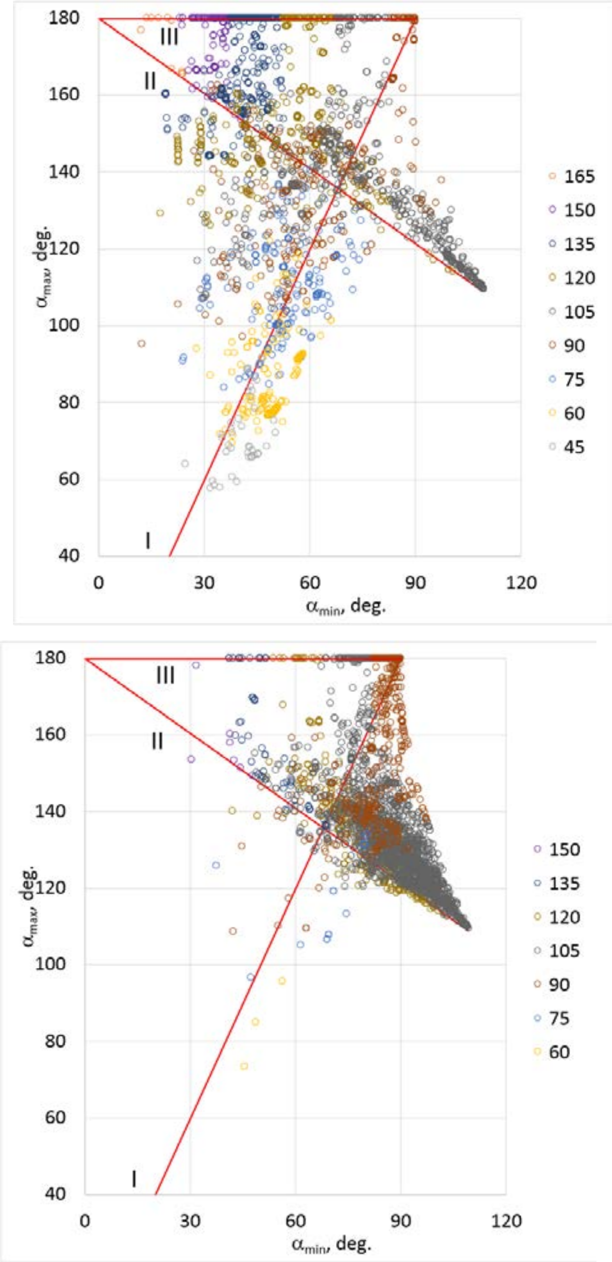


Fig. 3. Distribution of the minimal α_{min} and maximal α_{max} average angles for coordination figures with CN=4 for ligands (CF_{Lig}, left) and metal atoms (CF_{Me}, right). The average value α_{mid} of the intermediate angles α_3, α_4 of the coordination figure determines the color of the circles. Lines I, II and III correspond to different variants of the square pyramid SPY–4, tetrahedron T–4 and rectangular RAP–4, respectively

Table 1. Frequent pattern tree for associations "Dimension – Topology – $\mathrm{CF_{Lig}}$ – $\mathrm{CF_{Met}}$" are shown. Absolute (N_i) and relative ($w_i= N_i/N_{i\text{-}1}$) occurrences of N_0 = 126 coordination compounds with 4-c underlying net are given[a]

Dim	N_1	w_1	Topology	N_2	w_2	$\mathrm{CF_{Lig}}$	N_3	w_3	$\mathrm{CF_{Met}}$	N_4	w_4
3D	72	0.571	**pts**	25	0.347	RAP–4	22	0.880	T–4	10	0.455
									FAN–4	6	0.273
									SS–4	3	0.136
									RAP–4	2	0.091
			sra	7	0.097	RAP–4	3	0.429	T–4	2	0.667
			cds	6	0.083	RAP–4	6	1.000	RAP–4	6	1.000
			crb	5	0.069	T–4	3	0.600	T–4	3	1.000
			4,4,4 T9	4	0.056	RAP–4	4	1.000	SPY–4	3	0.750
			6/4/t7-4-I422	4	0.056	FAN–4	4	1.000	SS–4	3	0.750
			dia	4	0.056	SS–4	2	0.500	SS–4	2	1.000
						T–4	2	0.500	T–4	2	1.000
2D	54	0.429	**sql**	32	0.563	SPY–4	19	0.594	SS–4	7	0.368
									FAN–4	5	0.263
									T–4	5	0.263
						FAN–4	10	0.313	SS–4	3	0.300
									RAP–4	3	0.300
									FAN–4	2	0.200
									TPY–4	2	0.200
			(6,3)Ia	9	0.167	SPY–4	5	0.556	T–4	3	0.600
									SS–4	2	0.400
			4L4	5	0.093	FAN–4	3	0.600	T–4	3	1.000

[a] Only the correlations with $w_i > 0.035$ and $R = w_1 \cdot w_2 \cdot w_3 \cdot w_4 > 0.015$ are given. Underlying nets with these topologies are shown in Fig. 4.

In this case, SS–4, FAN–4 or T–4 can be coordination figure of the metal. It is important to note that the obtained data on fingerprints of coordination figures can be used as material for machine learning in order to more accurately predict the topological type. We plan to implement this in further research. As our next step, underlying nets with a different type of node coordination will be studied for the coordination compounds that are selected in this study.

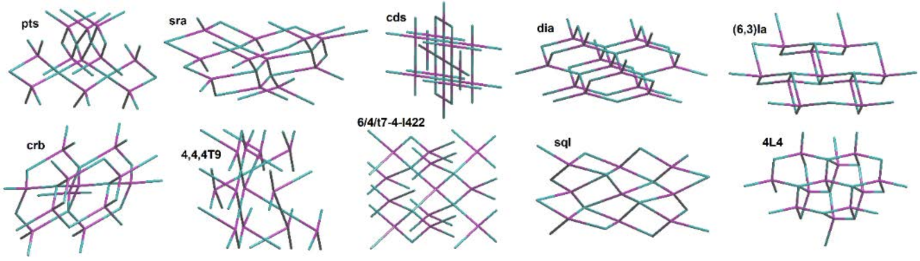


Fig. 4. The most abundant topologies in the studied structures are shown for the four coordinated underlying nets (see Table 1). Each net contains three coordination spheres of a metal atom (lilac nodes). The nodes that correspond to the ligand are colored turquoise

3 Conclusions

A new method for estimating the shape of a coordination figure using angular fingerprints has been proposed for coordination compounds with oxo-ligands. We have shown by the example of 4-coordination figures that a model based on this method allows us to designate any coordination figure, form a collection of reference coordination figures and automatically recognize an arbitrary set of coordination figures for this collection. The information obtained of the calculations can be used to classify the local environment of metals and ligands, as well as to predict the topological type of the crystal structure that is assembled from a given set of coordination figures.

Acknowledgements

The author thanks Russian Foundation for Basic Research (grants No. 18-07-00183 and 17-57-10001) for financial support.

References

1. Jain, A., Ong, S.P., Hautier, G., Chen, W., Richards, W.D., Dacek, S., Cholia, S., Gunter, D., Skinner, D., Ceder, G., and Persson, K.A.: A materials genome approach to accelerating materials innovation. The Materials Project. APL Mater. 1, 011002 (2013). doi: 10.1063/1.4812323
2. Ockwig, N.W., Delgado-Friedrichs, O., O’Keeffe, M., and Yaghi, O.M.: Reticular chemistry: occurrence and taxonomy of nets and grammar for the design of frameworks. Acc. Chem. Res. **38**, 176-182 (2005). doi: 10.1021/ar020022l
3. O’Keeffe, M., Peskov, M.A., Ramsden, S.J., and Yaghi, O.M.: The reticular chemistry structure resource (RCSR) database of, and symbols for, crystal nets. Acc. Chem. Res. **41** (12), 1782–1789 (2008). Reticular Chemistry Structure Resource. http://rcsr.anu.edu.au.
4. Öhrström, L.: Designing, describing and disseminating new materials by using the network topology approach. Chem. Eur. J. **22**, 1–7 (2016).

5. Wang, D., Yu, R., Kumada, N., and Kinomura, N.: Hydrothermal synthesis and characterization of a novel one-dimensional titanium glycolate complex single crystal: $Ti(OCH_2CH_2O)_2$. Chem. Mater. **11**, 2008–2012 (1999). doi: 10.1021/cm980579o
6. Alexandrov, E.V., Blatov, V.A., Kochetkov, A.V., and Proserpio, D.M.: Underlying nets in three-periodic coordination polymers: topology, taxonomy and prediction from a computer-aided analysis of the Cambridge Structural Database. Cryst. Eng. Comm, **13**, 3947−3958 (2011).
7. Blatov, V.A. and Proserpio, D.M.: Periodic-graph approaches in crystal structure prediction. Modern methods of crystal structure prediction, A.R. Oganov (ed), John Wiley & Sons, Ltd, Weinheim, 1–28 (2011).
8. Shevchenko, A.P., Blatov, I.A., Kitaeva, E.V., and Blatov, V.A.: Local coordination versus overall topology in crystal structures: deriving knowledge from crystallographic databases. Cryst. Growth Des. **17**, 774–785 (2017).
9. Lyakhov, A.O., Oganov, A.R., and Valle, M.: How to predict very large and complex crystal structures. Comp. Phys. Comm. **181**, 1623–1632 (2010). doi: 10.1016/j.cpc.2010.06.007
10. Hartshorn, R.M., Hey-Hawkins, E., Kalio, R., and Leigh, G.J.: Representation of configuration in coordination polyhedra and the extension of current methodology to coordination numbers greater than six (IUPAC Technical Report). Pure Appl. Chem. **79** (10), 1779–1799 (2007).
11. Python: A dynamic, open source programming language. Python Software Foundation. https://www.python.org
12. Groom, C. R., Bruno, I. J., Lightfoot, M.P., and Ward, S.C.: The Cambridge Structural Database. Acta Crystallogr., Sect. B: Struct. Sci., Cryst. Eng. Mater. **72**, 171−179 (2016).
13. Hellenbrandt, M.: The inorganic crystal structure database (ICSD) − present and future. Crystallogr. Rev. **10**, 17−22 (2004).
14. Blatov, V.A., Shevchenko, A.P., and Proserpio, D.M.: Applied topological analysis of crystal structures with the program package ToposPro. Cryst. Growth Des. **14**, 3576 (2014). doi: 10.1021/cg500498k
15. Alexandrov, E.V., Shevchenko, A.P., and Blatov, V.A.: Topological databases: why do we need them for design of coordination polymers? Cryst. Growth Des. **19**, 2604−2614 (2019). doi: 10.1021/acs.cgd.8b01721

Accelerated Materials Development Enabled by Collaborative Materials Information Management and Analytical Tools

Tatiana V. Vakhitova

Granta Design Ltd at Ansys, 300 Rustat House, CB1 7EG, Cambridge, UK

tatiana.vakhitova@grantadesign.com

Integrated computational materials engineering (ICME) is the full integration of processing-microstructure-property-performance relationships at various length scales, wherein the linkages from processing all the way up to performance can be made and utilized [1]. To realise the full potential of ICME, it is essential to integrate both computationally based and experimental data over the entire materials data life cycle and at various lengths in the same materials information management system.

More accurate modelling will require more data to be measured, predicted, compared, stored and fully tracked. As a provider of the effective enterprise level materials information management system, Granta Design has been involved in many collaborative projects on computational materials science, composites, additive manufacturing, advanced metallurgy etc. Collaborative ICME projects between academics and industries are good examples of accelerating materials development and creating values through sharing information.

In this contribution, we will share our views on how to successfully collaborate on martials knowledge, especially for projects with involvement of industrial end users. We will also give an example of the software for materials data analysis and visualization to enable materials-related decisions and communication of the results in a compelling manner.

References

1. Arnold, S.M., Bednarcyk, B.A., and Austin, N.: *NASA Technical report,* https://ntrs.nasa.gov/search.jsp?R=20160010071

Molecular Model Database of the Boltzmann–Zuse Society for Computational Molecular Engineering

Simon Stephan,[1] Martin Thomas Horsch,[2] Jadran Vrabec,[3] and Hans Hasse[1]

[1] Technische Universität Kaiserslautern, Laboratory of Engineering Thermodynamics, Erwin-Schrödinger-Str. 44, 67663 Kaiserslautern, Germany

[2] UK Research and Innovation, STFC Daresbury Laboratory, Keckwick Ln, Daresbury, Cheshire WA4 4AD, United Kingdom

[3] Technische Universität Berlin, Thermodynamics and Process Engineering, Ernst-Reuter-Platz 1, 10587 Berlin, Germany

simon.stephan@mv.uni-kl.de, martin.horsch@stfc.ac.uk

The openly accessible molecular model database (MolMod DB) of the Boltzmann–Zuse Society for Computational Molecular Engineering contains materials relations (force fields) for over 150 low-molecular fluids, meant for molecular modelling and simulation with molecular dynamics (MD) and Monte Carlo (MC) solvers [1].

The molecular models in the database have been published in about 30 articles over the past 20 years, which are associated with the respective entries and can be followed on the MolMod DB web front end. The database provides a wide range of search functionalities, e.g., for substances (names and CAS numbers) and model classes. Input files for several common environments can be downloaded via the web front end, including the file formats used by the molecular simulation codes ms2 [2], GROMACS [3], LAMMPS [4], and ls1 mardyn [5].

Acknowledgements

The co-authors from Berlin and Kaiserslautern acknowledge funding from the German Federal Ministry for Education and Research (BMBF) under grant no. 01IH16008, *Task-basierte Lastverteilung und Auto-Tuning in der Partikelsimulation (TaLPas),* the co-author from Daresbury acknowledges funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement no. 760907, *Virtual Materials Marketplace (VIMMP),* and the co-authors from Kaiserslautern acknowledge funding from the German Research Foundation (DFG) within the Reinhart Koselleck Programme and from the European Union’s Horizon 2020 research and innovation programme under grant agreement no. 694807, *Enrichment of Components at Interfaces and Mass Transfer in Fluid Separation Technologies (ENRICO).*

References

2. Stephan, S., Horsch, M., Vrabec, J., and Hasse, H.: MolMod – an open access database of force fields for molecular simulations of fluids. Molecular Simulation **45** (10), 806–814 (2019).
3. Rutkai, G., Köster, A., Guevara Carrión, G., Janzen, T., Schappals, M., Glass, C. W., Bernreuther, M., Wafai, A., Stephan, S., Kohns, M., Reiser, S., Deublein, S., Horsch, M., Hasse, H., and Vrabec, J.: ms2 – a molecular simulation tool for thermodynamic properties, release 3.0. Computer Physics Communications **221**, 343–351 (2017).

4. Abraham, M., Murtola, T., Schulz, R., Szilárd, P., Smith, J., Hess, B., and Lindahl, E.: GROMACS – High performance molecular simulations through multi-level parallelism from laptops to supercomputers. SoftwareX 1–2, 19–25 (2015).

5. Plimpton, S.: Fast parallel algorithms for short-range molecular dynamics. Journal of Computational Physics **117** (1), 1–19 (1995).

6. Niethammer, C., Becker, S., Bernreuther, M., Buchholz, M., Eckhardt, W., Heinecke, A., Werth, S., Bungartz, H.-J., Glass, C. W., Hasse, H., Vrabec, J., and Horsch, M.: ls1 mardyn – The massively parallel molecular dynmics code for large systems. Journal of Chemical Theory and Computation **10** (10), 4455–4464 (2014).

Databases on the Properties of Substances and Computer-assisted Design of Inorganic Compounds

N.N. Kiselyova[1][0000-0001-7243-9096], V.A. Dudarev[1][0000-0002-3583-0704],
and A.V. Stolyarenko[1]

[1] A.A. Baikov Institute of Metallurgy and Materials Science of RAS (IMET RAS), Moscow, 119334, Russia
kis@imet.ac.ru

Abstract. The virtually integrated distributed system of databases on the properties of inorganic substances and materials of the A.A. Baikov Institute of Metallurgy and Materials Science, Russian Academy of Sciences is considered. The information-analytical system for automation of process of new inorganic compounds computer-assisted design based on machine learning methods usage for search for regularities in information of the databases on inorganic substances and materials properties is discussed. The results this system application for compound design that have not yet been synthesized are presented.

Keywords: Database, Inorganic Substance and Material, Machine Learning

1 Introduction

Modern information technologies have made it possible to systematize and make available a huge array of data accumulated by chemistry over the centuries. Chemists and materials scientists make extensive use of the rich capabilities provided by numerous databases (DB), including the database on the properties of inorganic substances and materials (DBs PISM) [1], containing not only publications [2–5], but also data on the properties of substances [1, 6–11]. More detailed information on the information resources of inorganic chemistry is given in the IRIC database developed by us [12].

Information service does not limit the capabilities of the developed databases. One of the ways to make rational use of information on substances is the search for regularities that connect the properties of substances with the parameters of components. The objective existence of such regularities is a consequence of the Periodic Law. However, numerous attempts to present the desired complex regularities in an analytical form, as a rule, were unsuccessful, especially in the case of multicomponent substances. The methods for finding such complicated regularities in the data, based on the ideas of machine learning, were developed. In the mid-sixties, the idea of using machine learning to find regularities, that relate the properties of inorganic compounds to the parameters of components, was first proposed in our Institute of Metallurgy and Materials Science (IMET) [13]. Already the first calculations allowed us to find the relationship between the properties of binary compounds and the parameters of chemical elements, as well as to use the found regularities to predict compounds not yet obtained with an accuracy of about 90% [14]. Our further research in this area was associated with the use of more advanced machine learning programs [15–17] and complication of the composition of the compounds being predicted [18, 19].

1 Integrated Database System of IMET RAS on the Properties of Inorganic Substances and Materials

The source of information for the use of machine learning methods is the DBs PISM. In contrast to the databases usually used for these purposes, the information systems developed by the Institute of Metallurgy and Materials Science, Russian Academy of Sciences [1, 6, 11], by their functional structure, are focused on the selection of information for machine learning, which significantly reduces the time for preparation and analyzing the necessary data.

One of the most important problems in the application of machine learning to inorganic chemistry is the inconsistency of data obtained by different researchers. In this regard, the selection of information for machine learning is carried-ot by qualified experts in this subject area. This procedure is facilitated by providing the experts with the full texts of publications contained in our DBs PISM, from which examples are selected for machine learning, as well as through special programs for detecting sharply distinguished objects (outliers).

Now the integrated system of the DBs PISM includes the information systems developed in the IMET [1, 6, 11]: on the phase diagrams of semiconductor systems (Diagram), the properties of the acoustooptical, electro-optical, and nonlinear optical substances (Crystal), the band gap of inorganic substances (Bandgap), the properties of inorganic compounds (Phases), and the properties of chemical elements (Elements), the AtomWork database on the properties of inorganic substances, developed at the National Institute for Materials Science (NIMS, Japan) [8], and the TKV on substances thermal constants, developed in the Joint Institute for High Temperatures of RAS and Lomonosov Moscow State University cooperation.

The Phases database on the properties of inorganic compounds currently contains information on the properties of approximately 54000 ternary compounds and more than 34000 quaternary compounds, collected using more than 36000 publications. It includes brief information about the most common properties of inorganic compounds: crystal chemical (the type of crystal structure with indication of the temperature and pressure above which this structure is implemented, the crystal system, the space group, the number of formula units in the unit cell, and the lattice parameters) and thermophysical (the melting type and temperature, the temperature of decomposition of the compound in solid or gaseous phases, and the boiling point at atmospheric pressure) data. In addition, the database contains information on the superconducting properties of compounds. This database is available on the Internet for registered users [11].

The Elements database includes information about 90 of the most common properties of chemical elements: the thermal (the melting and boiling points at 1 atm and the standard values of thermal conductivity, molar heat capacity, enthalpy of atomization, entropy, etc.), size (the ionic, covalent, metal, and pseudopotential radii, the atomic volume, etc.), and other physical properties (the magnetic susceptibility, electrical conductivity, hardness, density, etc.); etc. The database is available on the Internet [11].

The Diagram database contains data on phase P,T,x-diagrams of binary and ternary semiconductor systems and the physicochemical properties of phases formed in them, collected and evaluated by highly qualified experts. The Diagram database is available on the Internet for registered users [11].

The Bandgap database includes information about the band gap of more than 3600 inorganic substances and is available on the Internet [11]. It has English version only.

The Crystal database includes information about the piezoelectric (piezoelectric coefficients, elastic constants, etc.), nonlinear optical (nonlinear optical coefficients, the Miller tensor components, etc.), crystal chemical (the type of the crystal structure, crystal system, space and point groups, the number of formula units per unit cell, and the crystal lattice parameters), optical (refractive indices, the transparency band, etc.), and thermal (melting point, specific heat, thermal conductivity, etc.) properties of more than 140 acousto-optical, electro-optical, and nonlinear optical materials, collected and evaluated by highly qualified experts in the subject area. It has Russian and English versions available for registered users on the Internet [11].

The AtomWork Inorganic Material Database (NIMS, Japan) contains information about more than 82000 crystal structures, 55000 values of the properties of materials, and 15000 phase diagrams; it is also available on the Internet [8].

The TKV DB on substances thermal constants contains information, available online from the Internet, on about 27 thousand substances formed by all chemical elements.

The complex integration approach that combines integration at data and user interfaces level is applied to these database integration [20]. The special single entry point allows a search for the all data on certain substance from different DBs.

2 Inorganic Compounds Computer-Assisted Design System

Machine learning procedure involves several stages:

1. The objects selection for machine learning.
2. The attribute description formation (including the most informative attributes selection and filling attribute values gaps also).
3. The best ML algorithms selection.
4. Machine learning including application of algorithms ensembles and collective solution synthesis in a case of several algorithms usage.
5. ML quality estimation.
6. New objects prediction and results interpretation.

The special information-analytical system (IAS), which, in addition to the information service for professionals, is designed to search for regularities in big chemical data and computer design of inorganic compounds was developed in IMET [21]. It includes (Fig. 1), along with the integrated system of DBs PISM, a subsystem of information analysis and predictions, bringing together a set of programs of machine learning, a base of found regularities (the knowledge base), a base of predictions of the possibility of forming and properties of inorganic compounds that have not been yet synthesized, and a management subsystem.

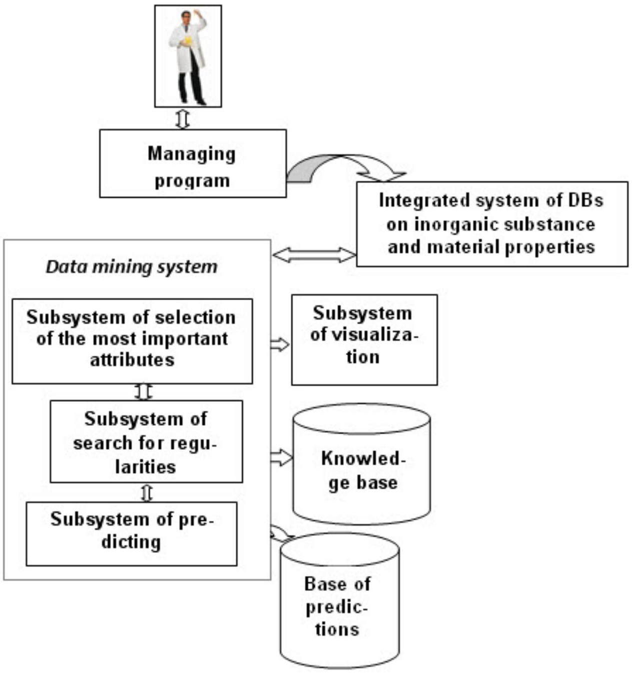

Managing program
Integrated system of DBs on inorganic substance and material properties
Data mining system
Subsystem of selection of the most important attributes
Subsystem of visualization
Subsystem of search for regularities
Knowledge base
Subsystem of predicting
Base of predictions

Fig. 1. The information-analytical system structure for inorganic compounds design

2.1 Subsystems for Searching for Classifying Regularities and Predictions

In the development of this subsystem, the most important task was the selection of the most appropriate mathematical methods for searching regularities in chemical data. Typically, this task is performed by the trial-and-error method. In the selection of machine learning methods for analysis of chemical information, many years' experience in the application of these methods to inorganic compounds design was taken into account [18]. The following methods and programs have been selected:

- a wide range of algorithms of the Recognition multifunctional system, developed at the Computing Center of the Russian Academy of Sciences [16] and bringing together, in addition to well-known techniques, the algorithms of pattern-recognition (based on calculation of estimates), voting algorithms based on deadlock tests, voting algorithms based on logical regularities, weighted statistical voting algorithms, etc.;
- a ConFor computer system for training a computer in the procedure for concept formation [15], which is based on an original organization of data in the computer memory in the form of growing pyramidal networks.

As a rule, it is not possible to specify in advance which algorithm would be the most efficient for solving a particular problem. Therefore, it seems promising to apply the methods of prediction by algorithms ensembles. In a collective decision creation, the possible prediction errors of individual algorithms can be compensated in many cases by correct results of other algorithms. Based on this, we included programs that implement different strategies for collective decision-making, for example, the Bayesian method, methods using clustering and selection, decision templates, logical correction, the method of a convex stabilizer, the Woods dynamic method, committee methods, etc., [16] in the developed IAS [21].

2.2 Subsystem for Searching the Classifying Properties of Components

For the selection of informative properties of the chemical compound's components, we included programs based on algorithms [22–24] in the IAS. The selection of the properties of the components, the most informative for the classification of substances, has a double meaning. On the one hand, it drastically decreases the volume of the information analyzed, which for multicomponent substances comprise hundreds of properties of elements and simpler compounds, as well as functions of these properties. On the other hand, the selection of properties of the components most important for the classification of chemical substances, enables the physical interpretation of the classifying regularities, which enhances the credibility of the predictions obtained and finding substantial causal links between the parameters of the objects and the development of the physical and chemical models of phenomena.

2.3 Visualization Subsystem

This subsystem facilitates the results interpretation, which constructs the projections of the points corresponding to the compounds in two-dimensional spaces of the properties of components, including not only the initial parameters but also user-specified algebraic functions of these parameters.

2.4 Knowledge and Prediction Bases

The knowledge base contains the obtained classifying regularities. The prediction base contains the results of previous computer experiments, as well as links to operation information stored in the knowledge base. Using the prediction base helped to improve the functionality of the databases on the properties of inorganic substances and materials, developed at the IMET, by providing the user with not only known data about already studied substances but also predictions for inorganic compounds not yet synthesized and evaluations of their properties.

2.5 Management Subsystem

The management subsystem organizes the computing process, ensures interaction between the functional subsystems of the IAS, and provides access to the system on the Internet. In addition, the management subsystem provides the expert with software for data preparation for analysis, outputting reports, and implementation of other service functions. In particular, we developed a special subsystem to retrieve information from the database, which, after evaluation of the expert, is used to learn the computer, and to prepare it for further analysis. It gives the expert the capability to edit the found information and to form training samples for analysis. In the latter case, the expert marks only the selected properties of the components in a special table (menu), and the subsystem for the sample preparation for analysis retrieves the selected property values from the Elements database. If needed, the algebraic functions of the initial properties are formed in the subsystem and the description of the compounds is assembled in the form of an Excel table, which is then input to the prediction subsystem. The subsystem of result delivery is intended to make predictions in a tabular form conventional to chemists and materials scientists.

3 Use the IAS for Predicting New Compounds and Evaluation of their Properties

The machine learning application allowed a search for inorganic compounds formation regularities, a prediction of thousands not yet synthesized substances and some their properties evaluation using obtained regularities. This approach efficiency to inorganic compounds design can be illustrated by comparison of the predictions results with newer experimental data obtained after publication of our predictions.

3.1 Prediction of the TiNiSi Crystal Structure Type for Compounds with the Composition ABAl

The equiatomic aluminides are of interest for the search for new magnetic materials. Thirty years ago, the prediction of new compounds of this type was carried out by us [25]. The algorithm based on the growing pyramidal networks learning (GPNL) [15] was used in the search for the criteria of this crystal structure type formation at ambient conditions. The learning set contained 39 examples of the compounds ABAl (hereinafter, A and B are various chemical elements) with the TiNiSi crystal structure type and 57 examples of the compounds with the structures different from TiNiSi. The following properties of elements A and B (attributes) were chosen for description of intermetallics: the distribution of electrons in the energy levels of isolated atoms of the chemical elements, the first three ionization potentials, the metal radii by Bokii and Belov, the standard entropies of individual substances, the melting points, the number of complete electronic shells, the number of electrons in incomplete s-, p-, d- or f-electronic shells for the atoms of elements.

Table 1. Part of a table illustrating the prediction of the crystal structure type TiNiSi for compounds with the composition ABAl [25]

A B	La	Ce	Pr	Nd	Pm	Sm	Eu	Gd	Tb	Dy	Ho	Er	Tm	Yb	Lu
Ru	-	O	-	-	-	+	-	+	+	+	+	+	+	-	-
Rh	O	O	∅	∅	-	+	-	©	+	+	©	+	©	∅	-
Os	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
Ir	©	©	©	©	+	©	+	©	©	©	©	©	©	+	©
Pt	©	©	©	©	+	⊕	+	⊕	⊕	⊕	⊕	⊕	⊕	©	

Table 1 is a result of comparing the predictions for each sets of properties and contains the comparison results of our predictions with newer experimental data. The following notations are used: + – prediction of the TiNiSi crystal structure type; - – prediction of the absence of the TiNiSi crystal structure type; ⊕ – a compound with composition ABAl has the TiNiSi crystal structure type and this fact was used for machine learning; © – prediction of the TiNiSi crystal structure type was confirmed experimentally; O – prediction of the crystal structure type different from TiNiSi was confirmed experimentally; ∅ – prediction of the crystal structure type different from TiNiSi was not confirmed experimentally; here and in other Tables the blank spaces correspond to the disagreement of the predictions with the use of different sets of the component's properties; all data and predictions are given for the substances under ambient conditions. A comparison of our predictions with newer experimental data has shown that the prediction error is lower than 12%.

3.2 Design of Compounds with Composition ABX_2 (X – S, Se, or Te)

The chalcogenides with composition ABX_2 are a class of compounds that is promising for the search for new semiconducting and nonlinear optical materials. Taking into account the perspective of these compounds practical use the design of their not yet synthesized analogues was made [26]. Previously we predicted new compounds of this composition [27] also.

The task solution was subdivided into to stages: (1) prediction of the formation of compounds with composition ABX_2; and (2) prediction of the crystal structure type of these compounds under ambient conditions.

Prediction of the formation of compounds with composition ABX_2. Data on 667 examples of the formation of ABX_2 (X=S, Se, or Te) and 504 examples of the absence of this composition compounds in the systems A_2X–B_2X_3 and AX–BX under ambient conditions were used for machine learning. The data were taken from the DB Phases. 84 properties of the elements A, B, and X, whose values were taken from the DB Elements, were used for the compounds representation in the computer memory.

For the data analysis, several machine learning algorithms that are included into IAS were used. The learning quality was estimated on the basis of examination recognition in the mode of cross-validation. The analysis of the results using various algorithms has shown that the best predictions under cross-validation have been obtained for the decision tree method (DT) [16] (accuracy of prediction being 72%), the logical regularities voting algorithm (LoReg) [16] (accuracy of prediction being 67.3%), and the deadlock test algorithm [16] (accuracy of prediction being 67.6%). These algorithms have been used for collective decisions using the committee method, in which the resulting prediction is calculated as an average arithmetic value of predictions obtained using different algorithms [16]. Using this procedure, the compound's formation predictions were obtained.

Prediction of the crystal structure type of compounds ABX_2. Data on 158 examples of the formation of ABX_2 with the crystal structure under ambient conditions α-$NaFeO_2$, 44 compounds with NaCl structure, 47 compounds with chalcopyrite structure, and 24 compounds with TlSe structure were used for machine learning. The same properties were used for the compound's representation.

The problem was solved in two ways. In the first case, multi-class learning and prediction, where the cumulative information on the four above-mentioned crystal phases has been used, was applied. In the second case, four problems of the dichotomy were solved – division into two classes, e.g., class 1, compounds with chalcopyrite crystal structure, and class 2, compounds with another structure. The results of predictions were compared, and a decision was made if the predictions obtained by multi-class prediction and dichotomies did not contradict each other. The results are summarized in Table 2.

Table 2. Part of a table illustrating the prediction of the crystal structure type for compounds with the composition ABX_2 (X – S, Se, or Te) [26]

X	S								Se								Te							
A B	Li	Na	K	Cu	Rb	Ag	Cs	Tl	Li	Na	K	Cu	Rb	Ag	Cs	Tl	Li	Na	K	Cu	Rb	Ag	Cs	Tl
B	#5	#5	#5	©3	#5	3	#5	#5	1	1		#5	1	3	1	4	1	1	4		1		1	4
Al	#5			#3		#3		©4	#5	#4	#5	#3		#3		#5		#4	#4	#3		#3	4	#4
Sc	#1	#1	©1	#5	©1			#1	1	1	1		1	#5	1		1		1		1		1	
Cr	#5	#1	#1	#5	#1	#5		#5	$2	#1	©1	#5	#1	#5	5				5	3	5	#5	5	#5
Fe	#5		#5	#3	#5	#3		#5		1	#5	#3	#5	#3		#5			1	#3	1	#3	1	©5
Ga	#5	#5	#5	#3	#5	#3	©5	#5	©5	4	#5	#3	4	#3		#5	#3	#4	#5	#3		#3		#4
As		#5		#5		#5	5	#5		#5	#5	#5	#5	#2	#5	©5	2		4	#6	$1	#2	5	
Y	#2	#1	#1	©5	©1	#5	5	#1	#1	#1	1		1	#5	1	#1	1	1		#5	1	#5	1	#1
In	#5	#1	#5	#3	#5	#3	#5	#4	#5	#1	©5	#3	#5	#3		#4	#3	#4	#4	#3		#3	$4	#4
Sb		©5	#5	#5	#5	$2	©5	©2	©2	©5	#5	#5	#5	#2		©5	#2	#2	4	#5	#5	#2	#5	#1
La		#2	#1		#1		#1	#6	#5	#1	©1	#5	#1		1	#6	2		©1		1		1	6
Ce		#2	#1	#5	#1		#1	#6	#5	#1	1	#5	#1	#6	1	#6			©1		©1		1	#6
Pr	#2	#1	#1	#5	#1		#5		1	#1	1	#5	#1	#6	1	#1	2		©1	#5	1		1	#1
Nd	#2	#1	#1	#5	#1		#5			#1	1		#1		1	#1	1	1	©1	#5	©1		©1	#1
Pm		1	1	5	1	5		1	1	1	1		1		1	1	1	1			1		1	
Sm	#2	#1	#1	#5	#1	#5	#5	#1		#1	©1	#5	#1	#6	1	©1		©1	©1		©1		1	#1
Gd	#2	#1	#1	#5	#1	#5	#5	#1	#1	#1	1	#5	#1	#5	1	#1	1	1	©1		1	#5	1	#1
Tb	#2	#1	#1	#5	#1	#5	#5	#1	#1	#1	1	#5	#1		1	#1	1	1			1	©5	1	#1
Ho	#1	#1	#1	#5	#1	#5	#5	#1	#1	#1	1		#1	#5	1	#1	1	1			1	#5	1	#1
Er	#1	#1	#1	#5	#1	©5	#5	#1	#1	#1	1		#1	#5	1	#1	1	1	#1		1	#5	1	#1
Tm	#1	#1	#1	#5	#1	#5	#5	#1	1	1	1		1	#5	1	#1	1	1			1	#5	1	#1
Yb	#1	#1	#1	#5	#1	#5	#5	#1		#1	#1		1	#5	#5	©1	2	1	1	#5	1		1	
Lu	#1	#1	#1	#5	#1	#2	#5	#1	1	1	1		#1	#5	1	#1	1	1			1	#5	1	#1
Bi	#2	#2	#2	#5	#1	#1		#1	#2	#2	#2	#2	#5		#5	#2	#2	#2		#5		$1		#1

In Table 2, the following notations were used: 1 – prediction of the structure of the α-$NaFeO_2$ type; 2 – prediction of the structure of the NaCl type; 3 – prediction of the structure of the chalcopyrite type; 4 – prediction of the structure of the TlSe type; 5 – prediction of the structure different from the ones mentioned above; 6 – prediction of the absence of ABX_2; the symbol # is used for objects for the machine learning; © – predictions was confirmed experimentally; $ – predictions was not confirmed experimentally.

40 compositions have been experimentally tested and only in five cases the predictions turned out to be incorrect, i.e., the prediction error was about 12.5 %. Beyond that the melting point and bandgap were evaluated for compounds with the chalcopyrite crystal structure type [28].

From ternary to quaternary compounds. Prediction of the crystal structure type of compounds A_2BCHal_6. Searching for and studying halide compounds having the composition A_2BCHal_6 (Hal=F, Cl, Br, or I) with the elpasolite crystal structure type is related to the development of new luminescent, laser, and magnetic materials.

The set for computer-assisted analysis included information about 289 (A≠C) compounds having the elpasolite structure; 20 compounds with Cs_2NaCrF_6 type of crystal structure; 57 compounds with crystal structures another than the ones given above under ambient conditions; and 81 AHal–$BHal_3$–CHal systems where compounds are not formed [19]. The 134 properties of chemical elements A, B, C, and Hal were included in the initial set of component parameters.

The problem of predicting new halo-elpasolites included solving three intermediate tasks. Formation of compounds with composition A_2BCHal_6 was predicted in the first of them (task 1). The next task included searching for regularities and predicting the formation of compounds with given composition and the most common types of crystal structures (elpasolite or Cs_2NaCrF_6). The latter task was divided into two smaller ones. When solving the first of them, the multi-class prediction of belonging to four classes (elpasolites, compounds with the Cs_2NaCrF_6 structure, compounds with the structure different from those shown above, and the systems containing no compounds with composition A_2BCHal_6 (task 2)) was performed. Next, halide systems were consecutively divided into three classes: the target class, e.g., 1 – elpasolites; class 2 – compounds with non-elpasolite structure; and class 3 – the AHal–$BHal_3$–CHal systems containing no compounds with composition A_2BCHal_6 (task 3). The final decision regarding the class that a compound being predicted belongs to, was made by comparing the predictions obtained when solving all three tasks. If the results were inconsistent, the prediction was regarded to be uncertain and the prediction table cell was left empty.

The algorithms LoReg, artificial neural network learning (ANN), K-nearest neighbor (KNN), and support vector machine (SVM) ensure the best accuracy of prediction of compound formation (task 1) in the cross-validation mode and the collective decision-making software based on the algorithm of generalized polynomial corrector [16] provided the best estimate for prediction accuracy, namely 95%.

When solving task 2 of multi-class prediction, the set of algorithms including DT, KNN, SVM, ANN, learning a multilayer perceptron, and the algorithm of the convex stabilizer [16] for collective decision-making, ensured the best accuracy of examination prediction: 89%. When forming the regularity that allows one to demarcate elpasolites from compounds with differing crystal structures and from systems where no

A_2BCHal_6 compounds are formed (task 3), the best accuracy (80%) was provided by the set of algorithms that included the algorithms LoReg, ANN, KNN, SVM, and the Bayesian method of collective decision-making [16].

Some results of comparing the predictions found by solving all three classification tasks are summarized in Table 3. The following notations are used: 1 – prediction of compounds with the elpasolite crystal structure; 2 – prediction of compounds with the Cs_2NaCrF_6 structure type; 3 – prediction of compounds having crystal structure another than the abovementioned ones; and 4 – prediction of the absence of an compound in the ACl–BCl_3–CCl system; the # symbol is used to denote previously studied compounds; the information about them was used for machine learning.

Table 3. Part of a table illustrating the prediction of the crystal structure type for compounds with the composition A_2BCCl_6 [19]

C	Li					Na					K				Rb			
A B	Na	K	Rb	Cs	Tl	Li	K	Rb	Cs	Tl	Li	Na	Rb	Cs	Li	Na	K	Cs
Al	4	#4	4	4	1	4	#4	4	#4	1	#4	#4	4	4	4	4	4	4
Sc		#3	#1	#2	#1			#1	#1	#3		4	1	#1	4	4		#4
Ti		3	3	3			1	1	#1	1	4			1	4	4	3	1
V		3	#3	#3			#1	#1	1	1	4	4		1	4	4		1
Cr			3	#3			#1	#1		#1	4		#3	#1	4	#4	3	3
Fe	4	#4		3		4	#4		#1		#4	#4			4	4	4	#4
Y			#1	#1			#4	#3	#1	3	4	#4	1	#1	4	4	1	1
In			#1	#3	3		1	#1	#3	3			3	#3	4			3
La	#4	#4		#1		#4	4		#1		#4	4	1	1	4	4	1	1
Ce	4	#4	1	#1		4	4		#1		#4	4	1	1	4	4	1	1
Pr	#4	#4	#4	#1		#4	#4	#4	#1		#4	#4	1	#1	#4	#4	1	1
Nd	4	4	4	#1		4	#4	4	#1		4	#4	1	#1	4	4	1	1
Pm			3	1				3	1		4	4	1	1		4	1	1
Sm	4		#3	#1		4	#4		#1			#4	1	#1	4	4		1
Eu		3	#3	#1	3			#3	#1	3		4	1	#1		4	1	1
Gd		3	#3	#1	3			#3	#1	3	4		1	#1	4	4		1
Tb		3	#1	#1	3			3	#1	3			1	#1		4	1	1
Dy		3	#1	#1	3		#4	#3	#1	3		#4	1	#1		4	1	1
Ho		3	#1	#1	3			#3	#1	3			1	#1		4	1	1
Er		3	#1	#1	3			#3	#1	3			1	#1		4	1	1
Tm		#3	#1	#1	#3			#3	#1	#3			1	#1		4	1	1
Yb		3	#1	#1	#3		3	#3	#1	3			1	#1		4	1	1
Lu			#1	#3	#3		3	#3	#1	3			1	1	4	4	1	1
Tl		1	1				1	1	#1		4		1	1	4	4	1	1
U	#4	#4	4	#1		#4		#3	#1		#4	4		1	4	4		1
Pu	4	4		1		4	4	3	#1		4	4	1	1	4	4		1

4 Conclusions

During half of the century the predictions of thousands of inorganic compounds in binary, ternary and more complicated chemical systems were obtained and some their properties (melting point, critical temperature of superconductivity, band gap energy, etc.) were estimated in IMET. The obtained predictions usage allows an essential progress provision in a search for new magnetic, semiconductor, superconductor, nonlinear optical, electro-optical, acousto-optical and other materials. Hundreds of predicted compounds were synthesized and our results experimental verification shows that the average prediction accuracy is higher than 80%. Machine learning methods application to search for regularities in big chemical data of DB PISM gives an opportunity for theoretic design of new inorganic compounds that allows substantially reduce the costs for search for new materials with predefined properties, replacing them by computations. It is important to note that only information on components properties (chemical elements or more simple compounds) is used in prediction process.

Acknowledgements

This work was partially supported by the Russian Foundation for Basic Research (project Nos. 17-07-01362 and 18-07-00080) and the State task No. 075-00746-19-00. We are grateful to V.V. Ryazanov, O.V. Sen’ko, and A.A. Dokukin for long-term help and collaboration.

References

1. Kiselyova, N.N., Dudarev, V.A., and Zemskov, V.S.: Computer information resources in inorganic chemistry and materials science. Russ. Chem. Rev. **79** (2), 145–166 (2010).
2. ACS Publications Homepage, https://pubs.acs.org/, last accessed 2019/04/22
3. ScienceDirect Homepage, https://www.sciencedirect.com/, last accessed 2019/04/22
4. Springer Nature Homepage, https://link.springer.com/, last accessed 2019/04/22.
5. Wiley Online Library Homepage, https://onlinelibrary.wiley.com/, last accessed 2019/04/22
6. Kiselyova, N.N., Dudarev, V.A., and Stolyarenko, A.V.: Integrated system of databases on the properties of inorganic substances and materials. High Temperature **54** (2), 215–222 (2016).
7. NIST Data Gateway, https://www.nist.gov/srd, last accessed 2019/04/22
8. NIMS Materials Database (MatNavi) Homepage, http://mits.nims.go.jp/index_en.html, last accessed 2019/04/22
9. SpringerMaterials Homepage, https://materials.springer.com/, last accessed 2019/04/22
10. Blokhin, E. and Villars, P.: The PAULING FILE Project and Materials Platform for Data Science: From Big Data Toward Materials Genome. In: Andreoni, W., Yip, S. (eds.) Handbook of Materials Modeling, pp. 1–26. Springer, Heidelberg (2018).
11. IMET RAS DBs Homepage, http://www.imet-db.ru/, last accessed 2019/04/22
12. DB IRIC (Information Resources on Inorganic Chemistry) Homepage, http://iric.imet-db.ru/, last accessed 2019/04/22
13. Savitskii, E.M., Devingtal’, Yu.V., and Gribulya, V.B.: Prediction of metallic compounds with composition A_3B using computer. Doklady Physical Chemistry **183** (5), 1110–1112 (1968).

14. Savitskii, E.M. and Gribulya, V.B.: Application of computer techniques in the prediction of inorganic compounds. Oxonian Press Pvt., Ltd., New Delhi-Calcutta (1985).
15. Gladun, V.P.: Processes of formation of new knowledge. SD "Pedagog 6", Sofia (1995).
16. Zhuravlev, Yu.I., Ryazanov, V.V., and Sen'ko, O.V.: RECOGNITION. Mathematical methods. Software system. Practical solutions. Phasis, Moscow (2006).
17. Pedregosa, F., Varoquaux, G., Gramfort, A., et al.: Scikit-learn: Machine learning in Python. J. Machine Learning Research 12 (Oct.), 2825–2830 (2011).
18. Kiselyova, N.N.: Komp'yuternoe konstruirovanie neorganicheskikh soedinenii. Ispol'zovanie baz dannykh i metodov iskusstvennogo intellekta (Computer Design of Inorganic Compounds: Use of Databases and Artificial Intelligence Methods). Nauka, Moscow (2005).
19. Kiselyova, N.N., Stolyarenko, A.V., Ryazanov, V.V., et al.: Prediction of New Halo-Elpasolites. Russ. J. Inorg. Chem. **61** (5), 604–609 (2016).
20. Dudarev, V.A.: Information systems on inorganic chemistry and materials science integration. Krasand, Moscow (2016).
21. Kiselyova, N.N., Stolyarenko, A.V., Ryazanov, V.V., et al.: A system for computer-assisted design of inorganic compounds based on computer training. Pattern Recognition and Image Analysis **21** (1), 88–94 (2011).
22. Senko, O.V.: An optimal ensemble of predictors in convex correcting procedures. Pattern Recognition and Image Analysis **19** (3), 465–468 (2009).
23. Yuan, G.-X., Ho, C.-H., and Lin, C.-J.: An improved GLMNET for L1-regularized logistic regression. J. Machine Learning Research. 13, 1999–2030 (2012).
24. Yang, Y. and Zou, H.: A coordinate majorization descent algorithm for L1 penalized learning. J. Statistical Computation & Simulation **84** (1), 1–12 (2014).
25. Kiseleva, N.N. and Burkhanov, G.S.: Search for new ternary phases with Al, Ga, and In using an information-prediction system. Russian Metallurgy (1), 223–226 (1989).
26. Kiselyova, N.N., Podbel'skii, V.V., Ryazanov, V.V., and Stolyarenko, A.V.: Computer-aided design of new inorganic compounds with composition ABX_2 (X = S, Se or Te). Inorg. Mater.: Applied Researches **1** (1), 9–16 (2010).
27. Savitskii, E.M. and Kiseleva, N.N.: Cybernetic prediction of the existence of ABX_2 phases. Inorg. Mater. **15** (6), 866–868 (1979).
28. Kiselyova, N.N., Stolyarenko, A.V., Gu, T., et al.: Computer-aided design of new inorganic compounds promising for search for electronic materials. In: Proc. The Sixth Int. Conf. on Computer-Aided Design of Discrete Devices (CAD DD 07). vol. 1, pp. 236–242. UIPI NASB, Minsk (2007).

POSTERS, DEMONSTRATIONS AND NETWORKING

Equilibrium and Thermodynamic Properties of Tin

Aigerim Syzdykova[1], Maxim Belov[1], and Igor Abrikosov[1,2]

[1]Materials Modeling and Development Laboratory, National University of Science and Technology "MISIS", Leninskii Pr 4, Moscow 119049, Russia
[2] Department of Physics, Chemistry and Biology (IFM), Linköping University, SE-581 83 Linköping, Sweden

syzdykova.aygerym@gmail.com

Abstract. Mechanical and thermodynamic properties of α and β phases of tin are calculated within the density functional theory. Large value of calculated energy barrier between two phases explains the slow rate of β→α phase transition. The vibrations of atoms in the α phase are almost harmonious, and in the β phase there is a strong anharmonicity.

Keywords: Grey tin, white tin, lattice dynamic, DFT, VASP

1 Introduction

There are two general allotropic forms of tin. The low temperature stable phase is α-Sn or gray tin with a diamond cubic structure, which is a zero-gap semiconductor. Above T=286 K, α-Sn transforms into β-phase or white tin, which is a body-centered tetragonal metal [1]. Transition between alfa and beta phases is strongly complicated process, therefore experimental and thermodynamic data are poor.

2 Methods

In this work ab-initio simulations in the framework of density functional theory (DFT) were carried out to define the equilibrium parameters and heat capacities at a constant pressure (Cp) in a quasi-harmonic approximation for both phases of tin. The interaction between ions and electrons were described using the projector augmented-wave (PAW) method [2, 3]. The correlation and exchange in the electron gas were taken into account via the local density approximation (LDA), the generalized gradient approximation (GGA) and a new improved SCAN-functional. All calculations were performed in the Vienna ab-initio simulation package VASP.

3 Results

We calculated mechanical and thermodynamic properties for both tin phases: Gibbs free energy, equilibrium lattice parameters, bulk modulus, entropy and enthalpy, heat capacity at constant pressure from 0 K up to room temperatures. The potential barrier along the transformation path between two phases was calculated, which turned out to be very large, about 200 meV.

4 Conclusion

It is shown that the LDA functional better describes all the calculated properties of both phases. It is shown that the atomic vibrations in the alpha phase are almost harmonic, while in the beta phase there is a strong anharmonicity. Calculated heat capacity fill a gap in the experimental heat capacity values for both phases. It was found that the α↔β phase transition is associated with the high energy barrier.

References

1. Pasquale Pavone, Stefano Baroni, Stefano de Gironcoli. α-β phase transition in tin: A theoretical study based on density-functional perturbation theory. Phys. Rev. B **57** (17) (1998).
2. Kresse, G. and Joubert, D.: From ultrasoft pseudopotentials to the projector augmentedwave method, Phys. Rev. B **59** (3), 1758–1775 (1999).
3. Blöchl, P.E.: Projector augmented-wave method, Phys. Rev. B **50** (1994).

Solubility of Carbon in Paramagnetic Fe-based Alloys

A.V. Ponomareva[1], B.O. Mukhamedov[1], and I.A. Abrikosov[1,2]

[1] *Materials Modeling and Development Laboratory, NUST „MISIS", Moscow, Russia*
[2] *Physics Department, Linköping University, Linköping, Sweden*

Abstract. We present a generalization of a model that takes into account the magnetic disorder of paramagnetic host with interstitial point defects towards the case of the alloys. In the framework of disordered local moment picture combined with magnetic sampling method, we calculate a solution enthalpy of carbon impurity in the paramagnetic fcc Fe-Mn and Fe – Mn – Al steels. By analyzing the local and global effects of impurity on the properties of the matrix, we discuss various factors that could increase the carbon solubility in high-manganese austenitic steels.

Keywords: solution energy, disordered alloys, substitution and interstitial impurities, paramagnetic state, first-principle calculations

Within the framework of the electron density functional theory, implemented using the VASP software package, the solution energy of carbon in fcc paramagnetic Fe as well as disordered Fe–Mn and Fe–Mn–Al alloys was calculated. To describe the properties of Fe-Mn-X alloys, we used a model that takes into account the contribution of thermal magnetic fluctuations in a paramagnetic matrix with point defects [1], generalized to the case of a material with magnetic and chemical disorder. The values of the energy of carbon dissolution in Fe-based alloys, the magnitude of dilatations, exchange interactions, charge density difference maps were obtained the effect of the impurity on the magnetic properties of the matrix was analyzed. It is shown that in alloys containing ~ 20 at. % Mn, the solution energy of carbon reduces compared to the pure γ-Fe; the addition of ~ 2 at.% Al in Fe-Mn-C alloys leads to an increase in the solution energy of carbon relative to ternary Fe-Mn-C alloys, but the energy value is lower as when carbon dissolves in pure paramagnetic γ-Fe. In ternary alloys with Mn, due to an increase in the lattice parameter, the deformation of Fe – C bonds decreases, and at the same time increases the spatial size of the octahedral sites where the carbon impurity is dissolved that reduces carbon solution energy. In Fe-Mn-Al-C alloys, the resulting strong dilatations around the aluminum atom compensate for the positive effect of reducing the main distortions and increasing the spatial size of the pore. Due to the presence of large negative exchange interactions in ternary Fe-Mn-C alloys, the tendency to polarization of the first coordination sphere of the impurity, which was observed in the Fe-C system, disappears and the solution energy becomes independent of the magnetization of the first coordination sphere of impurity. Analysis of charge density maps shows the delocalization of electron density in the first coordination sphere of carbon, in the presence of manganese atoms. This makes the screening of the impurity more efficient which causes the reduction of the carbon solution energy. A strong repulsion was found between the C and Al atoms in the first coordination sphere, considering the orientation arrangement of the atoms as the nearest neighbors, the energy of carbon dissolution increases by 30%.

Acknowledgements

Financial support from the Russian Science Foundation (Project No. 18-12-00492) is gratefully acknowledged.

References

1. Ponomareva, A.V., Gornostyrev, Yu.N., and Abrikosov, I.A.: Ab initio calculation of the solution enthalpies of substitutional and interstitial impurities in paramagnetic fcc Fe, Phys. Rev. B 90, 014439 (2014).

Molecular Dynamics Simulation of the Glass Transition of a Supercooled Zr-Nb Melt

Sergey Klyavinek[1,2] and Lada Kolotova[1,2]

[1] JIHT RAS, Moscow 125412, Russia
[2] MIPT, Moscow 117303, Russia
klyavinekss@gmail.com

Abstract. The amorphous states of the Zr-Nb alloy were studied using molecular dynamics simulations. The conditions for obtaining metallic glass were studied (threshold cooling rates for different Nb contents), the amorphous state was analyzed using various criteria, and glass transition temperatures were determined.

Keywords: Zirconium, Niobium, glass transition, molecular dynamics

1 Prerequisites for Job Creation

Amorphous compounds today are widespread in practical use. They are used in various areas of industry – defense industry (fencing production) , home appliances and so forth. In this paper, Zr-Nb alloy is investigated as it is used for the production of fuel rod housings, as well as in implantology (it has good corrosion parameters and a suitable Young's modulus). For this reason, determining the conditions for obtaining amorphous Zr-Nb is practically an important task. The possibility of conducting such a study appeared not so long ago, since the potential of interatomic interaction was developed only in 2017, so that such a large-scale study for this alloy is carried out for the first time.

2. Simulation Details

The main method for modeling the glass transition in this paper was the molecular dynamics method or MD. In this investigation, 2 types of potentials were used. First type of potential is the potential of an embedded atom or EAM (such as potentials for the interaction of Ni-Nb and Ni-Zr, that were developed by Mendelev and others in 2016). They were used in calculations for pure Zr and Nb and compare features of Ni-Nb and Ni-Zr alloys with the features of Zr-Nb. The second type of potential is the potential of an immersed atom with an angular dependence (ADP potential Smirnov and Starikov in 2017 [1]). It was used for study of Zr-Nb alloy and it was the main in this work. Calculations were performed under constant pressure conditions for a cubic computational cell in NVE-ensemble, for a percentage Nb range from 10 to 90 percent and cooling rates from 0.5 to 11 per 10^12 K/s. All calculations were performed using the

LAMMPS package [3]. To eliminate surface effects, periodic boundary conditions were used. An important method of analysis in my work is the construction of Voronoi polyhedra [2] – the set of points closest to this atom. Also in the work for the analysis used radial-distribution functions (or RDF) and the common neighbors analysis. A program has been written to separate icosahedral clusters from other atoms of the structure. The main data used are the coordinates of the atoms at different points in time (obtained using numerical integration). With their help, the Voronoi polyhedra and the RDF are constructed.

3. Main Results

During modeling, it was found that upon transition to the amorphous phase, the temperature dependence of the number of icosahedral clusters changes. Such a sharp change is observed only for this type of clusters. This gave the CNA and the Voronoi polyhedron methods. It was also determined that the splitting of the second peak of RDF is observed only forepy cross-function Zr-Nb. It is not observed for zirconium and for Nb RDF. When isolating separately the icosahedra from the system, the splitting of the second peak for the Niobium RDF was also obtained. Was found the correspondence between the peaks of the RDF and the distances between atoms in icosahedra. The formation of icosahedral clusters only by Niobium is explained by the relative size of the atoms – for niobium, the radius is smaller and as a result there are fewer neighbors. The number of neighbors was also verified using the Voronoi polyhedra; the relationship with the size of the atoms was confirmed. By changing the dependence of the icosahedra, glass transition temperatures were determined. The glass transition temperatures were also determined depending on the cooling rate and concentration, a diagram of the limiting cooling rates was plotted, depending on the percentage of Niobium. The cooling temperature was determined by two more methods – using the Wendt-Abraham criterion and changing the temperature dependence of the diffusion coefficient.

References

1. Smirnova, D.E. and Starikov, S.V.: An interatomic potential for simulation of Zr-Nb system. Computational Materials Science **129**, 259–272 (2017).
2. Stukowski, A.: Visualization and analysis of atomistic simulation data with OVITO – the Open Visualization Tool Modelling. Simul. Mater. Sci. Eng. **18**, 015012 (2010).
3. Plimpton, S.: Fast parallel algorithms for short-range molecular dynamics. Journal of Computational Physics **117** (1), 1–19 (1995).

Vibrational Free Energy in Various Methods and Its Reproducibility

Pavel Korotaev [1,2], Maxim Belov [2,] and Aleksey Yanilkin [1,3]

[1] Center for Fundamental and Applied Research, Dukhov Research Institute for Automatics,
127055, Sushchevskaya 22, Moscow, Russia
[2] Material Modeling and Development Laboratory, NUST “MISIS”, 119991 Leninskiy pr. 4,
Moscow, Russia
[3] Moscow Institute of Physics and Technology, 141700 Institutskiy per. 9, Dolgoprudny,
Moscow Region, Russia

Abstract. There are several modern methods for calculation of Helmholtz free energy, associated with atomic vibrations: quasiharmonic approximation, self-consistent lattice dynamics method, calculation based on velocity autocorrelation function, temperature-dependent effective potential and thermodynamic integration. In this paper, we review these methods and compare the results of calculations obtained by these methods with each other in a wide temperature range, and their applicability is discussed. As a tool we use classical molecular dynamics modeling on the example of bcc molybdenum and fcc aluminum. The relation is observed between how the vibrational spectrum changes with temperature due to the anharmonicity of the potential and how free energy is reproduced. All methods are consistent with each other within 10 meV/at. at relatively low temperatures. At high temperatures, the discrepancy reaches tens of meV/at., although the relative error is less than 5%.

Keywords: Helmholtz free energy, Vibrational properties, Anharmonicity, TDEP, SCAILD, Thermodynamic integration, VACF

1 Introduction

Knowledge of free energy in various external conditions is necessary for constructing phase diagrams, determining the stability of crystal structures and defects, prediction of chemical reactions. At this point the possibility of accurate calculation of the free energy without involving experimental data is important, since experiments in a wide range of external conditions are difficult. This question is also relevant in the development of new materials, when the criterion of phase stability should be satisfied.

A highly significant contribution to the free energy of crystals is caused by thermal motion of atoms. Therefore, we need a method for accurately reproducing the lattice dynamics. On the other hand, it must be computationally efficient, so that the calculation of free energy would not be a bottleneck, for instance, while searching for a new materials.

2 Methods

We review several modern methods for calculation of Helmholtz free energy, associated with atomic vibrations: quasiharmonic approximation and small displacements method [1, 2] self-consistent lattice dynamics method [3], calculation based on velocity autocorrelation function [4], temperature-dependent effective potential [5, 6] and thermodynamic integration [7]. As a tool we use classical molecular dynamics modeling via LAMMPS package [8] on the example of bcc molybdenum and fcc aluminum. The EAM potential was used for aluminum [9] and molybdenum [10].

3 Results

The results of free energy calculation by these methods are compared with each other in a wide temperature range, and their applicability is discussed. The relation is observed between how the vibrational spectrum changes with temperature due to the anharmonicity of the potential and how free energy is reproduced.

4 Conclusions

All methods are consistent with each other within 10 meV/at. at relatively low temperatures. At high temperatures, the discrepancy reaches tens of meV/at., although the relative error is less than 5%.

References

1. Fultz, B.: Vibrational thermodynamics of materials, Prog. Mater Sci. **55** (4), 247–352 (2010).
2. Grimvall, G., Magyari-Köpe, B. Ozoliņš, V., and Persson, K.A.: Lattice instabilities in metallic elements, Rev. Mod. Phys. 84 (2), 945–986 (2012).
3. Souvatzis, P., Eriksson, O., Katsnelson, M.I., and Rudin, S.P.: Entropy driven stabilization of energetically unstable crystal structures explained from first principles theory, Phys. Rev. Lett. **100** (9), 095901(2008).
4. Dickey, J.M. and Paskin, A.: Computer simulation of the lattice dynamics of solids, Phys. Rev. **188** (3), 1407–1418 (1969).
5. Hellman, O., Abrikosov, I.A., and Simak, S.I.: Lattice dynamics of anharmonic solids from first principles, Phys. Rev. B **84** (18), 180301 (2011).
6. Hellman, O., Steneteg, P., Abrikosov, I.A., and Simak, S.I.: Temperature dependent effective potential method for accurate free energy calculations of solids, Phys. Rev. B **87** (10), 104111 (2013).
7. Kirkwood, J.G.: Statistical mechanics of fluid mixtures, J. Chem. Phys. **3** (5) 300–313 (1935).
8. LAMMPS Molecular Dynamics Simulator. http://lammps.sandia.gov
9. Liu, X.-Y., Ercolessi, F., Adams, J.B.: Aluminium interatomic potential from density functional theory calculations with improved stacking fault energy, Model. Simul. Mater. Sci. Eng. **12** (4), 665 (2004).
10. Starikov, S.V., Insepov, Z., Rest, J., Kuksin, A.Y., Norman, G.E., Stegailov, V.V., and Yanilkin, A.V.: Radiation-induced damage and evolution of defects in Mo, Phys. Rev. B **84** (10) 104109 (2011).

Investigation of Elastic Properties of Ni_2AlX (X=Ti, Nb, Hf) Ternary Compounds

K. Sidnov[*1,2], S. Vorotilo[1], A. Rogachev[1,2], I. Abrikosov[1,3]

[1]Materials Modeling and Development Laboratory, National University of Science and Technology ‘MISIS’, Moscow, Russia

[2]Merzhanov Institute of Structural Macrokinetics, and Materials Science Russian Academy of Sciences, Chernogolovka, Russia

[3]Department of Physics, Chemistry, and Biology (IFM), Linköping University, Linköping, Sweden

*k.sidnov@misis.ru

In this work, using first-principles calculations we define the elastic properties of Ni_2AlM Heusler phases, where M=Ti, Nb, Hf. According to Pough’s criterion [1, 2] and Couchy pressure [3], Ni_2AlNb has the highest plasticity among the investigated compounds (including NiAl). To experimentally characterize these materials, we synthesized phase-pure NiAl, Ni_2AlTi, Ni_2AlNb, Ni_2AlHf using the combustion synthesis approach. Macrokinetic parameters of combustion synthesis were measured, including the ignition temperature T_i=650-660°C (approx. the melting point of Al) and combustion temperature T_c=1120–1330°C. Microstructural and XRD characterization revealed the formation of single-phase NiAl and Heusler alloys with a median grains size 15–20 µm. Micro-indentation of synthesized powders revealed that the Ni_2AlNb phase has the highest plasticity parameter [4], which corresponds well to the calculations and corroborates the applicability of the employed plasticity descriptors.

References

1. Pugh, S.F.: XCII. Relations between the elastic moduli and the plastic properties of polycrystalline pure metals, The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science **45**, 367 (1954).
2. Hill, R.: The elastic behavior of a crystalline aggregate. Proceedings of the Physical Society. Section A 65, 349 (1952).
3. Pettifor, D.: Mater. Sci. Technol. **8**, 392 (1992).
4. Milman, Yu., Galanov, A., and Chugunova, I.: Plasticity Characteristic Obtained Through Hardness Measurement. Acta Metallurgica Et Materialia **41**, 2523–2532 (1993). Doi 10.1016/0956-7151(93)90122-9

Investigation of Si-Au and Si-Al Melt Crystallization

I.S. Gordeev[1,2], L.N. Kolotova[1,2] and S.V. Starikov[2]

[1] Joint Institute for High Temperatures, Russian Academy of Sciences, Moscow 125412, Russia
[2] Moscow Institute of Physics and Technology (State University), Dolgoprudny 141700, Russia
Ruhr-Universität Bochum, Bochum 44801, Germany
gordeevilu@gmail.com

Abstract. Silicon nanoobjects are very interesting for nanophotonics due to unique optical properties depending on the atomic structure. Thus, it is a great interest in study the structural features of these particles, such as the grain size and the distribution of Au and Al atoms in crystal grains and between them. The influence of cooling rate on Si-Al nanoparticles (NP) structures was studied in this work for different NP sizes and Al concentrations. The simulations were carried out in the quasi-2D case: diameter of NP was up to 80 nm and thickness of cell was about 10 nm with periodic boundary conditions. In turn this may result smaller grain sizes in simulation in comparison to the experiment. That's why, additional one-dimensional simulations were performed to study the grain size dependence on the Au or Al concentration in NP. The movement of the crystal grain boundary and changes in the gold distribution during crystallisation were analysed. The results of simulations indicate that gold atoms try to leave crystallization area via atomic diffusion. So this effect leads to limitation of grain size for larger Au concentrations. The same simulations were carried out for Si-Al NP with novel interatomic potential.

Keywords: Silicon, Nanoparticle, Crystallization, gold, aluminium.

1 Introduction

Silicon nanoobjects are very interesting for nanophotonics due to unique optical properties depending on the atomic structure. Their optical properties can be tuned by changing the phase composition and doping by metal atoms. Thus, it is a great interest in study the structural features of these particles, such as the grain size and the distribution of Au and Al atoms in crystal grains and between them.

2 Crystalization of Si-Au and Si-Al Nanoparticles

The simulations were carried out in the quasi-2D case: diameter of NP was up to 80 nm and thickness of cell was about 10 nm with periodic boundary conditions. In turn this may result smaller grain sizes in simulation in comparison to the experiment. That's why, additional one-dimensional simulations were performed to study the grain size dependence on the Au or Al concentration in NP. The movement of the crystal grain boundary and changes in the gold distribution during crystallisation were analysed.

3 Results

The influence of cooling rate on Si-Au and Si-Al nanoparticles (NP) structures was studied in this work for different NP sizes and metal concentrations. The results of simulations indicate that gold atoms try to leave crystallization area via atomic diffusion. So this effect leads to limitation of grain size for larger Au concentrations. The same simulations were carried out for Si-Al NP with developed interatomic potential based on [1].

References

1. Starikov, S.V. et al.: Computational Materials Science **142**, 303–311 (2018).

Characterization of Data Provenance in Computational Engineering by an Ontological Representation of Simulation Workflows

Martin Thomas Horsch,[1] Christoph Niethammer,[2] Silvia Chiacchiera,[1]
Peter Schiffels,[3] Michael A. Seaton,[1] Ilian T. Todorov,[1] Jadran Vrabec,[4]
Philipp Neumann,[5] and Welchy Leite Cavalcanti[3]

[1] UK Research and Innovation, STFC Daresbury Laboratory, Keckwick Ln, Daresbury, Cheshire WA4 4AD, United Kingdom
[2] High Performance Computing Center Stuttgart, Nobelstr. 19, 70569 Stuttgart, Germany
[3] Fraunhofer Institute for Manufacturing Technology and Advanced Materials, Wiener Str. 12, 28359 Bremen, Germany
[4] Technische Universität Berlin, Thermodynamics and Process Engineering, Ernst-Reuter-Platz 1, 10587 Berlin, Germany
[5] Universität Hamburg, Scientific Computing, Bundesstr. 45A, 20146 Hamburg, Germany

`martin.horsch@stfc.ac.uk`

Abstract. This demo presents the Ontology for Simulation, Modelling, and Optimization (OSMO), i.e., a semantic asset that can be employed to formally represent simulation workflows in computational molecular engineering. It discusses technical aspects of the ontology and its OWL DL implementation as well as two related diagram notations for workflows: Reduced workflow graphs, i.e., MODA notation, and extended workflow graphs where MODA elements are connected to nodes that represent logical resources.

Keywords: Ontology, simulation workflow, materials modelling

1 Demo: Ontology for Simulation, Modelling, and Optimization

Where databases and platforms using different data structures and file formats interoperate, or where data from various sources are combined, agreement on semantics becomes a necessity. By an ontology, rules are formulated for entities (i.e., objects) and relations (i.e., properties of objects), which can guide the data ingest into databases and extend the possibilities of data extraction and evaluation by automated logical reasoning.

For data technology in computational molecular engineering, the characterization of workflows is relevant in two major ways. First, workflows are designed and communicated to simulation environments where materials models are evaluated to generate data by simulation. Second, in order to integrate data obtained in different ways (e.g., from simulation and experiment, or from simulations with different models or solvers), simulation results need to be stored together with metadata that describe their provenance, i.e., the process by which they have been produced. Hence, simulation workflows need to be described in a machine-readable way. There are a great variety of environments and languages for workflows (e.g., file formats and graph languages). Many of these, including AiiDA [1], MODA [2], Salome/YACS [3], and the TaLPas workflow and performance modelling environment [4], are applicable to simulation workflows in materials modelling.

OSMO [5] is compatible with RoMM [6], and it is directly based on MODA; by providing a common semantic basis for workflows that were designed with different tools, OSMO can be employed to consistently integrate data provenance descriptions for materials modelling data from diverse sources [5]. The demonstration at DACOMSIN will illustrate how OSMO may be used in practice to provide a common framework for previously non-interoperable workflow environments.

Acknowledgements

The co-authors M. T. H., S. C., P. S., M. A. S., I. T. T., and W. L. C. acknowledge funding within the Virtual Materials Marketplace (VIMMP) project; this project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No. 760907. The co-authors C. N., J. V., and P. N. acknowledge funding from the German Federal Ministry for Education and Research (BMBF) under grant agreement No. 01IH16008 (TaLPas).

References

1. Pizzia, G., Cepellotti, A., Sabatini, R., Marzari, N., and Kozinsky, B.: AiiDA – automated interactive infrastructure and database for computational science. Computational Materials Science **111**, 218–230 (2016).
2. CEN-CENELEC Management Centre: CEN Workshop Agreement 17284. CEN, Brussels (2018).
3. Ribes, A. and Caremoli, C.: Salomé platform component model for numerical simulation. In: Proceedings of COMPSAC 2007, pp. 553–563. IEEE Computer Society, Los Alamitos (2007).
4. Shudler, S., Vrabec, J., and Wolf, F.: Understanding the scalability of molecular simulation using empirical performance modelling. In: Programming and Performance Visualization Tools, pp. 125–143. Springer: Heidelberg (2018).
5. Horsch, M.T., Niethammer, C., Boccardo, G., Carbone, P., Chiacchiera, S., Chiricotto, M., Elliott, J.D., Lobaskin, V., Neumann, P., Schiffels, P., Seaton, M.A., Todorov, I.T., Vrabec, J., and Cavalcanti, W.L.: Semantic interoperability and characterization of data provenance in computational molecular engineering. To appear (2019).
6. De Baas, A.F.: What Makes a Material Function? EU Publications Office, Luxembourg (2017).

Author Index