

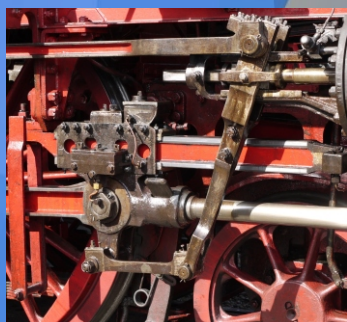
INTERNATIONAL SCIENTIFIC JOURNAL

INDUSTRY 4.0

ISSN (PRINT) 2534-8582

ISSN (WEB) 2534-997X

YEAR III ISSUE 5/2018



①



②



③



④

PUBLISHED BY
SCIENTIFIC TECHNICAL UNION OF MECHANICAL ENGINEERING
“INDUSTRY 4.0”, BULGARIA

INTERNATIONAL SCIENTIFIC JOURNAL

INDUSTRY 4.0

YEAR III, ISSUE 5 / 2018

ISSN (PRINT) 2543-8582, ISSN (WEB) 2534-997X

PUBLISHER

SCIENTIFIC TECHNICAL UNION OF MECHANICAL ENGINEERING "INDUSTRY 4.0"

108, Rakovski Str., 1000 Sofia, Bulgaria

tel. (+359 2) 987 72 90,

tel./fax (+359 2) 986 22 40,

office@stumejournals.com

WWW.STUMEJOURNALS.COM

EDITOR IN CHIEF

Prof. D.Sc. Georgi Popov, DHC, Technical University of Sofia, BG

Prof. Dr. Dr. Jivka Ovtcharova, DHC, Karlsruhe University of Technology, DE

EDITORIAL BOARD

Acad. Igor Bychkov, RU

Cor. member Alexey Beliy, BY

Cor. member Svetozar Margenov, BG

Prof. Alexander Afanasyev, RU

Prof. Alexander Guts, RU

Prof. Andrzej Golabczak, PL

Prof. Andrey Firsov, RU

Prof. Bobek Shuklev, MK

Prof. Boris Gordon, EE

Prof. Branko Sirok, SI

Prof. Claudio Melchiorri, IT

Prof. Cveta Martinovska, MK

Prof. Dale Dzemydiene, LT

Prof. Dimitar Yonchev, BG

Prof. Dimitrios Vlachos, GR

Prof. Galina Nikolcheva, BG

Prof. Galina Zhavoronkova, UA

Prof. Gerard Lyons, IE

Dr. Giovanni Pappalettera, IT

Prof. Henrik Carlsen, DK

Prof. Idilia Bachkova, BG

Prof. Idit Avrahami, IL

Prof. Iurii Bazhal, UA

Prof. Jürgen Köbler, DE

Prof. Jiri Maryska, CZ

Prof. Katia Vutova, BG

Prof. Lappalainen Kauko, FI

Dr. Liviu Jalba, RO

Prof. Luigi del Re, AT

Prof. Majid Zamani, DE

Prof. Martin Eigner, DE

Prof. Michael Valasek, CZ

Prof. Milija Suknovic, RS

Prof. Miodrag Dashic, RS

Prof. Mladen Velez, BG

Prof. Murat Alanyali, TR

Prof. Nina Bijedic, BA

Prof. Ninoslav Marina, MK

Prof. Olga Zaborovskaia, RU

Prof. Pavel Kovach, RS

Prof. Petar Kolev, BG

Prof. Peter Korondi, HU

Prof. Peter Sincak, SK

Prof. Petra Bittrich, GE

Prof. Dr. Predrag Dasic, RS

Prof. Radu Dogaru, RO

Prof. Raicho Ilarionov, BG

Prof. Raul Turmanidze, GE

Prof. René Beigang, DE

Prof. Rozeta Miho, AL

Prof. Sasho Guergov, BG

Prof. Seniye Ümit Oktay Firat, TR

Prof. Sreten Savicevic, ME

Prof. Stefan Stefanov, BG

Prof. Svetan Ratchev, UK

Prof. Sveto Svetkovski, MK

Prof. Tomislav Šarić, HR

Prof. Vasile Cartofeanu, MD

Prof. Vidosav Majstorovic, RS

Prof. Vjaceslavs Bobrovs, LV

Prof. Inocentiu Maniu, RO

Dipl.-Kfm. Michael Grethler, D

CONTENTS

TECHNOLOGICAL BASIS OF “INDUSTRY 4.0”

METHODS AND TECHNOLOGIES OF BUILDING AN INTELLIGENT SERVICE FOR ENERGY TECHNOLOGY FORECASTING

A.N. Kopaygorodsky 223

THE EFFICIENCY COMPARISON OF THE PRISM AND STORM PROBABILISTIC MODEL CHECKERS FOR ERROR PROPAGATION ANALYSIS TASKS

T. Fabarisov, N. Yusupova, K. Ding, A. Morozov, K. Janschek 229

MODEL REDUCTION ALGORITHM FOR FAST NEUTRALITY TESTS AND FAULT LOCALIZATION OF SIMULINK MODELS

X. Hu, A. Morozov, K. Ding, K. Janschek, N. Yusupova 232

USING THE FUZZY CONTROLLER TO CONTROL PROCESS PARAMETERS

E.A. Muravyova, D.S. Radakina 236

STREAM HANDLING LARGE VOLUME DOCUMENTS IN SITUATIONALLY-ORIENTED DATABASES

V. V. Mironov, A. S. Gusarenko, N. I. Yusupova 240

RECOGNITION OF TEXT INFORMATION IN THE BRONCHOPULMONARY DISEASES DIAGNOSIS SYSTEM

G.R. Shakhmametova, A.A. Evgrafov, R.Kh. Zulkarneev 245

DOMINANT TECHNOLOGIES IN “INDUSTRY 4.0”

THE MULTIPROCESSOR CONTROL SYSTEM OF DYNAMICS MULTIPLY CONNECTED TECHNOLOGICAL ELECTRIC DRIVES

O. Z. Khasanov, Z. M. Khasanov, I.R. Akhmetzyanova 248

MODELLING OF THE COGNITIVE MAP FOR THE MULTIDIMENSIONAL OBJECT'S CONTROL SYSTEM

Sazonova T.V., Khristodulo O.I. 251

BUSINESS & “INDUSTRY 4.0”

CLUSTER ANALYSIS TO IDENTIFY POTENTIAL EMPLOYERS OF UNIVERSITY GRADUATES

A.G. Karamzina 255

EXAMPLE OF PLM-SYSTEM ADOPTION AT PJSC «UEC-UMPO» IN THE NETWORK OF INTERACTION ON THE PROJECT PD-14

Kuznecov A.A., Mavrina A.S., Sapozhnikov A.Yu. 259

JUSTIFICATION OF VOLUME OF OUTPUT DEPENDING ON PRODUCT ENVIRONMENTAL CRITERION BY NONLINEAR MATHEMATICAL PROGRAMMING

Rodionova L.N. 262

SOCIETY & „INDUSTRY 4.0”

SCENARIO APPROACH FOR ANALYZING EXTREME SITUATIONS IN ENERGY FROM A CYBERSECURITY PERSPECTIVE

Aleksei Massel, Gaskova Daria 266

METHODS AND TOOLS FOR PROCESSING SEMI-STRUCTURED DATA ON THE EXAMPLE OF ACCOUNTING FOR EDUCATION IN THE SELECTION OF PERSONNEL IN THE IT INDUSTRY

I N. I. Yusupova, O. N. Smetanina, M. M. Gayanova 270

INTELLIGENT DECISION-MAKING SUPPORT IN ENERGY AND ECOLOGY IN VIEW OF THE QUALITY OF LIFE

L.V. Massel 280

FEATURE EXTRACTION FROM IMAGES WITH USE OF CONVOLUTIONAL NEURAL NETWORKS: APPLICATION TO SECURING PERSONAL DATA

T. R. Sokhin, K. V. Mironov 284

METHODS AND TECHNOLOGIES OF BUILDING AN INTELLIGENT SERVICE FOR ENERGY TECHNOLOGY FORECASTING

A.N. Kopaygorodsky

Melentiev Energy Systems Institute SB RAS, Irkutsk National Research Technical University, Irkutsk, Russia

e-mail: kopaygorodsky@mail.ru

Abstract: This article reports the approach and software tools for decision-making support in forecasting the energy infrastructure development. The author considers the problem of searching for information from various open sources, technology of information searching, knowledge detection and classification. The author describes architecture of Intelligent information system. For experts this classification and integrated warehouse simplify search for the knowledge required.

1. Introduction

Future-oriented research of industry development is becoming particularly relevant because of complicated economic conditions worldwide. This is necessitated by the task (a) to predict potential results of innovative projects with high reliability and (b) to save money. Because the Energy industry is basic production of final products largely depends on its development. In addition, in the recent years, alternative energy technologies have been actively developed, for instance those related to ecology and energy conservation, e.g. Green Energy.

The technological forecasting of the RF energy sector up to year 2035 considers three different world energy development scenarios [1]. Fidelity of each scenario cannot be properly estimated. This leads to the high uncertainty in innovative development strategic planning because of high investment risks in the energy sector. Improving evaluation of technological development and rolling forecasts will let to focus on the amplification of certain technologies with better effect in future. Thus, providing researchers with topical information and knowledge is an important part of development of expert assessment of new technological solutions in energy infrastructure.

The team employed at the Melentiev Energy Systems Institute, Siberian Branch, Russian Academy of Sciences (ESI SB RAS) has been developing the intelligent information resources and support software tools for decision-making and collective expert activity for innovative development forecasting of energy infrastructure. The latter is truly important for work at the ESI. The experts from various fields such as Energy, Mathematics, Economy, Ecology and Informatics are involved in research.

2. Related works

Since 2003, the international workgroup of leading scientists from the USA, Europe and East Asia coordinates research on Future Oriented Technology Analysis [2]. Researchers are carried out on such closely related areas as Technology Monitoring, Watch and Alerts, Technical and Competitive Intelligence, Technological Forecasting, Strategic Technology Assessment, development Technology Roadmapping, Technology Foresight and other.

Forecasting as a method of research is used in the domain of Energy Infrastructure to study the development and the functioning. In [3], authors were considering the technological prospects of various directions of decisions of the problem of resource restrictions of the development of wind and solar energy.

Recently, more and more often analysis results of Big Data and Linked Open Data as information source has been used. For example, the Big Data analysis was used for Electricity Load Forecasting [4]. This model of Electricity Load was designed to make predictions for time series with specific properties (strong seasonal dependence and concept drift). The prediction results via Big Data analysis described the behavior of the real Energy System in the near future very well.

Text Mining is the process of deriving information from text. Some applications of this technique are Social Media Monitoring and Scientific Discovery. Authors of [5] propose a conceptual approach

to the research into customer satisfaction based on a detailed analysis of consumer reviews written in natural languages.

Artificial Intelligence techniques successfully are used for forecasting the conduct of separate energy technologies. In [6], authors use patent indicators to predict the technological advances in Hydrogen Storage Materials (HSM). Patent analysis was been carried out using bibliometrics and Text Mining approaches in order to forecast the future trend of development. Authors were evaluating the technological life cycle stage, HSM class prominence and the role of different countries in HSM patenting.

3. Energy infrastructure development forecasting

The work on scientific and technological forecasting of Energy infrastructure is complex and requires solving the main problems:

- Need to support sustainable scientific and technological development;
- Acceleration of technological changes;
- Strengthening the impact of progress in science and technology on socio-economic processes;
- Increase in the complexity of facilities, counterparties and, as a consequence, the systems that unite them;
- Strengthening interdisciplinary effects, influence of interrelated areas of knowledge, including convergence of technologies;
- Need for a systematic adaptive forecast and mechanisms for "rolling" planning with continuous updating of scientific and technical information.

The ESI researchers are short of present data and topical knowledge to properly energy infrastructure forecasting.

Available basic methods of technological forecasting are currently used to study the Energy of RF:

1. Expert survey, e.g. Delphi and Foresight method. This is a simple cheap and quick solution. However, these methods have poor qualities as subjectivity, low validity and lack of responsibility.
2. Technological analysis by individual energy companies. This method is based on econometrics, monitoring, databases of development, operational analysis of problems and proposals.
3. System analysis of technology is research that requires a lot of time and effort of highly qualified scientific teams.

Because the last method very expensive and require a lot of information and explicit knowledge usually first two methods are applied for developing forecasts. And moreover using a second method available only within energy companies. Application of semantic technologies facilitates scientific and technological forecasting in the energy sector. This procedure includes:

- Semantic Information Search;
- Analysis of weakly structured texts and data, including the extraction of terms, concepts, their alignment and aggregation;
- Statistical processing and analysis of obtained semantically structured data, including computing of quantitative indicators and indicators of technology development;

- Cognitive technologies of visualization;
- Ontological modelling.

A new approach to extract and analyze knowledge and data lets us to improve quality of scientific results. The uncertainty can be reduced via Large Data Analytics (LDA) applied to global science, technology and innovation data sources. When applying LDA, it is feasible to identify the existing and raise new development trends, as well as foresee the technological breakthroughs by a comprehensive understanding of continuous innovation processes [7].

All possible sources of data and knowledge should be used for solving the problem of development forecasting. Even if the ESI has its own data sources and previous results, it is not enough to fulfill high-quality forecasts. Researchers should attract additional knowledge and data to improve the quality and accuracy of forecasts. Different sources contain information in diverse forms and information have varied level of reliability. Open government data and other open data may be used as additional information sources. For example, to improve the quality of results and probability of energy development forecasting models the researchers may apply to the database of patents and inventions or news feeds. Our research team stick to the governmental information systems, which integrate information about scientific and technical projects and developments, e.g. CITIS, RFBR and FIPS, and various commercial systems, e.g. SCOPUS, Web of Science, RSCI and Science Index as the sources for Tech Mining. The information in such sources is commonly represented in text form. The Tech Mining is a special form of LDA based on the intelligent text mining of electronic information sources. Such Intelligent technologies are capable to reduce large information flow and to classify the information items. It is possible to estimate qualitative and quantitative innovation indicators of potential technological growth in the areas of interest.

The task of elaboration software tools for energy infrastructure development forecasting consists of two main parts:

- Creating an Intelligent Information Environment for collaborative expert work and knowledge accumulation;
- Improving the existing and creating new methods of information search and forecasting.

Researchers use two classes of models to forecast: (1) Qualitative models show the relationship and influence of various factors; they are employed to determine some trends; (2) Quantitative model. The qualitative analysis as the first stage of forecasting development makes it possible to quickly and cheaply solve the problem. In some cases, quantitative modelling can be used as a more expensive approach, but giving better results in addition to the qualitative solution obtained. The main drawback of quantitative models is a long process of their construction and data preparation. The quantitative models are commonly used to compute the consequences of events and strategic decisions. In any case, researchers need to make relevant changes in the model to obtain adequate results, when researching forecast of energy development.

4. Technology of information searching

The most important task is providing relevant information to researchers. They apply some existing models and create new ones in the study process. Fresh information and knowledge permit the existing models to be updated. A large stream of poorly structured information significantly slows down and complicates building and updating of models. This is why researchers have to spend a lot of time searching for and identifying the facts of interest. This is extremely inefficient. Automatic classification is required to effectively extract knowledge from large data sets. Knowledge in explicit form may be described by set of knowledge element (digest). Each element from the set may be mapped to some concept from ontology. This can be represented as

$$F_K: T \rightarrow \{K_I\},$$

$$F_O: K_I \rightarrow O_I, O_I \subset O' \quad , \text{ where}$$

F_K presents the text fragment T to set of knowledge element K ; $\{K_I\}$ is digest of T ; F_O is a mapping for all knowledge element K_I to concepts O_I . And O' is a generalized ontology. Application of Ontology system allows us to simplify navigation and search of information.

Figure 1 exhibits the process of extracting interesting knowledge from the general information flow and their subsequent processing. Knowledge for energy infrastructure development forecasting arrives from various sources. Each information flow is unified and the text fragments are extracted. Some elements may contain keywords, abstracts, descriptions in explicit forms. The other elements may contain only raw text. At the stage of classification two important tasks are performed: (a) definition of word-set and (b) mapping with concepts of the Ontology system.

Methodical approaches to knowledge collection from text sources are employed to solve some tasks:

- to obtain raw semi-structured text data;
- to clean, identify and extract the text;
- to classify the text; draft the digest for each obtained element.

Figure 2 exhibits this process. Main steps of process are the extraction of text fragments and the building document digest. Thus process may be divided between two components: data search engine and data analysis engine. In this case data search engine encapsulate all exclusivity of data source. And data analysis engine perform common algorithm of classification by text fragment. The first stage of text analysis should be language detection. Using a method of character frequency allow to spot code page and text alphabet and then draw a conclusion about language.

On the next stage may be applied stemming algorithm [8] and improved TF-IDF metrics [9] to detect whole Term Frequency. This approach provide a simple way to define document class. But in this case estimate is rough.

For a better definition of Terms and their relations may be used specialized tools such as KLAN system [10, 11]. Context relation between words may be identified via analysis of word sequences. The result of this step is the set of normalized word-sets (patterns). The groups of patterns are mapped to keywords and concepts of ontology space. Thus, the text fragments of elements are mapped to some set of concepts and numerical characteristics (frequency). As a result of classification at the last step the digest contains keywords related to the Ontology system concepts only. This element holds information about the meaning of the text fragment of the knowledge element. The Knowledge warehouse keeps digests, links on raw documents and ontologies. If the text fragment has not been correlated with the key concepts of the Ontology system, it is not stored in the Knowledge warehouse. Thus, the warehouse contains only the elements of knowledge correlated with the concepts of ontologies. Researchers can access structured knowledge for model adjustment and research.

The texts of scientific and specialized literature were used for learning and testing the classification subsystem of text. Besides, the other texts were used to determine common phrases of the Russian language (words of general vocabulary). The databases of patterns consist of two parts: (1) general vocabulary (about 2000 elements); (2) scientific vocabulary (about 1500 elements). Maximum length of element of scientific vocabulary is 4 words.

Figure 3 exhibits two ways of knowledge warehouse enrichment. The first is collecting documents from structured sources such as government information system, paper collection et al. Source-oriented data adapters may extract information from such sources

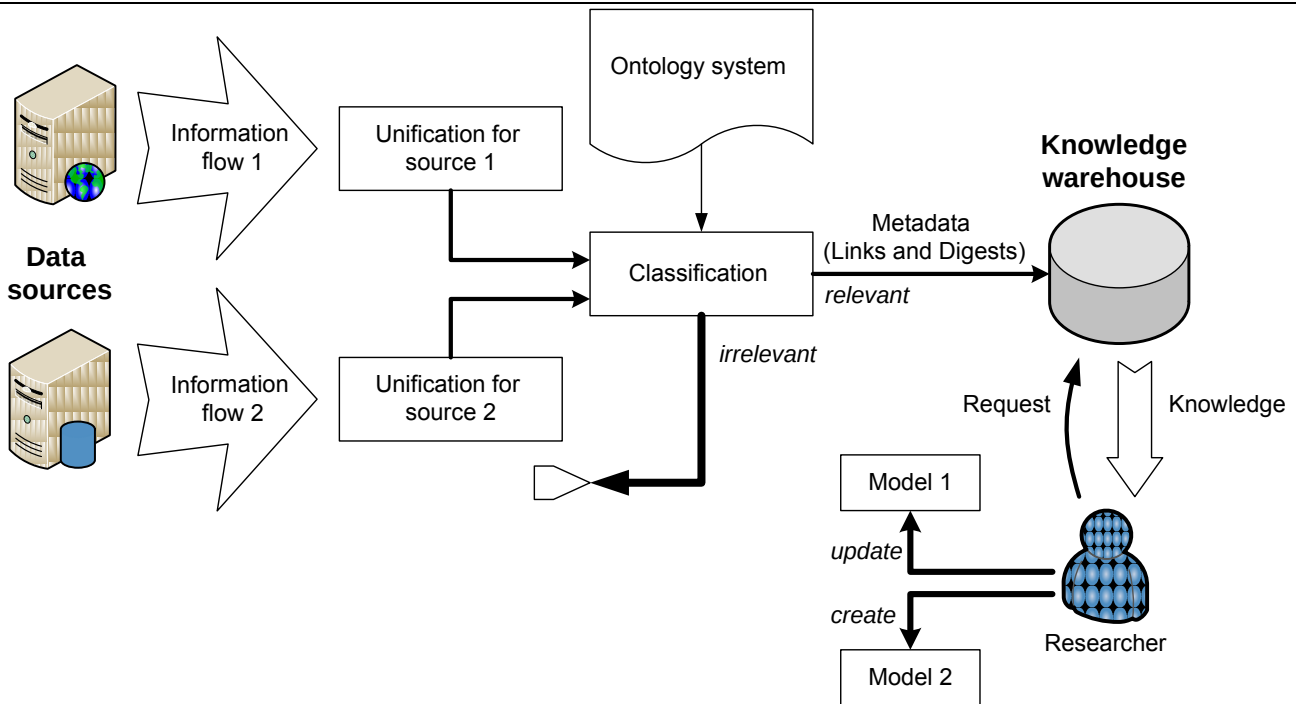


Fig. 1. The process of extracting knowledge from the sources, transformation and using for updating research models.

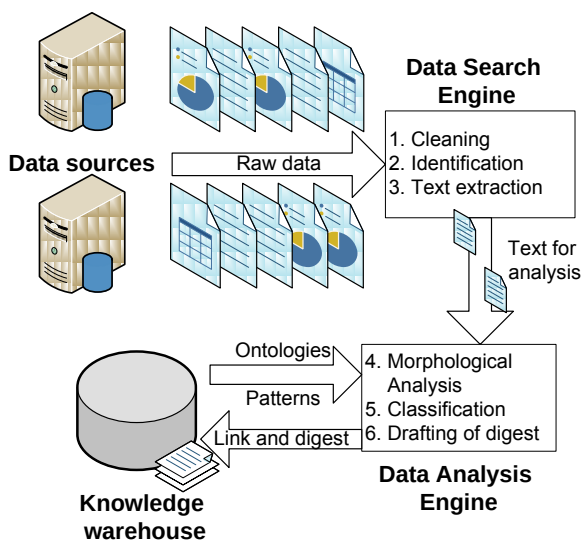


Fig. 2. The process of document classification based on analysis of text fragments.

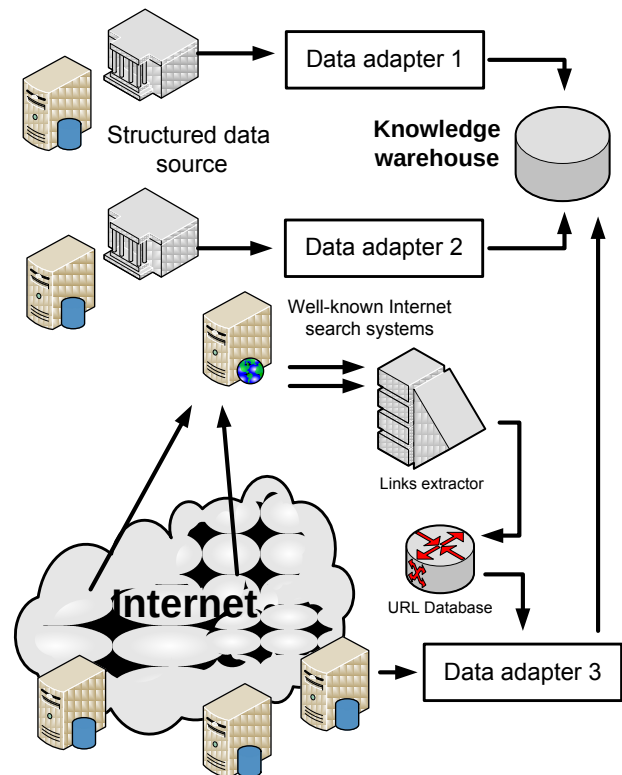


Fig. 3. The scheme of collecting information and enrichment of Warehouse.

The second is a blind Internet searching. Special constructed crawlers extract information from Internet resources and news feeds. Crawlers start job from top level link then looking through HTML extract new links from next level and add to work list. There is a problem of creating initial link set for supply crawlers. The most rational approach was to obtain relevant links directly from search engines by predefined queries. The link extractor interacts with the search system and generates an answer file, the format of which is presented below

```
BEGIN name_of_engine timestamp search_query
url 1
```

```

title 1
description 1
...
url n
title n
description n
END
    
```

The first string contain keyword “BEGIN”, name of search engine, time stamp of request and query. Further there are 3*N rows. Each 3 rows contain information about URL, page title and description.

This format has benefits: all packages are related to time and Internet search system. This makes it possible to observe change in mediaspace in time.

To support expert work and substantiate recommended solutions the author develops Intelligent Information Environment of Energy Innovative Development Forecasting (IIEIDF). They apply the tool set of IIEIDF to perform own studies. IIEIDF supplies interconnections between applications and stores and transforms the data and knowledge.

5. Architecture of Intelligent Information Environment

Intelligent Information Environment of Energy Innovative Development Forecasting consists of several components (Fig. 4): data search and data analysis engines, knowledge and data warehouses; FishTail server; Web-server supporting the server-side components of Web-applications. IIEIDF applies the some existing components of Intellectual Collective Expert Environment (ICEE) [12]. The data search engine automatically performs scheduled searching, obtains knowledge elements and then extracts the text content. The data analysis engine performs lexical and morphological analysis of the text fragments and mapping to concepts.

Knowledge and data warehouse describes not only scientific knowledge, but also contains initial information for investigation. In addition, the knowledge warehouse store digests and links the knowledge elements from open Internet data and sources. Access to

data sources and separate servers of database management systems is reached by virtual integration methods [13]. These database management systems contain information and results of some other ESI SB RAS investigations.

The FishTail Server developed at ESI SB RAS is applied to store and manage RDF-triples associated with some knowledge and concepts of Ontology space. The FishTail is employed for enrichment by the new RDF-triples based on some predefined rules and existing RDF-triples. Each rule has body and head. Head contain one or more groups of triplets, body has only one triplets. If head is true and body is absent then FishTail adds body triplet in the storage.

Figure 5 exhibits placing a IIEIDF services. The most valuable services are located in ESI SB RAS. Data adapters and crawlers required fast Internet connection and low latency. Therefore, this components was be placed in Datacenter. Data exchange between sites is not large by reason of applying local copy of Ontology database. One important consequence of the selected architecture is a easy increasing count work-servers

Implementation of Web-based application components IIEIDF are performed in a Private Cloud on the basis of approach of Rich Internet Applications (RIA). RIA Web-application is a full functionality of traditional desktop applications. The implementation of the RIA client Web-based applications is performed on the Java platform as the most suitable and convenient one for development. The technology of Java Web Start (JWS) and Java Network Launching Protocol (JNLP) lets downloading and running quazilocal applications that are automatically updated through the Internet and work outside of the Web-browser environment.

Researchers commonly apply relevant modelling tools to support the proposed methodical approach, e.g. OntoMap, CognitiveMap, EventMap and BayNet. The modelling software tools for qualitative analysis are based on graphical core GrModeling [14]. This approach provides shared space for building and transforming models.

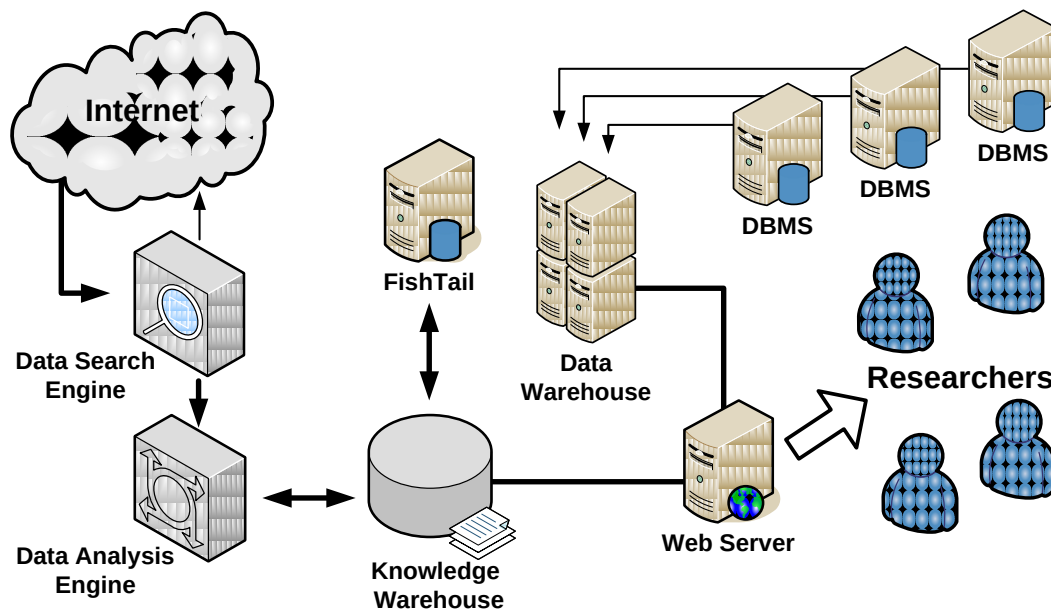


Fig. 4. Logical Architecture of Intelligent Information Environment of Energy Innovative Development Forecasting.

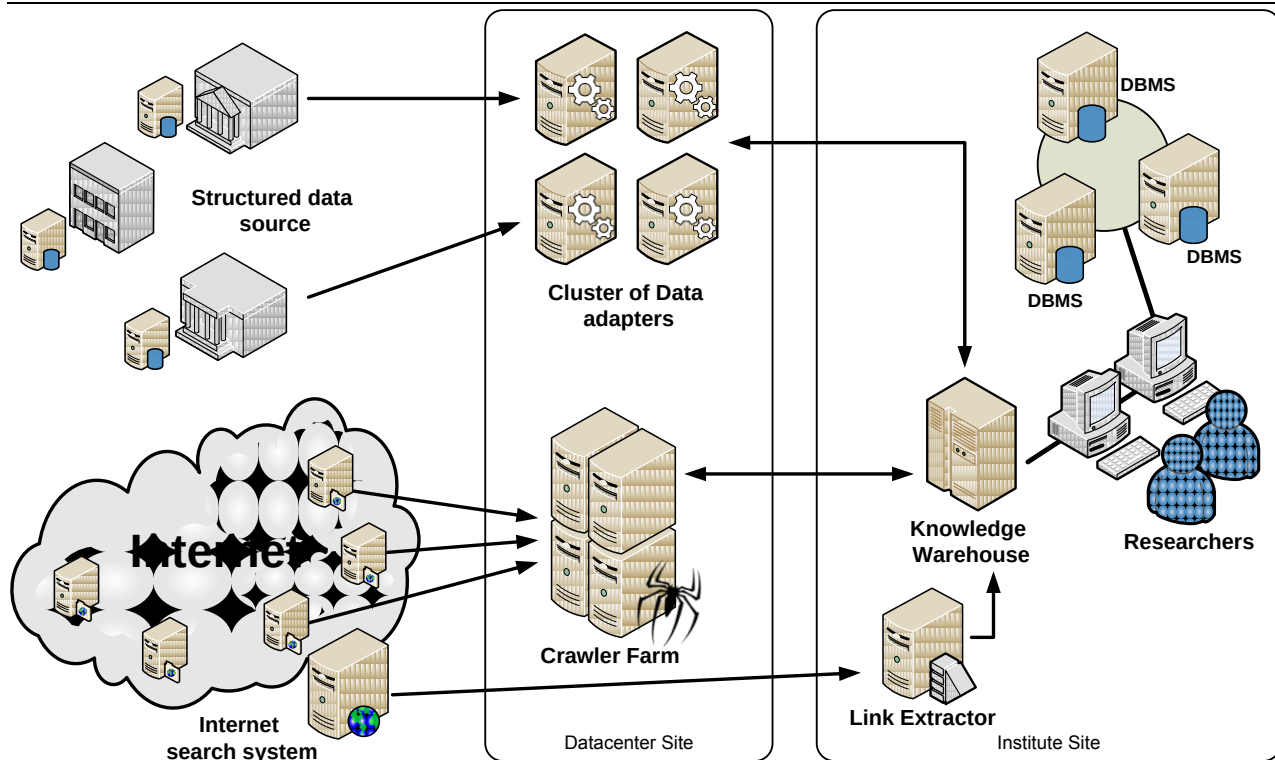


Fig. 5. Placing a services of Intelligent Information Environment of Energy Innovative Development Forecasting.

Implementation of native Web-application that contained not so many various functions is produced with such technologies as JavaScript (jQuery, Bootstrap et al.) and HTML5. Server-side components are developed in PHP and Java. To simplify the creation of PHP-applications is using its own MVC framework and such technology as TWIG.

6. Conclusion

The article considers methods, technology, algorithms and software tools to performing searching and analysis information from various sources:

- The Open data, Internet and other commercial systems may be employed as the additional knowledge sources for the purpose of the Energy Infrastructure Development Forecasting.
- The process of searching and extracting new knowledge is based on lexical and morphological analysis, TD-IDF method and frequencies analysis.
- The analysis of text fragments and creation of ontology-based digest improve the efficiency and quality of forecast estimation via simplified access to knowledge. The results of qualitative modelling are transferred into the process of numerical computing.
- The Intelligent Information Environment of Energy Innovative Development Forecasting includes the mechanisms for data search and analysis. Researchers use Web-applications to access knowledge and data warehouse, they also use the RIA-tools for creating own models.

These techniques surely improve quality of forecast results in the studies of energy infrastructure development.

Acknowledgments

The author is grateful to the Russian Foundation for Basic Research (RFBR) for partial financial support of the research via projects 17-07-01341, 16-07-00569. The results were obtained during the implementation of the basis scientific project of the fundamental

research programs of SB RAS III.17.2.1, reg. № AAAA-A17-117030310444-2, III.17.1.4, reg. № AAAA-A17-117030310436-7.

This research was performed using the servers of the multi-access scientific center "High Temperature Circuit" of Melentiev Energy Systems Institute SB RAS.

References

1. The technological forecast of sectors development of the Russian fuel and energy complex for period till 2035. Ministry of energy RF, 2016, Available at <http://minenergo.gov.ru/system/download-pdf/6365/66647>.
2. Cagnin C., Keenan M., Johnston R., Scapolo F., Barre R. "Future-oriented technology analysis. strategic intelligence for an innovative economy", Springer, 2008, 170 p., DOI: 10.1007/978-3-540-68811-2.
3. Nizhegorodtsev R. M., Ratner S. V., "Trends in the Development of Industrially Assimilated Renewable Energy: The Problem of Resource Restrictions", *Thermal Engineering*, Springer, 2016, pp. 197-207, DOI:10.1134/S0040601516030083.
4. Grmanova G., Laurinec P., Rozinajova V. et al., "Incremental Ensemble Learning for Electricity Load Forecasting", *Acta Polytech Hungarica*, vol. 13, no. 2, 2016, pp. 97-117. DOI: 10.12700/APH.13.2.2016.2.6.
5. Yussupova N., George Kovacs G., Maxim Boyko M. et al., "Models and Methods for Quality Management Based on Artificial Intelligence Applications", *Acta Polytech Hungarica*, vol. 13, no. 3, 2016, pp. 45-60. DOI: 10.12700/APH.13.3.2016.3.3.
6. Chanchetti L. F., Diaz S. M. O., Leiva D. R. et al., "Technological Forecasting of Hydrogen Storage Materials using Patent Indicators", *International Journal of Hydrogen Energy*, vol. 41, Elsevier, 2016, pp. 18301-18310, DOI:10.1016/j.ijhydene.2016.08.137.

7. Daim T.U., Chiavetta D., Porter A.L., Saritas O., "Anticipating future innovation pathways through large data analysis", *Innovation, technology, and knowledge management*, Springer, 2016, 360 p., DOI 10.1007/978-3-319-39056-7.
8. Willett P., "The Porter stemming algorithm: then and now", In: *Program: Electronic Library and Information Systems*, 2006, pp. 219-223.
9. Huilong Fan, Yongbin Qin., "Research on Text Classification Based on Improved TF-IDF Algorithm", *The International Conference on Network, Communication, Computer Engineering*, 2018, pp. 501-506.
10. Sidorova E. "Multipurpose dictionary subsystem for extraction of subject lexicon", *Proc. International conference Computer linguistics and intellectual technologies (Dialogue'2008)*, Moscow, 2008, pp. 475-481.
11. Kononenko I.S., Sidorova E.A. "Language Resources in Ontology-Driven Information Systems", *Proc. First Russia and Pacific Conference on Computer Technology and Applications*, 6-9 September, Vladivostok, Russia, 2010, pp. 18-23.
12. Kopaygorodskiy A.N. "Knowledge management via intellectual collective expert environment", *Proceeding of the 18th Workshop on Computer Science and Information Technologies (CSIT'2016)*, Vol. 1. USATU, Ufa, Russia, 2016, pp. 65-69.
13. Kopaygorodsky A.N. "Virtual integration of distributed data of energy research", *Proceedings of the XIII Russian conference of the Baikal Information and mathematical technologies in science and management*, ESI, Irkutsk, Russia, 2008, Vol. 3, pp. 260-266.
14. Kopaygorodsky A.N. "Design and implementation of a graphical modeling system", *Proceedings of the XV Russian conference of the Baikal Information and mathematical technologies in science and management*, ESI, Irkutsk, Russia, 2010, pp. 22-28.

THE EFFICIENCY COMPARISON OF THE PRISM AND STORM PROBABILISTIC MODEL CHECKERS FOR ERROR PROPAGATION ANALYSIS TASKS

T. Fabarisov¹, N. Yusupova¹, K. Ding, A. Morozov², K. Janschek²

Faculty of Computer Science and Robotics, Ufa State Aviation Technical University Ufa, Russia¹

Technische Universität Dresden Germany²

flatagir@gmail.com, yussupova@ugatu.ac.ru, kai.ding@tu-dresden.de, andrey.morozov@tu-dresden.de, klaus.janschek@tu-dresden.de

Abstract: Dual-graph Error Propagation Model (DEPM) is a stochastic framework developed in our lab that captures system properties relevant to error propagation processes such as control and data flow structures and reliability characteristics of single components. The DEPM helps to estimate the impact of a fault of a particular component on the overall system reliability. The probability of a system failure, Mean Time Between Failures and Mean Time to Repair, and the expected number of failures are the quantitative results of the DEPM-based analysis. In addition, the DEPM provides an access to the powerful techniques from Probabilistic Model Checking (PMC) that allow the evaluation of advanced, highly customizable, time-related reliability properties, e.g., “the probability of a system failure during certain time units after the occurrence of a certain combination of errors in certain system components”.

For this reason, the DEPM models are automatically transformed to discrete-time Markov chain (DTMC) models and are analyzed using the state of the art probabilistic model checker PRISM. However, the new promising model checker Storm has been recently released.

This paper presents the results of the efficiency comparison of these two model checkers based on the automatically generated set of the DTMC models that are typical for the error propagation analysis tasks. Several computation engines that are supported by both model checkers have been taking into account. The paper also gives the general description of the DEPM workflow with the focus on the PMC interface.

KEYWORDS: ERROR PROPAGATION MODEL, RELIABILITY, SAFETY, DEPENDABILITY, MODEL-BASED SYSTEMS, MODEL-BASED ANALYSIS, PROBABILISTIC MODEL CHECKING, CONTROL FLOW, DATA FLOW, OPTIMIZATION.

1. Introduction

Evaluation of the system reliability characteristics is an important part of the safety analysis which helps to assess the degree of dependability, performance, and safety assurance of the future system and helps to determine whether or not system (or system unit) will meet the qualitative requirements, e.g., “the expected mean time between failures for a given scenario of particular system part” or “the minimum time before the battery leakage given certain operating temperature”.

Estimation of these characteristics allows safety engineers to identify hazards, analyze and assess safety risks prior to implementation of the system design. The safety assessment must be evaluated to ensure that safety requirements have been met. The general goal is to minimize the hazard-related risks in safety-critical industrial domains, where a failure or malfunction may result in environmental harm, severe equipment damage or even serious injuries of the people.

2. Probabilistic Model Checking and DEPM

Fault likelihood of the system components and overall system behavior in the context of error propagation can be expressed in formalized metrics – reliability properties.

These metrics can be divided into three categories [1]:

- quantitative reachability: the probability of reaching some state of the system model, e.g., “probability that more than 25 percent of outputs are erroneous”;
- qualitative reachability: whether the probability of reaching some set of states of the system model is 0 or 1, e.g., “probability that a leader is eventually elected”;
- expected reachability: expected accumulated reward or cost to reach some set of states of the system model, e.g., “expected number of steps until termination”.

The Dual-graph Error Propagation Model (DEPM) is a mathematical abstraction [2, 5] of the system properties such as control and data flow structures and reliability characteristics of single components, which are vital for the determination of the

error propagation processes. It is a useful analytical tool for assessing the impact of specific errors and errors on the overall behavior of the system. DEPM-based analysis is used to assess the reliability properties of a system in the context of error propagation. DEPM is represented by the set-based mathematical notation:

$$\text{DEPM} := \langle E, D, A_{CF}, A_{DF}, C \rangle$$

Where E is a non-empty set of elements; D is a set of data storages; A_{CF} is a set of directed control flow arcs, extended with control flow decision probabilities; A_{DF} is a set of directed data flow arcs; C is a set of conditions of the elements.

The OpenErrorPro [3, 4] is an analytical software developed in our lab that supports the system analysis with the DEPM. On the input there are baseline models that describe the target system. Several parsers transform the baseline models into the DEPM models [5]. A screenshot of the OpenErrorPro is shown in Figure 1.

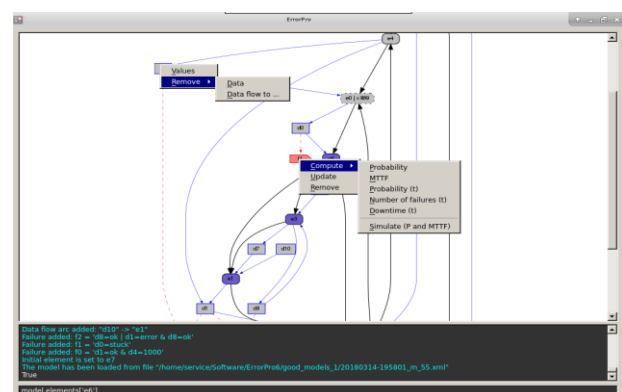


Fig 1. A screenshot of the OpenErrorPro

Reliability properties can be used for evaluation of such system characteristics as Mean Time to Failure (MTTF), Mean Time Between Failure (MTBF), Failure in Time (FIT), Mean Time to Repair (MTTR), and etc [6]. These reliability metrics refer to the system itself or to a specific component. Despite the fact that MTBF remains the main metric of system reliability for most cases, it is often being considered as no more relevant and, in general, widely incomprehensible. However, it is still a useful metric in consideration of its restrictions. For that reason, the DEPM also

provides an access to the powerful techniques from the Probabilistic Model Checking (PMC) [7, 8].

These techniques allow the evaluation of advanced, highly customizable, time-related reliability properties, e.g., “the probability of a system failure during certain time units after the occurrence of a certain combination of errors in certain system components”.

In order to calculate probabilistic properties DEPM models must be transformed to discrete-time Markov chain models (DTMC). In the DTMC models each state of the system has a probability distribution to its successor states. Whenever the system is in a certain state, a type of random experiment is conducted to determine the state to which the system moves at the next step. The number of steps that the system takes is counted by non-negative integers, therefore, in the DTMC models time is discrete. DTMC models are described in a PRISM language. PRISM model checker [9] is a powerful stochastic probabilistic model analyser. It uses its own notation to describe DTMC models and PRISM's property specification language to describe reliability properties of the system. A good example of PRISM model and properties is given in [10]. Generated DTMC models consist of the initial state; of the set of states to which model can move to; of the rewards: values associated with certain states of the model; of the labels: sets of states that are of particular interest. The reliability properties can be calculated for DTMC model.

3. Storm model checker

Storm [11] is a tool for the analysis of systems involving random or probabilistic phenomena. Given an input model and a quantitative specification, it can determine whether the input model conforms to the specification. It has been designed with performance and modularity in mind. The tool has been developed at the Chair for Software Modeling and Verification at RWTH Aachen University.

Storm is built around discrete- and continuous-time Markov models:

- Discrete-Time Markov Chains
- Markov Decision Processes
- Continuous-Time Markov Chains
- Markov Automata [12]

According to the author's website, Storm model checker has several characteristics that made it a promising tools, which has the efficient computation core, modularity of the modules, support of various input languages, and the python interface. Because the properties of the analyzing systems are very different Storm has several engines that determine which data structures are used to build and analyze the model accordingly. Supporting of the several input modeling languages helps to overcome the problem that occurs when the target model is already created by another tool and cannot be transferred to another toolset without translation [13].

In order to help a user to abstract from some of core code and focus on the computation algorithms, the developers of the Storm had published a Stormpy python interface, which provides a python API to the main functionality of the Storm code. For the efficient benchmarking of Storm model checker, we used Stormpy – python API for Storm. It allows to check models and verify properties right in the python environment [14]. On the input Stormpy requires a path to the file with a symbolic description of the model in the PRISM language format and a string that contains properties in the prism-property format. An example of the code is shown in Listing 1. An example of the symbolic description of the simple model in the PRISM language format is shown in Listing 2.

```
path = "/path/to/model/"
prism_program = stormpy.parse_prism_program(path)
formula_str = "P=? [F s=2]"
```

```
properties = stormpy.parse_properties_for_prism_program(
    formula_str, prism_program)
model = stormpy.build_model(prism_program, properties)
result = stormpy.model_checking(model, properties[0])
print(result)
```

Listing 1. An example of the Stormpy usage

```
dtmc
module die
s : [0..7] init 0;
d : [0..6] init 0;
[] s=0 -> 0.5 : (s'=1) + 0.5 : (s'=2);
[] s=1 -> 0.5 : (s'=3) + 0.5 : (s'=4);
[] s=2 -> 0.5 : (s'=5) + 0.5 : (s'=6);
[] s=3 -> 0.5 : (s'=1) + 0.5 : (s'=7) & (d'=1);
[] s=4 -> 0.5 : (s'=7) & (d'=2) + 0.5 : (s'=7) & (d'=3);
[] s=5 -> 0.5 : (s'=7) & (d'=4) + 0.5 : (s'=7) & (d'=5);
[] s=6 -> 0.5 : (s'=2) + 0.5 : (s'=7) & (d'=6);
[] s=7 -> (s'=7);
endmodule
rewards "coin_flips"
[] s<7 : 1;
endrewards
```

Listing 2. An example of the analyzed model

It has been tested which one of two model checkers is more efficient. In order to test time efficiency of Storm and PRISM model checkers, two scripts have been written. For the testing, 2464 PRISM models have been generated. For every model two properties have been checked. Every model has been checked four times. Therefore, for every model we got 8 checked properties. Mean time of property analysis per every model has been calculated. The snippet of results and the total result of the comparison are presented in the Table 1. The comparison has shown that Storm takes 0,96 seconds to analyze one property, PRISM analyzes one property in 2,64 seconds which is almost three times more than Storm. Despite the impressive results and promising novelty of Storm, there are several reasons that restrict us from implementing Storm into the EroPro.

Storm is not capable of computing Steady-state and Transient probabilities. In PRISM if the model is a CTMC or DTMC, it is possible to compute corresponding vectors of steady-state or transient probabilities directly (rather than indirectly by analyzing a property which requires their computation) [15]. Because the Storm has many different input formats and it depends on the format how this distribution would have to be specified. PRISM has an advantage in the sense that it supports only one input format.

Table 1. Results of the Storm and PRISM comparison (seconds)

Model	Mean_time	Mean_time
	STORM	PRISM
m_1209.xml_prism.pm	1,1129	2,5478
m_1208.xml_prism.pm	0,6677	2,6055
m_1207.xml_prism.pm	1,2605	3,2407
...	...	...
m_2.xml_prism.pm	1,2075	2,6420
m_55.xml_prism.pm	1,1163	2,8417
m_33.xml_prism.pm	0,7144	2,3773
Min	0,64	2,27
Max	1,98	7,78
Mean	0,96	2,64

Table 2. Results of the different PRISM engines comparison (seconds)

Model	Mean time (explicit)	Mean time (MTBDD)	Mean time (sparse)	Mean time (hybrid)
<i>m_47.pm</i>	2,4231	2,3996	2,4512	2,4499
<i>m_32.pm</i>	2,4276	2,4223	2,4540	2,3831
<i>m_38.pm</i>	2,4608	2,3630	2,3937	2,3855
...	...	...	...	...
<i>m_1096.p</i>	2,4816	2,3926	2,4845	2,4879
<i>m</i>				
<i>m_1200.p</i>	2,3484	2,3605	2,4053	2,4523
<i>m</i>				
<i>m_1261.p</i>	2,7282	19,3512	2,5472	2,4892
<i>m</i>				
<i>Mix</i>	2,3026	2,3230	2,3254	2,3595
<i>Min</i>	2,8577	108,3340	3,7128	3,6529
<i>Mean</i>	2,4819	5,7274	2,5179	2,5190

Storm has to load a model from file. In order to parse a PRISM model, one has to give a path to the file with a symbolic description of the model to the `stormpy.parse_prism_program()`. But it is not possible to parse a PRISM model directly from the string. Though it has to be said that PRISM requires additional parsing of the results.

In addition to Storm vs PRISM comparison, the comparison between different types of PRISM modes was also carried out: explicit, MTBDD, sparse, hybrid and exact. Unfortunately, exact mode didn't work and every time a following type of error showed: "Error: Probabilities sum to 49999995999999/5000000000000 instead of 1". For the testing 100 models have been generated. For every model two properties have been checked. Every model has been checked four times. Therefore, for every model we got 8 checked properties. Mean time of property analysis per every model has been calculated. The snippet of results and a total result of the comparison is presented in Table 2. The comparison has shown that explicit mode of PRISM is the fastest one. The hybrid engine is enabled by default in PRISM [16]. It uses a combination of symbolic and explicit-state data structures (as used in the MTBDD and sparse engines, respectively). The explicit engine is similar to the sparse engine, in that it can be a good option for relatively small models, but will not scale up to some of the models that can be handled by the hybrid or MTBDD engines. Therefore, it is recommended to continue using the hybrid mode of PRISM.

4. Conclusion

An importance of the risk assessment in the systems has been discussed in this article. General description of the DEPM workflow with the focus on the PMC interface has been presented and various reliability metrics along with the example of the advanced reliability properties had been shown. Usage of the PRISM model checker in the context of reliability characteristics evaluation has been illustrated. Storm model checker has been reviewed. In order to evaluate its efficiency and assess the necessity of using Storm instead of PRISM, a set of experiments has been conducted. The article has presented the results of the efficiency comparison of these two model checkers based on the automatically generated set of the DTMC models. Several computation engines that are supported by both model checkers have been taken into account. The review of the Storm model checker has shown that Storm does not support some of the PRISM features, such as: computing of the steady-state and transient probabilities, probabilistic timed automata, and multi-objective model checking. The effectiveness comparison has shown that Storm model checker is almost three times faster than PRISM. But due to the abovementioned restrictions of its functionality, it is not worth to be implemented to OpenErrorPro yet. At least, until these functions will be developed by the authors.

Acknowledgment

This work has been supported by Erasmus+ EU funds.

References

1. Maria Svorenova and Marta Kwiatkowska. Quantitative Verification and Strategy Synthesis for Stochastic Games. European Journal of Control, Elsevier, 2016.
2. A. Morozov and K. Janschek. Dual graph error propagation model for mechatronic system analysis. In 18th IFAC World Congress, Milano, Italy, pages 9893–9898, 2011.
3. A. Morozov, R. Tuk, and K. Janschek. ErrorPro: Software Tool for Stochastic Error Propagation Analysis. In 1st International Workshop on Resiliency in Embedded Electronic Systems, Amsterdam, The Netherlands, pages 59–60, 2015.
4. Fabarisov, T., Yussupova, N., Ding, K., Morozov, A., Janschek, K.: Analytical Toolset for Model-based Stochastic Error Propagation Analysis: Extension and Optimization Towards Industrial Requirements. Proceedings of the 19th international workshop on computer science and information technologies, Germany, Baden-Baden, 2017.
5. A. Morozov and K. Janschek. Probabilistic error propagation model for mechatronic systems. Mechatronics, 24(8):1189 – 1202, 2014.
6. IMC Networks: MTBF, MTTR, MTTF & FIT Explanation of Terms, Aarschot, Belgium, 2011.
7. Marta Kwiatkowska, Gethin Norman, Dave Parker Dave Parker. Probabilistic Model Checking Probabilistic Model Checking, University of Oxford.
8. Vojtěch Forejt, Marta Kwiatkowska, Gethin Norman and David Parker. Automated Verification Techniques for Probabilistic Systems. In M. Bernardo and V. Issarny (editors), Formal Methods for Eternal Networked Software Systems (SFM'11), volume 6659 of LNCS, pages 53–113, Springer. June 2011.
9. Marta Kwiatkowska, Gethin Norman and David Parker. PRISM 4.0: Verification of Probabilistic Real-time Systems. In Proc. 23rd International Conference on Computer Aided Verification (CAV'11), volume 6806 of LNCS, pages 585–591, Springer, 2011.
10. Marta Kwiatkowska, Gethin Norman and David Parker. Probabilistic Verification of Herman's Self-Stabilisation Algorithm. Formal Aspects of Computing, 24(4), pages 661–670, Springer. July 2012.
11. Christian Dehnert, Sebastian Junges, Joost-Pieter Katoen, Matthias Volk. A Storm is Coming: A Modern Probabilistic Model Checker. Proc. of CAV, Volume 10427 of LNCS, pages 592–600, Springer, 2017.
12. Yuxin Deng, Matthew Hennessy, On the semantics of Markov automata, Information and Computation, Volume 222, 2013, Pages 139–168, ISSN 0890-5401.
13. Tim Quatmann, Christian Dehnert, Nils Jansen, Sebastian Junges, Joost-Pieter Katoen. Parameter Synthesis for Markov Models: Faster Than Ever. Proc. of ATVA, Volume 9938 of LNCS, pages 50–67, Springer, 2016.
14. Christian Dehnert, Sebastian Junges, Joost-Pieter Katoen, Matthias Volk. The Probabilistic Model Checker Storm (Extended Abstract). CoRR abs/1610.08713, 2016.
15. Gethin Norman and David Parker. Quantitative Verification: Formal Guarantees for Timeliness, Reliability and Performance. A report by the London Mathematical Society and the Smith Institute. Edited by Robert Leese and Tom Melham. 2014.
16. Marta Kwiatkowska, Gethin Norman and David Parker. Advances and Challenges of Probabilistic Model Checking. In Proc. 48th Annual Allerton Conference on Communication, Control and Computing, pages 1691–1698, IEEE Press. Invited paper. October 2010.

MODEL REDUCTION ALGORITHM FOR FAST NEUTRALITY TESTS AND FAULT LOCALIZATION OF SIMULINK MODELS

X. Hu, A. Morozov¹, K. Ding, K. Janschek¹, N. Yusupova²

Technische Universität Dresden, Germany¹

Faculty of Computer Science and Robotics, Ufa State Aviation Technical University, Russia²

xinrui.hu tu-dresden.de, andrey.morozov tu-dresden.de, kai.ding tu-dresden.de, klaus.janschek@tu-dresden.de, yussupova@ugatu.ac.ru

Abstract: A minor change of a Simulink model can result in an unexpected consequence, so the Simulink model is usually required to be rerun and tested, which increases the development cost and time. Compared with the reference model, only the changed parts of the updated model could result in a failure at the outputs. So, a two-stage model reduction algorithm is designed to isolate the changed parts, that speeds up the processes of neutrality test and fault localization. The first reduction is based on the changed parts, the second reduction is based on the bad outputs. The changed parts and the bad outputs are the blocks of interest of the reduction. The blocks related to the blocks of interest are reserved, the others are deleted. The thesis proposes a way of conversion of the Simulink model to a digraph based on extended data dependence to find the related blocks. After the model reduction, the faults are located with the help of signal comparison.

1. Introduction

The methodology of Model-Based Development is widely used in modern mechatronic and cyber-physical systems of higher heterogeneity and larger scale. One of the world-wide recognized tools is MATLAB/Simulink, a common tool for the system- and component-level design and simulation, automatic code generation and its deployment on target hardware [13], which allows the engineer to verify it early in the life cycle.

Even after minor changes, the reliability standards require that the Simulink model should be tested to ensure no fault is introduced. One of the model test activities is the fault localization. The commonly used method, regression test, is often used to ensure that there aren't any newly introduced faults after the update of the model. It often requires rerunning the whole model and the whole test cases, which consumes an amount of time. In order to decrease the costs of the fault localization, the researchers have investigated a lot of useful ways like statistical debugging [1, 2, 7, 10], Simulink Design Verifier [26], Falsification-based testing using Temporal Logic [3, 5].

However, it is assumed that only the changed blocks of the Simulink model may be faulty. In contrast with rerunning the entire updated model, the thesis designs the two-stage of model reduction algorithm to reduce the scale of the models and the faults are located based on the reduced models.

2. State of the Art

Simulink/Simulink API: As the world-wide most recognized tool for model-based system design, MATLAB/Simulink [13] is widely used in industrial application. Simulink and Simulink Model-Based Design aim at helping to implement the system with high quality, especially for the extremely complicated systems. It allows developers to design, analyze and simulate the models before moving to the real hardware world. After the model test, the codes, e.g. C, C++, can be generated from those verified models automatically and deployed on target hardware. The Simulink model is a network of blocks of different types. By drawing blocks and lines, the designer can have experience of intuitive graphical design. The configuration of the blocks and models can be implemented manually by a user interface or programmatically by Simulink API. Simulink provides different types of solvers [16], such as fixed-step and variable-step solvers, to simulate dynamic systems for different requirements.

Simulink model reduction: Simulink Model Slicing is widely used in many approaches of faults localization [5, 9]. Ways of model slicing can be basically classified into two groups: the

official tool - Simulink Model Slicer provided by Simulink [14], which also allows users to get access programmatically using Simulink API, and the unofficial tool based on dependence analysis [12]. The former isolates problematic behavior in a model and allows the user to highlight and trace ports, signals, and blocks, and then slices a large model into smaller models for further analysis [14]. The latter slices the Simulink models by using a dependence graph, which is calculated by analysis of the data dependence and the control dependence, so that the models can be reduced for the given blocks of interest.

Fault localization for Simulink model: So far there are lots of investigation to fault localization for Simulink models. They can be roughly classified into two categories: official tool from Simulink - Simulink Design Verifier (SDV) [14] and unofficial tools. The unofficial tools are mainly developed based on the statistical debugging method and falsification testing using temporal logic. Statistical debugging is a well-studied and wide-used debugging technique in software engineering domain [1, 2, 7]. The statistical debugging is extended to fault localization for the Simulink model and search-based test suite generation algorithm are proposed in [9].

The other tool is based on Signal Temporal Logic (STL). STL is originally proposed for hardware circuits. With growing interest in model checking for more sophisticated fields such as analog/digital mixed signal circuits, cyber-physical systems this approach has been also widely used in fault localization for Simulink/Stateflow models [5]. It uses execution-context-based model slicing [12] and finds the faults that result in a violation of STL at the outputs of the sliced model.

SDV is a powerful product of Simulink, which includes abundant products such as Simulink Requirements, Simulink Check, and Simulink Coverage, aiming at generating test cases based on design requirements, detecting basic modeling errors such as dead logic, integer, and fixed-point overflows, division by zero, etc. With help of Model Slicer tool in SDV, the potentially problematic parts of a time interval of interest will be isolated. But basically, the user should define an interval of interest manually. Usually after the first model slicing the experienced user need to adjust the interval in order to find the problematic position more precisely. This is hence an iterative process.

Neutrality test and regression test: Inspired by the neutrality theory of biology domain, modifications or changes which do not affect the system are neutral. Neutrality testing aims at determining whether changes in model interfere the parts of interest by using neutrality as a null hypothesis [6]. In the software engineering domain, even a small tweak can result in unexpected consequences. The regression test is a test method to make sure that changes don't

cause unintended effects. Usually, regression test requires rerunning test cases and check if violations at observed position appear. Due to the large scale of test suits regression test is a quite time-consuming progress. So many techniques have been proposed to make the regression test more cost-efficient. They can be mainly classified into three groups [17]: test suite minimization [15], test case selection [11] and test case prioritization [8].

3. Design and Implementation

In order to reduce the validation time of an updated or modified model, the model is reduced two times. After two reductions, the faults will be located. The first stage of model reduction is based on the changed parts and the blocks related to the changed parts. After the first reduction, the models are executed to find the bad outputs. The second reduction is based on the bad outputs and the blocks related to the bad outputs. The conversion of the Simulink model to the digraph is designed. With the help of the digraph and the corresponding adjacency matrix, signals are traced and the related blocks are found. After two reductions, the scale of the model is smaller. Based on signal comparisons of the reserved blocks, faults are located. Figure 1 shows the algorithm of the two-stage model reduction for fault localization.

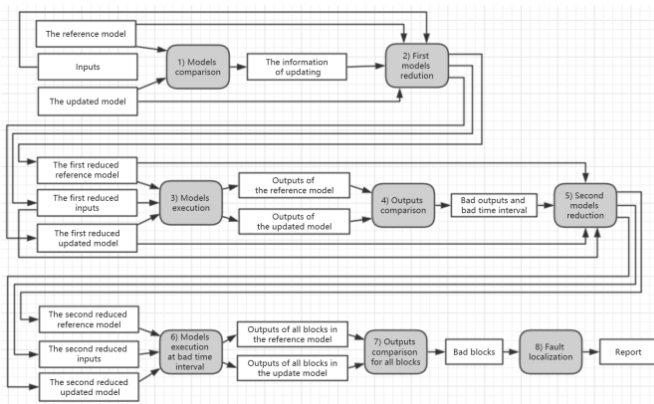


Fig. 1 Algorithm of the model reduction and fault localization

3. 1. Signal Tracing and Model Reduction

a) *Conversion of the Simulink model to the digraph*: The two model reductions have the similar idea, i.e. to reduce the models to the given blocks of interest. The blocks related to the blocks of interest are found and reserved, and the other blocks and redundant lines are deleted. To find the related blocks, the Simulink model is first converted to the digraph. It is defined that, the nodes in the Simulink model is a vertex in the digraph. The signal transmission is an arc in the digraph. To make sure that the models can be executed normally, the signal transmission is defined as the follows: i) The directed lines in the Simulink transmit signals from the start to end of the line. ii) The *Goto* block transmits signals to *From* block with the same *GotoTag*. iii) The blocks, connected with the condition port of the conditionally execution subsystems (the *triggered*, *enabled*, and *action* subsystem), transmits signals to the *trigger*, *enable*, and *action* port inside the subsystems. iv) The *trigger*, *enable*, and *action* port transmit signals to all blocks inside the conditionally execution subsystems.

In graph theory, directed graph (or just digraph) D consists of a non-empty finite set $V(D)$ and finite set $A(D)$, where [4],

- $V(D)$ (vertex set) is a finite set of vertices. It is denoted as $V(D) = \{v_1, v_2, \dots, v_n\}$.
- $A(D)$ (arc set) is a finite set of ordered pairs of distinct vertices namely arcs. It is denoted as $A(D) := \{(v_i, v_j) | v_i, v_j \in V(D)\}$.

An example in figure 2 shows the conversion of the Simulink model to the digraph.

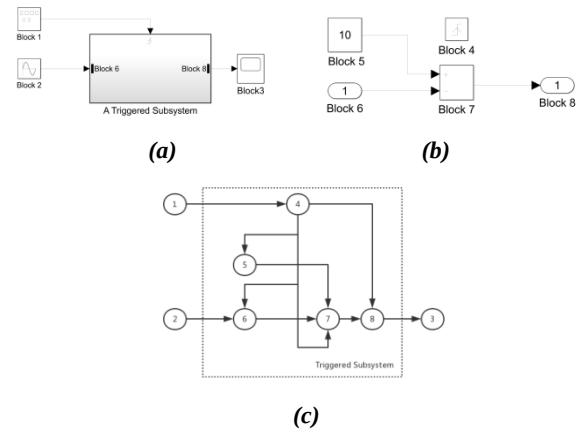


Fig. 2 An example of the transformation of the Simulink model to digraph: (a) Simulink model with triggered subsystem (b) Internal structure of triggered subsystem (c) Digraph of the Simulink model

b) *Signal tracing based on the digraph*: As explained, the nodes in the digraph represent the blocks in the Simulink model. In order to find the related blocks of the block interest, the adjacency matrix of the digraph is used. Based on the digraph, the related nodes of the node of the interest are defined in three groups:

1. The nodes that are backward-connected with the node of interest.
2. The blocks that are forward-connected with the block of interest.
3. The blocks that are backward-connected with the blocks from the group 2.

Figure 3 shows an example of a digraph of a Simulink model. Assume that the node 4 is the node of interest. The related nodes from group 1 are nodes 1, 2, 3; the related nodes from group 2 are nodes 5, 8; the related node from group 3 is node 6. Only node 7 is not related to node 4. After the model reduction, node 7 is deleted.

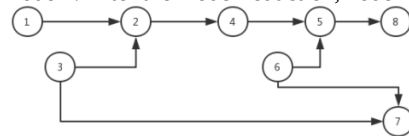


Fig. 3 An example digraph for signal tracing

3. 2. Signal Comparison

Simulink provides several ways to get the simulation data, such as *Simulink.sdi.sim*. The former provides an interface to get access to Simulation Data Inspector [18] programmatically. It helps to access Run and Signal objects in Simulation Data Inspector by using functions *Simulink.sdi.getRun* and *Simulink.sdi.getSignal* respectively. However, the inputs of the functions are *runID* and *sigID*, meaning the ID number of the signal object and the simulation run object, which make it quite difficult to distinguish a *sigID* represents which signal object and a *runID* represents which run object. The latter is a function to simulate a model programmatically. Using *sim* makes it possible to specify parameter name-value pairs, to configure simulation, to access simulation metadata, etc.. The parameter name-value pairs can be specified before simulation, meaning that the signals, which will be compared, can be pre-set names so that they can be accessed by names. Based on this reason, *sim* will be used for getting the simulation data. *Simulink.SimulationData.Signal* stores simulation data, it has a *Values* property. The *Values* field represents a timeseries type of variable. It contains data variable and the corresponding time variable. Signal comparison is based on these two variables. Two signals will be considered as "bad" if any one of the situations occurs: a) time variables aren't the same, b) time variables are the same, but the data variables differ.

3. 3. Fault Localization

Based on the signal comparison, the models will be run only at the "bad" sample time intervals, which saves a lot of time. It might happen that some "bad" time intervals are too long, the upper limits of "bad" time interval will be given. The inputs of the reference and updated models will also be reduced to the given "bad" time interval. However, in a Simulink model, there are some delay. For example, assume the simulation step time is T , if there is a block b_2 which receive signal from block b_1 , b_2 is a delay block, which delays the signal e.g. $1T$. So if b_1 outputs a fault at time point kT , b_2 will output the fault however at the time point $(k+1)T$. The outputs of the blocks forward-connected with b_2 will output the fault at $(k+1)T$ or even later. In this case, the bad time interval shouldn't start at time point kT , otherwise, it could happen that after reduction of the inputs, the fault will disappear. So, the margin time should be taken into account when generating the bad time intervals. Noted that the calculation process of the Simulink block doesn't use the value at the future point, the detected error at the output is only dependent on the current time point or earlier time points, thus, the margin time should be added at the front of the bad time interval.

After finding the bad time intervals, the model will be run to perform fault localization. Each non-subsystem blocks in Simulink model is a system element, in our thesis called blocks b . Each element b_i has inputs and outputs, defined as [17]:

- $I_{b_i} := \{i_1, i_2, \dots, i_{N_{b_i}^{in}}\}$,
- $O_{b_i} := \{o_1, o_2, \dots, o_{N_{b_i}^{out}}\}$

, where I_{b_i} and O_{b_i} are the sets of inputs and outputs of the block b_i , $N_{b_i}^{in}$ and $N_{b_i}^{out}$ are numbers of inputs and outputs of b_i . Then the faulty block b_f satisfies the following criteria:

- It is a related block of the bad outputs.
- If it has inputs, and all the inputs are correct (equal to the reference), but not every output is correct.
- If it doesn't have input and not every output is correct.
- If it is an added or deleted block, it is considered as potentially faulty.

Figure 4 shows the example of the faulty blocks. The red colored lines represent the incorrect signals, the black colored are correct.

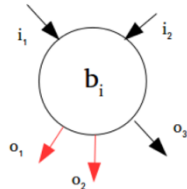


Fig. 4 An example of a faulty block

In Simulink, it's not allowed to log signals, such as *action* signals, signals that feed into a *Merge* block, *function-call* signals, or *sate* signals. This thesis doesn't handle with the models containing *function-call* signals and *sate* signals. Also, there are some blocks don't have real data signal lines, and some blocks' lines are control signal lines.

For example:

1. *Inports* in a subsystem don't have inputs signal lines;
2. *Outports* in a subsystem don't have outputs signal lines;
3. If and *SwitchCase* blocks only have control or more explicitly action output signals.
4. The *Action* ports in *Action* subsystems have neither inputs nor outputs.
5. The *Trigger* ports in *Triggered* or *Function-Call* subsystems have neither inputs nor outputs.
6. The *Enable* ports in *Enabled* subsystems have neither inputs nor outputs.

The ports mentioned in situations 4, 5, and 6 accept the signals

to start the execution of conditional execution context [12]. The process is done by the internal schema of Simulink, it can be considered as always correct. So, these special ports won't be checked. The outputs of *If* and *SwitchCase* are *action* signals so that their outputs can't be logged directly. However, the Action subsystems which are predicted by them can give some information. By logging the output value of *Inports* blocks, the activation time points of the subsystems can be obtained. It gives information about if an action signal is given. So, *If* or *SwitchCase* blocks are blocks,

- whose inputs are all correct (equal to the reference),
- and the activation time points of the subsystems connected with its *outports* are incorrect.

4. Results and Discussion

To validate the program, a model with lookup tables in figure 5 is chosen. Since one of the inputs of the 2-D lookup tables is constant, which means some breakpoints may not be used. One breakpoint of *Pumping Constant*, which is not used, is modified. One breakpoint of *Ramp Rate Ki*, which is used, is modified. In addition, two blocks are deleted in the updated model.

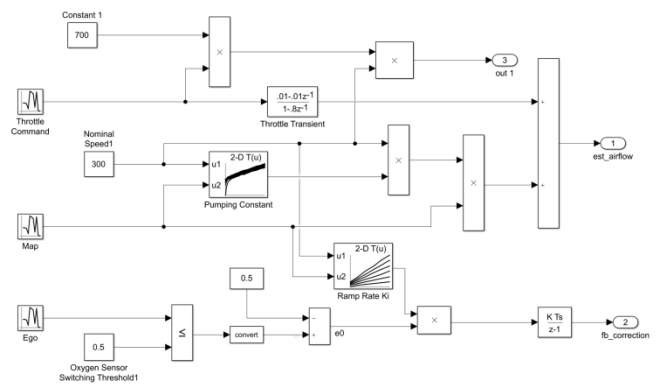


Fig. 5 A validation model with lookup tables

After the execution of the program the faults are located, and the reduced model is in figure 6, the report of fault localization is in figure 7. Table 1 shows the number of the blocks after each model reduction.

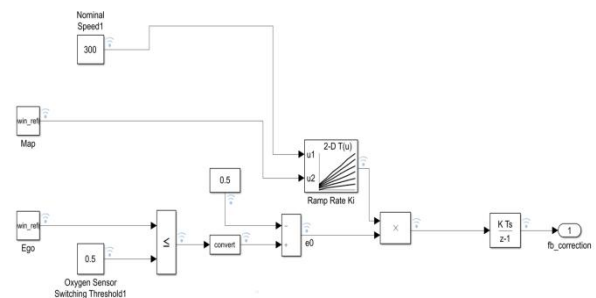


Fig. 6 The reference model after two reductions

```
>> fault_localization
The faults may locate at the following blocks:
1. RampRateKi
2. deleted Sum1
3. deleted Constant
```

Fig. 7 The report of fault localization

From this example, it can be seen that, the models are first reduced to the changed parts. After the first reduction, the models are run. The first output *est_airflow* of the model only receives the signal from *Pumping Constant*, the second output *fb_connection* only receives the signal from other changed parts. However, the changed breakpoint in the lookup table *Pumping Constant* is never used, no error is produced and propagated to the first output, so the first output is correct. The second reduction is based only on the second output. After two reductions, the scale of

the model is obviously reduced. The deleted blocks are related to the bad outputs, so they are considered as potentially faulty.

Table 1: Numbers of blocks after each model reduction

Stage	Number of blocks in the reference model	Number of blocks in the updated model
Before reduction	23	21
After first reduction	19	17
After second reduction	12	10

5. Conclusion

Since the Simulink model is modelled by drawing the visible and intuitive lines and blocks, so the idea of doing fault localization of updated models based on signal tracing and signal comparison comes up naturally. In addition, faults can only be located at the positions, where they are not the same as a previously validated model. So, the updated model will be compared with the previously validated model, or the reference model. Hence, the neutrality testing based on model reduction is proposed.

The work consists of three main tasks: model reduction, signal tracing, and signal comparison. A controller model contains normally two parts: control logic and data signals. The latter can be analyzed along the lines in the model. The former however is an internal mechanism, which isn't so intuitive. The built-in Simulink Model Comparison tool provides a way of model comparison. However, it doesn't give information about which blocks use the modified variables in the model workspace, which should be solved additionally. After model comparison, the model is reduced to a new model with only those blocks which are relevant to the modified blocks. Hence, an approach of transforming of the Simulink model to digraph is proposed to solve the problem of signal tracing and finding the related blocks.

Furthermore, in order to do fault localization, the signal tracing and signal comparison are needed to find which blocks are guilty for the detected bad outputs of a model. With the help of signal comparison and the adjacency matrix of the digraph. The fault blocks can be found.

However, the project has some limitations that will be solved further. i) The comparison of the two models is based on the names of the blocks. If the names of blocks are changed or added with an unintentional whitespace symbol, then they will be reported as faulty blocks. ii) If there are faults in two cascaded connected blocks, each execution of the program can detect only the first faulty block. However, if the program is executed iteratively, the second faulty block can be found. iii) If several faulty blocks result in bad signals at the same output one after the other, then only the block, which produces the first error, will be detected. This can be also solved by the interactive execution of the program. iv) The program can't locate faults for the model with S-Function and Stateflow Chart.

Reference

[1] James A. Jones and Mary Harrold. "Empirical evaluation of the Tarantula automatic fault-localization technique". In: 20th IEEE/ACM International Conference on Automated Software Engineering, ASE 2005. Jan. 2005, pp. 273-282.

[2] R. Abreu, P. Zoetewij, and A. J. C. van Gemund. "On the Accuracy of Spectrum-based Fault Localization". In: Testing: Academic and Industrial Conference Practice and Research Techniques - MUTATION (TAICPART-MUTATION 2007). Sept. 2007, pp. 89-98.

[3] Yashwanth Annpureddy, Che Liu, Georgios Fainekos, and Sriram Sankaranarayanan. "S-TaLiRo: A Tool for Temporal Logic Falsification for Hybrid Systems". In: Tools and Algorithms for the Construction and Analysis of Systems. Ed. by Parosh Aziz Abdulla and K. Rustan M. Leino. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 254-257.

[4] Jrgen Bang-Jensen and Gregory Z. Gutin. "Basic Terminology, Notation and Results". In: Digraphs: Theory, Algorithms and Applications. Springer Publishing Company, Incorporated, 2008. Chap. 1.

[5] Ezio Bartocci, Thomas Ferrere, Niveditha Manjunath, and Dejan Nickovic. "Localizing Faults in Simulink/Stateflow Models with STL". In: Proceedings of the 21st International Conference on Hybrid Systems: Computation and Control (Part of CPS Week). HSCC '18. Porto, Portugal: ACM, 2018, pp. 197-206.

[6] Yun-xin Fu. "New Statistical Tests of Neutrality for DNA Samples From a Population". In: Genetics 143.1 (1996), pp. 557-570.

[7] J. A. Jones, M. J. Harrold, and J. Stasko. "Visualization of test information to assist fault localization". In: Proceedings of the 24th International Conference on Software Engineering. ICSE 2002. May 2002, pp. 467-477.

[8] Muhammad Khatibsyarbini, Mohd Adham Isa, Dayang N.A. Jawawi, and Rooster Tumeng. "Test case prioritization approaches in regression testing: A systematic literature review". In: Information and Software Technology 93 (2018), pp. 74-93.

[9] B. Liu, Lucia, S. Nejati, and L. C. Briand. "Improving fault localization for Simulink models using search-based testing and prediction models". In: 2017 IEEE 24th International Conference on Software Analysis, Evolution and Reengineering (SANER). Feb. 2017, pp. 359-370.

[10] Chao Liu, Xifeng Yan, Long Fei, Jiawei Han, and Samuel P. Midki. "SOBER: statistical model-based bug localization". In: ESEC/SIG-SOFT FSE. 2005.

[11] Everton Note Narciso, Marcio Delamaro, and Fatima De Lourdes Dos Santos Nunes. "Test Case Selection: A Systematic Literature Review". In: 24 (May 2014), pp. 653-676

[12] R. Reicherdt and S. Glesner. "Slicing MATLAB Simulink models". In: 2012 34th International Conference on Software Engineering (ICSE). June 2012, pp. 551-561.

[13] Simulation and Model-Based Design. <https://www.mathworks.com/products/simulink.html>. Access: 2018-08-30.

[14] Simulink Design Verifier. <https://www.mathworks.com/products/sldesignverifier.html> Accessed: 2018-08-30.

[15] Rajvir Singh and Mamta Santosh. "Test Case Minimization Techniques: A Review". In: 2 (Dec. 2013), pp. 1048-1056.

[16] Simulink Solvers. <https://www.mathworks.com/help/simulink/ug/solvers.html> Accessed: 2018-08-30.

[17] S. Yoo and M. Harman. "Regression Testing Minimization, Selection and Prioritization: A Survey". In: Softw. Test. Verif. Reliab. 22.2 (Mar.2012), pp. 67-120.

[18] Simulation Data Inspector in Your Workflow. <https://www.mathworks.com/help/simulink/ug/simulation-data-inspector-overview.html>. Accessed: 2018-07-12

USING THE FUZZY CONTROLLER TO CONTROL PROCESS PARAMETERS

E.A. Muravyova, D.S. Radakina

Department of automated technological and information systems, Ufa State Petroleum Technological University, Branch in the Sterlitamak, Sterlitamak, Russia

e-mail: muraveva_ea@mail.ru, darya.radakina@yandex.ru

Abstract: In the course of writing this article, the goal was to develop the algorithm for setting up an adaptive fuzzy controller with a double rule base. To achieve this goal, a fuzzy controller designed to control the temperature of the steam leaving the boiler. As the boiler was considered the model of boiler BKZ-75-39GMA. In the course of writing, the fuzzy controller was synthesized and adapted. On the basis of the work done, the algorithm for adaption the fuzzy controller was determined.

1. Introduction

Fuzzy controllers generate control signals based on the application of the fuzzy logic [1-6]. Over the years fuzzy controllers have been of still greater use in modern automatic control systems [2]. Fuzzy controllers can be used to control all the parameters of complex process objects. In this paper the fuzzy controller is used to control one of the parameters of process in production of superheated steam in the boiler unit BKZ-75-39 GMA. As a parameter, whose value is chosen for controlling of the valve opening gap, the boiler unit outlet steam temperature is selected. Since this parameter could be influenced by parametric interferences such as steam-generating capacity variation of the boiler unit hence steam temperature variation downstream the superheater first stage as well as the condensate temperature variation, it is then presumed as necessary for the fuzzy controller to be capable, unattended, to automatically compensate for the above interferences. In that way, the main objective in this paper is the synthesis and the fuzzy controller adaptation. The fuzzy controller adaptation is designed so that it could control automatically in a way to match parametric interferences.

2. The fuzzy control adaptation

The steam temperature toutlet steam is controlled by varying the injection flow rate of steam spray attemperator the internal condensate located between the first and second stages of the superheater.

In order to adjust the fuzzy controller adaptation one must have a clear idea of how the process works, i.e. a model should be built which could characterize the steam spray attemperator operation.

From the regression model the following equation has been obtained:

$$y=3,33+1,1*x_1-0,48333*x_2-0,81556*x_3, \quad (1)$$

where X_1 – the steam temperature downstream the first superheater stage t_{s1st} ,

X_2 – the condensate temperature t_c ;

X_3 – the valve opening gap α ;

y – steam temperature downstream the superheater second stage t_{s2st} .

The model has been obtained on the basis of 45 design points, which will determine the valve opening gap.

The fuzzy controller input will be supplied with the steam temperature t_{s1st} coming from the superheater first stage and the condensate temperature t_c . The valve opening gap α in percentage represents the fuzzy controller output variable by means of which the steam temperature is exactly controlled.

The fuzzy controller conceptual model built determines which parameter values must be put together for the fuzzy controller to function properly in the process of the expert information gathering. As a result of the above process the reference points table for the fuzzy controller was derived. This table describes the entire scope

of the fuzzy controller inputs and outputs. This table is constructed based on experimental data.

Table 1. Table of reference points

Input		Output
t_{s1st}	t_c	α
360	50	20
360	55	40
360	60	60
360	65	80
360	70	100
370	60	20
370	65	40
370	70	60
370	75	80
370	80	100
380	70	20
380	75	40
380	80	60
380	85	80
380	90	100
390	80	20
390	85	40
390	90	60
390	95	80
390	100	100
400	90	20
400	95	40
400	100	60
400	105	80

The next step in the fuzzy controller synthesis is the determination of linguistic variables inputs and outputs, which variables are described by the range of current values and by membership functions for each of the terms. A triangular shape terms are used to describe the input linguistic variables, the vertex of which is located at the reference point, the base is located between the two nearest reference points [6]. Terms applied are triangular shape because they help to obtain smooth static characteristics.

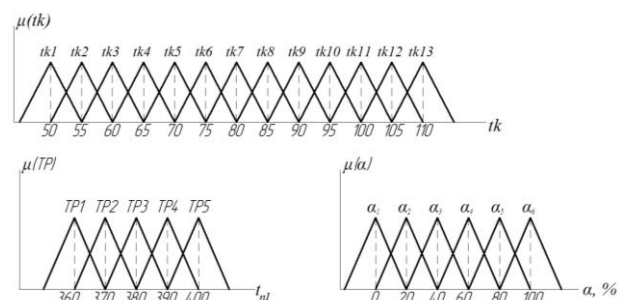


Fig 1. Linguistic variables inputs and outputs

Furthermore the production rule synthesis must be carried out by means of creating table of desired values for the fuzzy controller

output variable that is the valve opening gap α , for each value the steam temperature t_{s1st} and condensate temperature t_c term.

Table 2. The desired values of the valve opening gap α

№	t_{s1st}	t_c	α
	TP ₁	tk ₁	23
	TP ₁	tk ₂	36
	TP ₁	tk ₃	58
	TP ₁	tk ₄	82
	TP ₁	tk ₅	98
	TP ₁	tk ₆	93
	TP ₁	tk ₇	96
	TP ₁	tk ₈	99
	TP ₁	tk ₉	97
	TP ₁	tk ₁₀	94
	TP ₁	tk ₁₁	93
	TP ₁	tk ₁₂	91
	TP ₁	tk ₁₃	92
	TP ₂	tk ₁	22
	TP ₂	tk ₂	21
...	...	...	...
64	TP ₅	tk ₁₂	77
65	TP ₅	tk ₁₃	97

Production rules were constructed from table 2 (table 3). 65 production rules were obtained, this number has come out as a result of multiplying the number of the fuzzy controller terms input variables ($5 \cdot 13 = 65$), i.e. 5 values of the steam temperature downstream the attemperator first stage f t_{s1st} (360, 370, 380, 390, 400°C), 13 value of the condensation temperature t_c (50, 55, 60, 65, 70, 75, 80, 85, 90, 95, 100, 105, 110°C). These rules determine which of consequents are basic or additional, the truth degree is determined for additional consequent. Both the technique of determining consequents and their truth degrees are described in [1].

Table 3. The desired values of the valve opening gap α

№ of rules	Production rules
1	If $t_v = T_{v1}$ and $t_c = T_{c1}$ then $\alpha = \alpha_2$ and $\alpha = \alpha_3^{0.12}$
2	If $t_v = T_{v1}$ and $t_c = T_{c2}$ then $\alpha = \alpha_2^{0.17}$ and $\alpha = \alpha_3$
...	...
65	If $t_v = T_{v5}$ and $t_k = T_{c13}$ then $\alpha = \alpha_5^{0.12}$ and $\alpha = \alpha_6$

Consider the first rule: If $t_v = t_{v1}$ and $t_c = t_{c1}$ then $\alpha = \alpha_2$ and $\alpha = \alpha_3^{0.12}$. As can be seen from of this rule consequent (then $\alpha = \alpha_2$ and $\alpha = \alpha_3^{0.12}$), for α_2 the degree is equal to 1, whereas for α_3 it is equal to 0.12. This means that α_2 is the basic consequent and α_3 is the additional one.

So after looking into each of rules isolatedly, the matrix (1) is derived, compiled by the truth degrees of of the basic and additional consequential. The column number of the matrix indicates the number of the rule, the line number is the number of the term the valve opening gap α .

Adaptation will be produced for the developed fuzzy controller by means of a training algorithm.

As a training algorithm the gradient descent algorithm is selected in which as an initial point for algorithm the Mamdani fuzzy controller can be used.

If we denote as the integer N the "left" of the terms which are used in the production rule consequents, then the "right" term will have a number (N+1) [6]. The left term is a term with a lower number, that is, that in figure 1 to the left with respect to the other term of a production rule. Therefore, the right term is a term with a larger

number being located with respect to another term of the rule on the right side.

Let's introduce the designation:

$$W = N + v, \quad (2)$$

where W – the production rule characteristic, which are obtained from the truth degree values of the basic and additional consequents, i.e. on the basis of this characteristics one can return to the production rule;

N – the left term number;

v – number as determined by the formula:

$$v = \frac{C+r}{2}, \quad (3)$$

where $r = 1$, if the consequent "right" term is used, and $r = 0$ if the "left" term is applied;

C – the additional consequent truth degree.

Consider the determination of the production rule characteristics using the first rule: If $t_v = t_{v1}$ and $t_c = t_{c1}$ then $\alpha = \alpha_2$ and $\alpha = \alpha_3^{0.12}$.

The left term number is 2. The additional consequent truth degree 0.12, $r=0$, because it's the left term which is basic. Therefore, we get the production rule characteristics: $W_1 = 2.06$.

Using the formula (2) we now get the characteristics of rest production rules from table 3: $W_2 = 2.585$; ...; $W_{65} = 5.06$.

In the process of the fuzzy controller synthesis both the basic and additional consequents as well as the additional consequents truth degrees were determined [2, 3]. In determining the additional consequents truth degree the reference point were take as basic (table 1) which represent the maxmums of the input linguistic variables terms (figures 1) [2, 3]. It is to be understood that in practice, the input parameters can take different values, therefore, the fuzzy controller adaptation must be carried out to allow the controller to be automatically adjusted to match parametric interferences. On the basis of experimental data the model (1) has been developed which characterizes the steam spray attemperator operation. The model as developed is conducive to the fuzzy controller synthesis, the production rules were obtained with their characteristic calculated on the basis of which the fuzzy controller will be adapted. During adaptation arbitrary values of the steam temperature downstream the superheater first stage t_{s1st} in the range from 360° to 400°C, the condensate temperature t_c from 50° to 110°C, will be used the outlet steam temperature may be preset in the range from 270° to 400°C.

Arbitrary values of the steam temperature, condensate temperature t_{sl} and desired outlet steam temperature t_{sII} are entered into the model (1), the controller in its turn automatically adjusts to match variations and produces the current valve for the valve opening gap α (table 4).

Table 4. Values calculated by means of the model

t_{s1st}	t_c	t_{sII}	α
363	54	362	18
365	56	345	40
368	58	310	86
372	63	327	68
378	59	315	93
376	82	340	46
374	84	270	128
376	92	334	47

371	83	352	24
373	104	300	78
372	102	325	47
377	98	350	25
374	106	322	51
379	108	306	76
373	87	290	100
382	52	343	68
384	59	376	26
387	63	356	52
383	57	335	76
386	67	313	101
392	78	324	89
389	84	370	25
393	88	352	50
394	92	331	75
397	102	309	100
388	108	359	23
396	104	368	25
387	104	337	51
398	107	328	75
392	98	358	36
392	53	369	49
398	61	394	22
391	72	357	51
393	56	347	75
394	78	317	101
396	82	321	96
398	94	375	25
392	87	351	51
397	98	331	76
391	107	364	22
396	108	366	25
394	102	340	58
392	99	325	76
399	109	368	26

Let's introduce the designation:

$$W^* = W - \Delta W \cdot \mu(C_p) \cdot \mu(C_k), \quad (4)$$

where W^* – new production rules characteristics as derived from table 7;

W – production rules characteristics from table 3;

$\mu(C_p) \cdot \mu(C_k)$ – the input variables degree membership, where

C_p is the membership degree of the steam temperature downstream the superheater first stage t_{slst} , C_k is the membership degree of the condensate temperature, i.e. C_p must be multiplied by C_k ;

ΔW – adaptation step, which is determined by an expert to specify the controller precision.

To create both new production rules and their characteristics, by selecting in turn the lines from the table 4, it is necessary to determine among which terms lie the steam temperature downstream the superheater first stage t_{slst} and the condensate temperature t_c . Further, from the previously obtained rules (table 3) we find the rules, in which terms of the steam temperature t_{sl} and of the condensate temperature t_c coincide with those from table 4. We change over to the next production rule using the characteristics of the selected rule and using the formula (2), where W is substituted by W^* .

Thus, in obtaining the production rule characteristics, we change over to a next rule. New rules are needed to create the fuzzy controller adaptation.

The process of obtaining a new rule:

1. By selecting the 1st line from table 4, we determine that the steam temperature downstream the superheater first stage t_{slst} lies between the TS1 and TS2 terms (figure 1), the condensate temperature of t_c too lies between the TC1 and TC2 terms (figure 1). Further on we find a rule from table 3, where the temperatures occur lying among the same terms numbers. The first production rule from table 3 is now to be selected.

2. At first, we have to determine a new production rule characteristics $W_1^* = 2.06 - 0.15 \cdot 0.7 \cdot 0.1 = 2.0495$.

3. Using the formula (6) we now may calculate it to change over to a new production rule: If $t_v = T_{v1}$ and $t_c = T_{c1}$ then $\alpha = \alpha_2$ and $\alpha = \alpha_3^{0.099}$. Similarly, the remaining 44 production rules (table 5) are calculated.

Table 5. New production rules and their characteristics W^*

№	W^*	New production rules
1	1,0495	If $t_n = T_{n1}$ and $t_k = T_{k1}$ then $\alpha = \alpha_1$ and $\alpha = \alpha_2^{0.099}$
2	2,5175	If $t_n = T_{n1}$ and $t_k = T_{k2}$ then $\alpha = \alpha_2^{0.035}$ and $\alpha = \alpha_3$
3	3,042	If $t_n = T_{n1}$ and $t_k = T_{k3}$ then $\alpha = \alpha_3$ and $\alpha = \alpha_4^{0.084}$
4	4,992	If $t_n = T_{n2}$ and $t_k = T_{k3}$ then $\alpha = \alpha_4^{0.984}$ and $\alpha = \alpha_5$
5	5,013	If $t_n = T_{n2}$ and $t_k = T_{k3}$ then $\alpha = \alpha_5$ and $\alpha = \alpha_6^{0.026}$
6	5,561	If $t_n = T_{n2}$ and $t_k = T_{k7}$ then $\alpha = \alpha_5^{0.122}$ and $\alpha = \alpha_6$
7	4,488	If $t_n = T_{n2}$ and $t_k = T_{k7}$ then $\alpha = \alpha_4$ and $\alpha = \alpha_5^{0.976}$
8	5,591	If $t_n = T_{n3}$ and $t_k = T_{k9}$ then $\alpha = \alpha_5^{0.182}$ and $\alpha = \alpha_6$
9	4,639	If $t_n = T_{n3}$ and $t_k = T_{k8}$ then $\alpha = \alpha_4^{0.278}$ and $\alpha = \alpha_5$
10	4,6315	If $t_n = T_{n3}$ and $t_k = T_{k8}$ then $\alpha = \alpha_4^{0.263}$ and $\alpha = \alpha_5$
11	5,573	If $t_n = T_{n3}$ and $t_k = T_{k9}$ then $\alpha = \alpha_5^{0.146}$ and $\alpha = \alpha_6$
12	5,627	If $t_n = T_{n3}$ and $t_k = T_{k9}$ then $\alpha = \alpha_5^{0.254}$ and $\alpha = \alpha_6$
13	2,004	If $t_n = T_{n3}$ and $t_k = T_{k5}$ then $\alpha = \alpha_2$ and $\alpha = \alpha_3^{0.008}$
14	1,973	If $t_n = T_{n3}$ and $t_k = T_{k6}$ then $\alpha = \alpha_1^{0.946}$ and $\alpha = \alpha_2$
15	4,627	If $t_n = T_{n3}$ and $t_k = T_{k8}$ then $\alpha = \alpha_4^{0.254}$ and $\alpha = \alpha_5$
16	1,944	If $t_n = T_{n3}$ and $t_k = T_{k5}$ then $\alpha = \alpha_1^{0.888}$ and $\alpha = \alpha_2$
17	1,504	If $t_n = T_{n4}$ and $t_k = T_{k7}$ then $\alpha = \alpha_1^{0.008}$ and $\alpha = \alpha_2$
18	2,045	If $t_n = T_{n3}$ and $t_k = T_{k6}$ then $\alpha = \alpha_2$ and

		$\alpha = \alpha_3^{0,09}$
19	4,636	If $t_n = T_{n3}$ and $t_k = T_{k8}$ then $\alpha = \alpha_4^{0,272}$ and $\alpha = \alpha_5$
20	3,049	If $t_n = T_{n4}$ and $t_k = T_{k8}$ then $\alpha = \alpha_3$ and $\alpha = \alpha_4^{0,098}$
21	1,513	If $t_n = T_{n4}$ and $t_k = T_{k7}$ then $\alpha = \alpha_1^{0,026}$ and $\alpha = \alpha_2$
22	3,004	If $t_n = T_{n4}$ and $t_k = T_{k9}$ then $\alpha = \alpha_3$ and $\alpha = \alpha_4^{0,008}$
23	4,022;	If $t_n = T_{n4}$ and $t_k = T_{k9}$ then $\alpha = \alpha_4^{0,044}$ and $\alpha = \alpha_5$
24	2,074	If $t_n = T_{n5}$ and $t_k = T_{k9}$ then $\alpha = \alpha_2$ and $\alpha = \alpha_3^{0,148}$
25	5,583	If $t_n = T_{n4}$ and $t_k = T_{k11}$ then $\alpha = \alpha_5^{0,166}$ and $\alpha = \alpha_6$
26	5,562	If $t_n = T_{n4}$ and $t_k = T_{k11}$ then $\alpha = \alpha_5^{0,124}$ and $\alpha = \alpha_6$
27	4,022	If $t_n = T_{n5}$ and $t_k = T_{k11}$ then $\alpha = \alpha_4$ and $\alpha = \alpha_5^{0,044}$
28	5,589	If $t_n = T_{n4}$ and $t_k = T_{k11}$ then $\alpha = \alpha_5^{0,178}$ and $\alpha = \alpha_6$
29	4,488	If $t_n = T_{n5}$ and $t_k = T_{k12}$ then $\alpha = \alpha_4$ and $\alpha = \alpha_5^{0,976}$
30	5,062	If $t_n = T_{n4}$ and $t_k = T_{k10}$ then $\alpha = \alpha_5$ and $\alpha = \alpha_6^{0,124}$
31	1,561	If $t_n = T_{n4}$ and $t_k = T_{k7}$ then $\alpha = \alpha_1^{0,122}$ and $\alpha = \alpha_2$
32	1,567	If $t_n = T_{n4}$ and $t_k = T_{k7}$ then $\alpha = \alpha_1^{0,134}$ and $\alpha = \alpha_2$
33	2,977	If $t_n = T_{n4}$ and $t_k = T_{k8}$ then $\alpha = \alpha_2^{0,954}$ and $\alpha = \alpha_3$
34	2,074	If $t_n = T_{n5}$ and $t_k = T_{k9}$ then $\alpha = \alpha_2$ and $\alpha = \alpha_3^{0,148}$
35	3,03	If $t_n = T_{n5}$ and $t_k = T_{k10}$ then $\alpha = \alpha_3$ and $\alpha = \alpha_4^{0,06}$
36	3,061	If $t_n = T_{n4}$ and $t_k = T_{k8}$ then $\alpha = \alpha_3$ and $\alpha = \alpha_4^{0,122}$
37	2,964	If $t_n = T_{n5}$ and $t_k = T_{k10}$ then $\alpha = \alpha_2^{0,928}$ and $\alpha = \alpha_3$
38	2,098	If $t_n = T_{n5}$ and $t_k = T_{k9}$ then $\alpha = \alpha_2$ and $\alpha = \alpha_3^{0,196}$
39	3,977	If $t_n = T_{n5}$ and $t_k = T_{k11}$ then $\alpha = \alpha_3^{0,954}$ and $\alpha = \alpha_4$
40	4,533	If $t_n = T_{n5}$ and $t_k = T_{k12}$ then $\alpha = \alpha_4^{0,066}$ and $\alpha = \alpha_5$
41	4,548	If $t_n = T_{n5}$ and $t_k = T_{k12}$ then $\alpha =$ $\alpha_4^{0,096}$ and $\alpha = \alpha_5$
42	5,506	If $t_n = T_{n5}$ and $t_k = T_{k13}$ then $\alpha =$ $\alpha_5^{0,012}$ and $\alpha = \alpha_6$
43	4,524	If $t_n = T_{n5}$ and $t_k = T_{k12}$ then $\alpha =$ $\alpha_4^{0,048}$ and $\alpha = \alpha_5$
44	4,016	If $t_n = T_{n5}$ and $t_k = T_{k11}$ then $\alpha = \alpha_4$ and $\alpha = \alpha_5^{0,032}$
45	5,548	If $t_n = T_{n5}$ and $t_k = T_{k13}$ then $\alpha = \alpha_5^{0,096}$ and $\alpha = \alpha_6$

The program to control the outlet steam temperature t_{sII} and as well as for the fuzzy controller adaptation is developed in the software Unity Pro XL using ST language based on PLC Modicon M340. Thus, this paper deals with the fuzzy controller synthesis designed to control the outlet steam temperature t_{sII} in the boiler unit. Besides the fuzzy controller adaptation has been effected based on the method of selection of that consequent in production rules with a double consequent, which is optimal for the current values of the boiler unit parameters.

3. Conclusion

Algorithm of adaptation fuzzy controller is developed, which includes the following steps:

1. The definition of a conceptual model of a fuzzy controller.
2. Development of a model that characterizes the flow of the process.
3. Synthesis of fuzzy controller. Synthesis includes: collection of expert information, construction of the table of reference points, definition of linguistic variables, construction of the table of desirable values, synthesis of the received production rules.
4. The fuzzy controller adaptation: obtaining characteristics obtained at the stage of synthesis of production rules, the use of a model characterizing the flow of the process to obtain new characteristics and new production rules.
5. Develop program code that implements the adaptation of a fuzzy controller.

References

1. Synthesis methods for fuzzy controller with double base of rules Soloviyov K. A., Muravyova E. A. Proceedings of the 4th International conference: Information technologies for intelligent decision-making support (ITIDS'2016). 2016.P. 200-203.
2. Fuzzy controller adaptation Muravyova E.A., Sharipov M.I., Radakina D.S. В сборнике: IOP Conference Series: Earth and Environmental Science Сер. "Innovations and Prospects of Development of Mining Machinery and Electrical Engineering - Mechanical Engineering" 2017. С. 082033.
3. Muravyova E. Intellectual control of oil and gas transportation system by multidimensional fuzzy controllers with precise terms / Muravyova E., Sagdatullin A., Emekeev A. // Applied Mechanics and Materials. 2015. Т. 756. С. 633.
4. Muravyova E., Sagdatullin A., Sharipov M. Modelling of Fuzzy Control Modes for the Automated Pumping Station of the Oil and Gas Transportation System // IOP Conference Series: Materials Science and Engineering, 2016. – 132 (1), 012028.
5. Muravyova E.A. Fuzzification concept using the any-time algorithm on the basis of precise term sets // Industrial Engineering, Applications and Manufacturing (ICIEAM) : International Conference on Industrial Engineering, Applications and Manufacturing, May 2017 / IEEE. - СПб., 2017.
6. Muravyova E.A. Two fuzzy controller synthesis methods with the double base of rules: Reference points and training using // Industrial Engineering, Applications and Manufacturing (ICIEAM) : International Conference on Industrial Engineering, Applications and Manufacturing, 16-19 May 2017 / IEEE. - СПб., 2017.

STREAM HANDLING LARGE VOLUME DOCUMENTS IN SITUATIONALLY-ORIENTED DATABASES

V. V. Mironov, A. S. Gusarenko, N. I. Yusupova

Department of Computer Science and Robotics, Ufa State Aviation Technical University, Ufa, Russia

e-mail: mironov@ugatu.su, gusarenko@ugatu.su, yussupova@ugatu.ac.ru

Abstract: *The article discusses the streaming data processing in large quantities for the situation-oriented databases (SODB). Traditionally SODB provides cached processing of heterogeneous data. This article considered the work with large files that do not fit entirely in memory - containing a number of similar fragments, which can be processed portions. Portions of data are extracted from the input stream, processed in the buffer, and then sent to the output stream. Three variants of implementing this scheme are considered in the framework of the hierarchical situational model of the HSM. The tools offered at the conceptual level are illustrated by the example of processing XML data using PHP software tools, such as XMLReader, XMLWriter, DOM.*

1. Introduction

At the present stage of the information systems development accessible through the network, large data sets accumulated in the course of their life cycle are used in their architecture. As part of such large-scale information systems, databases are used as information support that help to cope with the tasks of processing, storing data and supporting automated functions [1, 2]. Recently, two areas of DBMS development have been identified, the first SQL database and NoSQL, which recently received well-deserved attention by using not only SQL-tools, but also their own data structures in the form of documents suitable for processing and storage, as well as language tools that do not use SQL. Situationally-oriented databases (SODB) is the integrator of diverse (heterogeneous) data. As part SODB formalized dynamic model HSM [3] and programmed functionality to handle the current application state, based on SODB opportunities. In model states, data from relational database tables, such as MySQL [4, 5], are processed to which they are bound. At the initial stage, these were cached processing technologies [6], the technology was based on the DOM (Document Object Model), used in a variety of modern applications that target web applications and other desktop applications. DOM technology used to create a dynamic model states SODB for processing linked to the states of the model data. As soon as the application changed the current state to the state with which the data was associated, dynamic DOM objects were automatically created to process them. With this approach, the application for executing the natural operations of the DBMS for data processing was occupied by the operative memory. In order to avoid the overhead associated with memory utilization, another type of data processing objects based on XMLReader technology was proposed and developed, which was useful for streaming data processing and developing algorithms [6] for processing data in DBMS. Over time, the number of sources for SODB grew into new types of objects [7] have been proposed as a response to, for example, JSON to meet the needs in data processing. The introduction of new types of objects has been further progress in the direction of coverage processing capabilities of heterogeneous data sources that have been considered in the work on SODB [8]. In the next stage of development SODB, when there external sources of heterogeneous data started to develop a web presence in the form of independent SODB services were built into the dynamic model and the remote execution of natural reading query tool for processing data from remote RESTful-service [9, 10]. As the number of sources and the number of data processed increased, it became necessary to organize VDA (Virtual Data Array) [10] structured [11] and invariantly mapped to heterogeneous data stores [12]. In the present conditions, processing in virtual data arrays is required using DPO (Data Processing Objects), such as DOM and Smarty objects [13-15]. The volume of data grows over time, so there are sources [16-19] big data.

2. Stream Handling Large Volume Documents

Custom Data Sources. Basically, this is data received from users who trust and store their data in web applications. Sources of such data can be presented in the form of databases, data stores, individual files and documents in common formats, as well as archives. Such data is also generated by web-based systems for automatic backup, thus creating a dump of information already available in databases, as well as metadata for them.

Web service data. There is another common way of working with sources, that is, not the Web system's own data, but a request for external data distributed on the terms of remote services, for example, by subscription, since Scopus (the international database of metadata and full texts of scientific journals) or Wikipedia, which has its own backup storage - database dumps, it includes its own generated data from the contents of articles and the table of contents and other semantic parts of the encyclopedia. Dumps occupy a significant size for each version, stored in the form of archives, in which are placed article files, with a branched structure, it can be either HTML files or metadata in the form of XML. Such data occupy a large volume even in archives. Wikipedia article files stored together in the archive take up a large amount of data, while the metadata itself, which is a single XML file, occupies a comparable amount of data. Such a dump file can not be normally read in a browser or text editor to view large files, because there is a shortage of memory, there are difficulties in navigating the file.

The task of obtaining and binding data from own sources and remote services. The task of obtaining such data arises if you need to process the associated information of your own source with the data of the remote service. An example can be the user's activity or his work in another service, for example, writing a Wikipedia article or publishing a scientific article in a publication indexed by Scopus. The natural needs to link such data are met by their receipt and processing. When information is received, there are difficulties associated with the large amount of data transmitted over the network, while the download time consumes them.

Processing of received data. When downloading and local data unpacking, the extraction and processing time is doubled, and a significant portion of the external storage will be occupied. To circumvent this kind of difficulty in local processing, so-called wrappers (wrappers) are implemented for streaming reading of deleted data. Thus, data files are not downloaded entirely, but instead are opened on a remote service and read by the program. If the data is not large, it can be read and processed in memory (cached processing) entirely [8], but if they occupy a large part of the memory, the stream processing is used [8]. Stream processing helps to save memory and receive results portion by piece in the right amount.

The problem of obtaining and processing data from a service SODB. The task of processing large data sets when there is a file with a large number of monotonous objects in the tree view of XML, etc. is solved and implemented in modern information systems tools. In most systems, XMLReader's stream processing technology is implemented and used to process large data sets. At

the moment, there is no SODB advanced tools can make efficient use of memory to handle the same type of objects contained in the XML-files. The problem of obtaining and processing large amounts of data from the services received for SODB requires decisions by equipping a dynamic model of the specialized stream processing means based on the existing technology of streaming read and write XMLReader and XMLWriter, as well as the possibility of creating a DPO-objects based on the DOM for the intermediate data conversion.

Streams and stream processing. Currently most systems implement programming tools for streaming. The flow regulates the specifics of the data that will be transferred to the application for processing, as well as in the reverse direction with respect to the data that will be output after processing. These streams are oriented to files transmitted over the network, data sent in a compressed form, such as archives. In this case, the flow is meant to be remote resources or their groups, in which the data is stored in a sequential manner, such as an XML file containing sequentially described monotonous objects. Data organized sequentially requires sequential reading by specialized means, for this there are particular cases of implementing such tools XMLReader/XMLWriter, oriented to work with XML. In those conditions, when there are non-standard protocols and data types, it is required to implement additional program code, called a wrapper in the PHP programming system. There are standard wrappers that have their own context, where it is determined which parameters are required to handle a particular known data type. If there are not enough wrappers for non-standard protocols and data types, the programming system is given the opportunity to implement their own wrappers in the form of a separate script, and there are no restrictions on the number of such wrappers. In this paper, we use standard thread wrappers such as `php://output` that is responsible for the standard output of the PHP system.

3. HSM-stream processing model

Let's consider in general form the idea of the HSM-model for stream processing of documents in the DBMS. The idea is that virtual documents are defined in the model, which, on the one hand, are mapped to the data being processed, and on the other hand, they are associated with the data flow handler. In Fig. 1 summarizes three variants of the organization of the HSM model for streaming data processing: with explicit specification of the thread processing cycle (Fig. 1, a); with the organization of a cycle of processing by means of a loop transition (Fig. 1, b); with implicit specification of the processing cycle (Fig. 1, c). In the `sta:Doc-Stream-`

Processing state, two virtual documents (VD) are defined: `doc:Input` and `doc:Output` are the source and destination of the data, respectively. In all three versions of the declaration of virtual documents are the same, the differences are manifested in the organization of the processing of their data. VDs are defined using `doc`-elements the same as in the case of cached processing. The attributes of these elements (for simplicity, not shown in Figure 1) define the physical data store to which the virtual document is displayed. The difference is in setting the VD processing method using the method attribute, which indicates the thread's handler used. The `action` attribute opens a stream with certain parameters specifying those portions that will be read from the data source or sent to the receiver. When processing is complete, open threads are closed using the same attribute. The VD processing is performed in the `sub:Proc` submodel, which includes two states: `sta:DocProc` – actual document processing; and `sta:ErrProc` – error handling.

Variant a - with an explicit cycle of stream processing. In this variant, stream processing is organized using a `for`-element, intended for cyclic interpretation of the embedded model fragment (`for:EachPiece`, see Fig. 1, a). The `scan` attribute here points to VD `doc:Input`, so the `for` element sets the data to be cyclically read from this VD and loads them by default into the buffer of the same name as a `dpo:EachPiece` data object. This object provides a transition to the `sta:ErrProc` state in case of an error (`onLoadErr` attribute) and saving the content in the `doc:Output` document if the `saveDoc` attribute is successfully completed. The actual processing of the contents of the buffer is specified by the nested elements of the `dpo`-element (in Figure 1 are not shown).

Variant b - with a loop transition to organize a cycle of stream processing. In this case, the `dpo`-element referencing the VD `doc:Input` (the `srcDoc` attribute) specifies the loading of the next portion of data for processing into the buffer. The processing cycle is provided by the jump element `jmp:DocProc`, which after the buffer processing executes a second transition to the `sta:DocProc` state until the end of the file of the readable document (the `when` attribute) is reached.

Variant c - with an implicit definition of the thread processing cycle. In this case, the model fragment related to the data processing looks exactly the same way as for the cached document processing. This implies an implicit loop, such as in case of variant b, but without a loopback. Thus, here during the interpretation of the `dpo`-element, there is a call to the input stream handler, loading into the buffer portions of data, processing them and repeating this

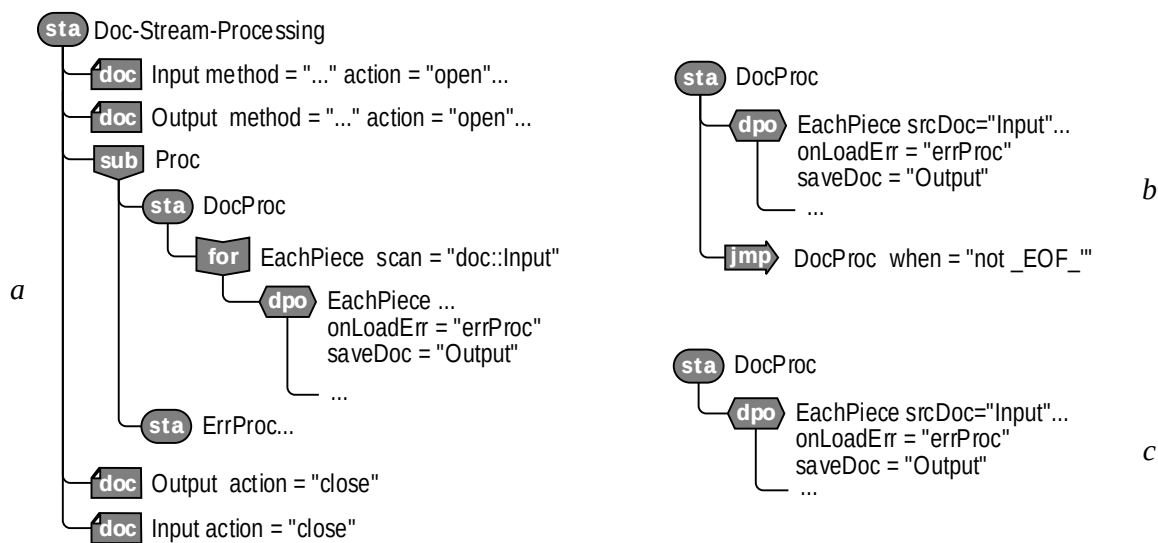


Fig. 1. General view of the HSM-model of streaming document processing:
a - with an explicit cycle; b - with a loop; c - with an implicit cycle

procedure if the end of the file is not reached.

Comparison of variants. It is currently difficult to say which option is preferable. In variant *a*, the input flow handler used (the request to read the next portion of data) is executed during the interpretation of the *for*-element, in variants *b* and *c* during the interpretation of the *dpo*-element. Variant *c* expands the general concept of interpreting the *dpo*-element, introducing the notion of cyclicity. In this case, the cyclicity of the flow processing is not reflected explicitly in the model. However, in this case, the principle of DPO processing invariance is respected in relation to the definition of VD [5]: the part of the model that specifies the processing of the document remains unchanged when the mapping is changed to physical storage. For example, the processing model does not change externally if, when defining VD, a cached data handler is specified instead of streaming.

4. Example: streaming XML processing

XML-data are "native" to SODB - they used them initially. Let's illustrate the stream processing on this data format. In Fig. Figure 2 shows an example of an HSM model for streaming a large XML document. The XML document used is a file from the archive of the Russian-language Wikipedia. The file size exceeds 100 GB, its cached processing on a normal personal computer is impossible. The structure of the document is shown in Fig. 2, *a*. The root element contains the child *siteinfo* (the content is not shown) and a number of child *pages* that reflect information about changes made to Wikipedia articles. Thus, this document refers to a fairly common form of large files containing a large number of small similar fragments. This circumstance makes it possible to effectively organize the stream processing. Let it be required to select from the document those page elements that belong, say, to the author with identifier 121755, and not all information, but only some, in another format is needed. Thus, the processing of the document must include filtering the pages by content, as well as transforming the content. In Fig. 2, *b* shows the corresponding HSM model.

In the *sta:XML-Stream-Processing* state, two virtual

multidocuments are specified: *doc:Input*, *ent:RuWiki*, mapped to the XML file of the document being processed, and *doc:Output*, *ent:Tmp*, mapped to the resulting output XML file. With both VD the flow handlers are connected: with the first - XML-Reader - the streaming reader, and with the second - XML-Writer - respectively, the means of streaming XML documents. In the *sub:Proc* sub-model, the virtual documents are streamed according to the scheme of variant *a* (see Figure 1, *a*). In the *sta:DocProc* state, the input and output documents are opened, while reading portion of the XML page element, and for the *Output.Tmp* output, the root XML element name. The element *for:Page* cyclically reads the XML elements of the page from the input document and places them by default in the DOM-object *dom:Page* of the same name for processing the *Input.RuWiki* input document specifies the *saveDoc* attribute of the *dom*-element prescribes to save the result in the output document (to send to the output stream) at the end of the processing of the data portion. The actual processing of the data portion is specified by the *jmp:_CONT_* and *rcv:Transform* elements. Using *jmp:_CONT_*, the contents of the page being processed are checked, namely the value of the author's identifier (contributor) using the XPath expression in the when attribute.

XML-data are "native" to SODB - they used them initially. Let's illustrate the stream processing on this data format. In Fig. Figure 1 shows an example of an HSM model for streaming a large XML document. The XML document used is a file from the archive of the Russian-language Wikipedia. The file size exceeds 100 GB, its cached processing on a normal personal computer is impossible. The structure of the document is shown in Fig. 2, *a*. The root element contains the child *siteinfo* (the content is not shown) and a number of child *pages* that reflect information about changes made to Wikipedia articles. Thus, this document refers to a fairly common form of large files containing a large number of small similar fragments. This circumstance makes it possible to effectively organize the stream processing. Let it be required to select from the document those page elements that belong, say, to the author with identifier 121755, and not all information, but only

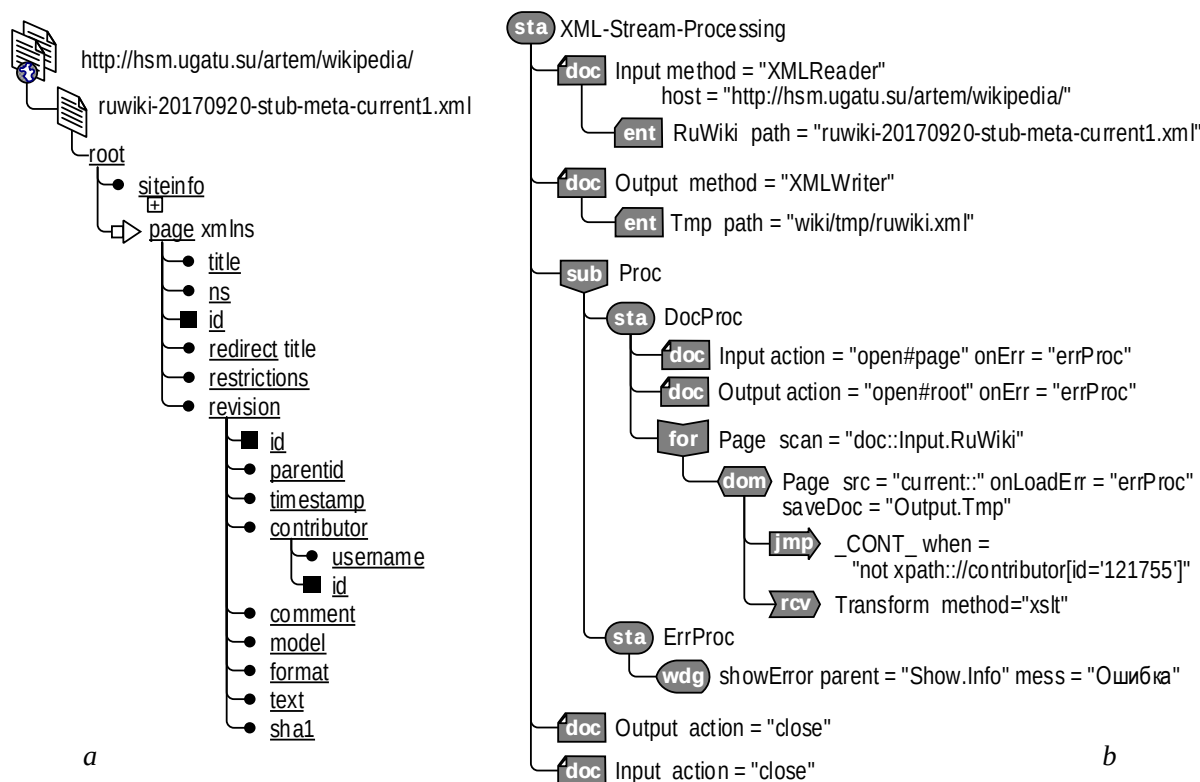


Fig. 2. An example of an HSM model for streaming XML data:
a - the structure of the XML document; *b* - stream processing model

some, in another format is needed. Thus, the processing of the document must include filtering the pages by content, as well as transforming the content. In Fig. 2, *b* shows the corresponding HSM model. In the `sta:XML-Stream-Processing` state, two virtual multidocuments are specified: `doc:Input, ent:RuWiki`, mapped to the XML file of the document being processed, and `doc:Output, ent:Tmp`, mapped to the resulting output XML file. With both VD the flow handlers are connected: with the first - XML-Reader - the streaming reader, and with the second - XML-Writer - respectively, the means of streaming XML documents. In the `sub:Proc` sub-model, the virtual documents are streamed according to the scheme of variant a (see Figure 1, a). In the `sta:DocProc` state, the input and output documents are opened, while reading portion of the XML page element, and for the `Output.Tmp` output, the `root` XML element name. The element `for:Page` cyclically reads the XML elements of the page from the input document and places them by default in the DOM-object `dom:Page` of the same name for processing the `Input.RuWiki` input document specifies the `saveDoc` attribute of the `dom`-element prescribes to save the result in the output document (to send to the output stream) at the end of the processing of the data portion. The actual processing of the data portion is specified by the `jmp:_CONT_` and `rcv:Transform` elements. Using `jmp:_CONT_`, the contents of the page being processed are checked, namely the value of the author's identifier (contributor) using the XPath expression in the `when` attribute. XML-data are "native" to SODB - they used them initially. Let's illustrate the stream processing on this data format. In Fig. Figure 1 shows an example of an HSM model for streaming a large XML document. The XML document used is a file from the archive of the Russian-language Wikipedia. The file size exceeds 100 GB, its cached processing on a normal personal computer is impossible. The structure of the document is shown in Fig. 2, *a*. The `root` element contains the child `siteinfo` (the content is not shown) and a number of child `pages` that reflect information about changes made to Wikipedia articles. Thus, this document refers to a fairly common form of large files containing a large number of small similar fragments. This circumstance makes it possible to effectively organize the stream processing. Let it be required to select from the document those page elements that belong, say, to the author with identifier 121755, and not all information, but only some, in another format is needed. Thus, the processing of the document must include filtering the pages by content, as well as transforming the content. In Fig. 2, *b* shows the corresponding HSM model. In the `sta:XML-Stream-Processing` state, two virtual multidocuments are specified: `doc:Input, ent:RuWiki`, mapped to the XML file of the document being processed, and `doc:Output, ent:Tmp`, mapped to the resulting output XML file. With both VD the flow handlers are connected: with the first - XML-Reader - the streaming reader, and with the second - XML-Writer - respectively, the means of streaming XML documents. In the `sub:Proc` sub-model, the virtual documents are streamed according to the scheme of variant a (see Figure 1, a). In the `sta:DocProc` state, the input and output documents are opened, while reading portion of the XML page element, and for the `Output.Tmp` output, the `root` XML element name. The element `for:Page` cyclically reads the XML elements of the page from the input document and places them by default in the DOM-object `dom:Page` of the same name for processing the `Input.RuWiki` input document specifies the `saveDoc` attribute of the `dom`-element prescribes to save the result in the output document (to send to the output stream) at the end of the processing of the data portion. The actual processing of the data portion is specified by the `jmp:_CONT_` and `rcv:Transform` elements. Using `jmp:_CONT_`, the contents of the page being processed are checked, namely the value of the author's identifier (contributor) using the XPath expression in the `when` attribute. If the author's identifier does not match the specified value, the processing is interrupted, thereby this `page` element is ignored in the output

stream. Using the `rcv:Transform` receiver, XSL transforms the contents of the DOM object before sending it to the output stream. XSLT is a powerful tool for converting XML data based on style sheets (Stylesheets), both in XML and in other formats. The conversion method is specified by the `method` attribute, by default the style sheet, which is the same as the element, is used, i.e., `Transform.xsl`. The details of the transformation are omitted here for brevity. When errors occur, the state goes to `sta:ErrProc`, where the corresponding message is generated to the user.

5. Conclusion

In the article the questions of the task of streaming data processing in DBMS on the basis of declarative HSM-models of integration of heterogeneous data sources are considered. For data processing in the HSM, it is provided, firstly, to set VD - virtual documents displayed on the physical storage, and secondly, to set DPO - data processing objects, into which data from VD is loaded. The previously provided HSM capabilities are focused on cached processing, in which VD is completely loaded into the main memory. This imposes restrictions on the amount of processed documents. For processing large data, computing platforms provide streaming processing capabilities, in which data is downloaded in batches from memory to the memory from the source, processed there, and uploaded to the data receiver. The task is to develop declarative means of specifying stream processing within the framework of the HSM. The introduction of stream processing required additional provision in the HSM:

1) The ability to specify a particular streaming handler when specifying VD, and specifying the portions of the data extracted from the physical store or placed in the peg. For this, the `method` attribute is provided in the virtual document definition;

2) Possibility of setting cyclic processing of data portions VD. For this, three organization variants are proposed: with an explicit cycle, which is provided by the `for`-element; with the cycle provided by the `jmp`-element; with an implicit cycle provided by a modified `dpo`-element.

The possibilities of specifying stream processing are demonstrated in the example of processing a large XML document. Portions of the document are read by the XMLReader tool, loaded into the DOM-object and processed in it, and then output to the output stream by the XMLWriter tool. Thus, the general concept of VD / DPO, used in DBMS for cached processing, with minor modifications is also applicable for the stream processing of large documents.

Acknowledgments

This work is supported by RFBR grant № 16-07-00239.

References

1. Pokorný J., "NoSQL databases: a step to database scalability in web environment", *International Journal of Web Information Systems*, 2013; 9(1):69–82.
2. Mironov V. V., Gusarenko A. S., Dimetiev R. R., Sarvarov M. R. "The Personalized Documents Generating Using DOM-objects in Situation-Oriented Databases", *Vestnik UGATU*. 2014; 65:191–197. (In Russian).
3. Kosar T., Bohra S., Memik M. Domain-Specific Languages: A Systematic Mapping Study // *Information and Software Technology*. 2016; 71:77–91.
4. Gusarenko A. S. "Improvement of Situation-Oriented Database Model for Interaction with MySQL", *Izvestiya vysshikh uchebnykh zavedeniy. Priborostroenie*. [Journal of Instrument Engineering] 2016; 59:355–363. (In Russian).
5. Mironov V. V., Gusarenko A. S., Yusupova N. I., "Displaying virtual the XML-documents to MySQL tables in situation-oriented databases, "distributed approach"", *Informacionnye*

- tehnologii i vychislitel'nye sistemy* [Information technology and computer systems], 2017; 1:77–89. (In Russian).
6. Mironov V. V., Gusarenko A. S., Yusupova N. I. "Situation-oriented databases: current state and prospects for research", *Vestnik UGATU*. 2015; 68:188–199. (In Russian).
7. Gusarenko A. S., Mironov V. V. "Smarty-objects: Use Case of Heterogeneous Sources in Situationally-Oriented Databases", *Vestnik UGATU*. 2014; 64:242–252. (In Russian).
8. Gusarenko A. S., Mironov V. V. "Heterogeneous Document Sources in Situationally-Oriented Databases", *Vestnik UGATU*. 2015; 19:124–131. (In Russian).
9. Mironov V. V., Gusarenko A. S. "Using of RESTful-Services in Situationally-Oriented Databases", *Vestnik UGATU*. 2015; 67:232–239. (In Russian).
10. Mironov V. V., Gusarenko A. S., Yusupova N. I. "Situation-Oriented Database: Integration of XML Data and Relational Environment" // *Sistemy upravleniya i informacionnye tehnologii* [Automation and Remote Control]. 2016; 65:48–56. (In Russian).
11. Mironov V. V., Gusarenko A. S., Yusupova N. I. "Structuring Virtual Multi-Documents in Situationally-Oriented Databases by Means of Entry-Elements", *Trudy SPIIRAN* [SPIIRAS Proceedings]. 2017; 53:225–243. (In Russian).
12. Mironov V. V., Gusarenko A. S., Yusupova N. I., "The Invariance of The Virtual Data in The Situationally Oriented Database When Displayed on Heterogeneous Data Storages", *Vestnik komp'yuternykh i informatsionnykh tekhnologii* [Herald of Computer and Information Technologies], 2017; 151:29–36. (In Russian).
13. Osvaldo S. S. Jr. etc. "Developing software systems to Big Data platform based on MapReduce model: An approach based on Model Driven Engineering", *Information and Software Technology*. 2017; 92:30–48.
14. Cobo M. J., López-Herrera A.G., Herrera-Viedma E. "A relational database model for science mapping analysis", *Acta Polytechnica Hungarica*. 2015; 6:43–62.
15. Arevalo C. etc. "A metamodel to integrate business processes time perspective in BPMN 2.0", *Information and Software Technology*. 2016; 77:17–33.
16. Amanatidis T., Chatzigeorgiou A. "Studying the evolution of PHP web applications", *Information and Software Technology*. 2016; 72:48–67.
17. Agh H., Ramsin R. "A pattern-based model-driven approach for situational method engineering", *Information and Software Technology*. 2016; 78:95–120.
18. Mironov V. V., Gusarenko A. S., Yusupova N. I. Situation-oriented databases: document management on the base of embedded dynamic model / CEUR Workshop Proceedings (CEUR-WS.org): Selected Papers of the XI International Scientific-Practical Conference Modern Information Technologies and IT-Education (SITITO 2016), Moscow, Russia, November 25-26, 2016. P. 238–247.
19. Djukic V., Lukovic I., Popovic A., Ivancevic V., "Model Execution: An Approach based on extending Domain-Specific Modeling with Action Reports", *Computer Science and Information Systems*, 2013; 4:1585–1620.

RECOGNITION OF TEXT INFORMATION IN THE BRONCHOPULMONARY DISEASES DIAGNOSIS SYSTEM

G.R. Shakhmametova¹, A.A. Evgrafov¹, R.Kh. Zulkarneev²

Computer Science and Robotics Department, Ufa State Aviation Technical University Ufa, Russia¹

Faculty of General Medicine, Bashkir State Medical University Ufa, Russia²

e-mail: shakhgouzel@mail.ru, evgrafov.alexander92@yandex.ru, zrستم@ufanet.ru

Abstract: In article the problem of recognition of text information is considered. The obtained information is processed in the form of the scanned pages with a large number of medical terms for the purpose of further processing in the system of diagnosis of bronchopulmonary diseases.

1. Introduction

Active penetration of information technologies into medical branch stimulated creation of program complexes which main goal is essential improvement of quality of rendering medical services due to introduction of modern technologies in process of interaction of medical personnel with patients of clinics. One of such developments – the system of diagnosis of bronchopulmonary diseases. This system assumes existence of the following functions: the analysis of data of the patient in the form of the block of text information, statement of the estimated diagnosis on the basis of the carried-out analysis and the revealed regularities, formation and issue of medical recommendations for increase in efficiency of treatment of the patient [1].

2. Problem definition

One of functional blocks of the developed system is the block of recognition of text information. Existence of this block is caused by specifics of collecting and storage of information on the patients asking for the help in medical institutions. The individual data of patients including both the general information, and the data obtained during delivery of health care (primary survey, holding necessary procedures and their results), are often stored in a type of the unstructured electronic, or printing text and also occupy essential volumes. At the same time it is necessary to understand that qualitative and timely analysis of this sort of information is capable to lead to essential increase in efficiency of diagnosing and treatment of patients. A prime task is transformation of the saved-up information on patients to a format of electronic text documents that will allow to start the further analysis.

3. State of Art

Process of recognition of the text represents a special case of a problem of recognition of images. Recognition of images - the scientific direction connected with development of the principles and creation of the systems intended for definition of belonging of a concrete object to one of in advance allocated classes of objects [2]. The possibility of recognition relies on similarity of the same objects. In spite of the fact that all objects and situations are unique in strict sense, between some of them it is always possible to find similarities on this or that sign. From there is a concept of classification — splitting all set of objects into not crossed subsets - classes which elements have some similar properties distinguishing them from elements of other classes [3].

There are several main approaches described in references to process of recognition of images:

- Heuristic (approach experience and an intuition of the person is the cornerstone, means transfer of members of a class and the principle of community of properties);
- Mathematical (approach rules of classification which are formulated and removed within a certain mathematical formalism by means of the principles of community of properties and a clustering are the cornerstone);

- Linguistic (The image can be described by means of hierarchical structure of subimages, similar syntactic structure of language) [4].

In relation to a solvable task preference is given to mathematical methods as they are the most universal and effective.

4. Proposed solution

During the work with text information the problem of optical recognition of symbols which consists in mechanical or wire transfer of images of the hand-written, typewritten or printing text in the text data which are used for representation of symbols in the computer is solved (for example, in a text editor). Recognition is widely applied to transformation of books and documents to an electron look, to automation of accounting systems in business or to the publication of the text on the web page. Optical recognition of symbols allows to edit the text, to carry out search of words or phrases, to store it in more compact form, to show or unpack material, without losing quality, to analyze information and also to apply to the text wire transfer, formatting or transformation to the speech [5].

In the developed system of diagnostics of bronchopulmonary diseases as entrance data for recognition of the text use of the files of the JPEG format containing the scanned documents of the A4 format with existence of the printing text is supposed. These documents contain necessary information on the patient, symptoms of an estimated disease and also results of different types of the carried-out analyses.

During primary analysis of the documents which are subject to processing the following aspects complicating recognition of text information have been marked out:

- symbols of Cyrillics can settle down, both in lower, and in top registers;
- tracing of symbols can significantly differ, owing to use at the press of various fonts or their sizes;
- as a result of scanning documents with a non-uniform brightness of symbols and also distortions of the text and perspective distortions can be received;
- the recognized text may contain various schedules, charts, tables, designations using Latin symbols, and the additional signs coded, for example, in Unicode.

The above-stated aspects considerably complicate process of recognition of text information and require special attention at the solution of tasks of this sort.

We will consider in more detail process of recognition of texts, having divided him into separate blocks from which it is accepted to distinguish the following:

- segmentation block (localizations and allocations) text elements;
- block of preprocessing of the image;

- block of allocation of signs;
- block of recognition of symbols;
- block of post-processing of results of recognition [6].

It is easy to notice that process of recognition of symbols is only one of components of the described task. First of all the initial image is exposed to the analysis which is designed to localize all significant fragments containing text information with the highest probability. At the same time it should be noted that the block realizing preprocessing of images can be carried out not only after the segmentation block, but, including, to it and, in some cases, together with a problem of localization of symbols.

Process of segmentation of elements of the text comes down to splitting the image of the document for certain areas, at the same time there is an allocation of structural text units, such as lines, words and symbols. Allocation of fragments of high levels (lines and words) can be carried out on the basis of the analysis of intervals between dark areas [7]. At first there is a division of the text into lines by means of the procedure of search between them white strips. Further the approximate value of average distance between symbols is set that allows to allocate separate letters, and distances, big in comparison with averages, will confirm the gaps concluded between words. At the same time it is supposed that offers of the text are located horizontally and have no crossings with each other.

There are several widespread methods of segmentation:

1. Dummy algorithm - the dummy segmentation algorithm takes the whole page as one segment. The purpose of this algorithm is to see how well we can perform without doing anything. Then the performance of other algorithms can be seen as gains over that achieved by the dummy algorithm;
2. The X-Y cut segmentation algorithm [8], also referred to as recursive X-Y cuts (RXYC) algorithm, is a tree-based top-down algorithm. The root of the tree represents the entire document page. All the leaf nodes together represent the final segmentation. The RXYC algorithm recursively splits the document into two or more smaller rectangular blocks which represent the nodes of the tree. At each step of the recursion, the horizontal and vertical projection profiles of each node are computed.
3. Run-length smearing algorithm RLSA algorithm (RLSA) [9] works on binary images where white pixels are represented by 0's and black pixels by 1's.
4. The whitespace analysis algorithm described by Baird [10] analyzes the structure of the white background in document images. The first step is to find a set of maximal white rectangles (called covers) whose union completely covers the background. Breuel's algorithm for finding the maximal empty whitespace is used in our implementation for this step.
5. Constrained text-line detection - the layout analysis approach by Breuel [11] finds text-lines as a two step process: Find tall whitespace rectangles and evaluate them as candidates for gutters, column separators, etc. The whitespace rectangles representing the columns are used as obstacles in a robust least square, globally optimal text-line detection algorithm [12]. Then, the bounding box of all the characters making the text-line is computed.
6. The algorithm Docstrum – is carried out a clustering of symbols, distances between them are calculated, at the same time three minimum distances on increase are allocated: 1 – distance between letters in a word, 2 – interlower case distance, 3 – a space;

7. The Voronoi-diagram based segmentation algorithm by Kise et al. [13] is also a bottom-up algorithm [14].

Finally the image is divided into certain sectors (squares) in which some symbol which is subject to further recognition with guarantee contains.

Further there is a problem of preprocessing of the image which consists in carrying out operations of smoothing over him, filtrations and normalization. Smoothing is meant as process of restoration of the symbols containing the gaps or gaps received during scanning. Also the return process – reduction of the lines significantly differing from others on thickness that allows to simplify further recognition of a symbol is carried out. Methods of binary filtration which give the chance to get rid of roughnesses of borders of a symbol are applied. Normalization of the image, in turn, is intended for elimination of distortions of separate lines and symbols and also reduction of symbols to the uniform height and width after their processing.

The following two operations realized by blocks of allocation of signs and recognitions of symbols have various purpose, but at the same time are closely connected among themselves as directly depends on quality of the marked-out signs whether the system will be able to distinguish this or that symbol. At this stage allocate four main techniques of recognition:

1. The search method – consists in comparing of a defined image with numerous number of the reference samples which are contained in certainly made database;

The deep analysis of characteristics of an image – a method is the cornerstone determination of different geometrical characteristics of characters, their key features;

Use of artificial neural networks - this method requires essential quantity of examples of the task of recognition when training and also special structure of the neural network considering specifics of an objective. It is necessary to mark that use of artificial neural networks is the most perspective and effective method;

The expert method – a method is based on the continuous training of expert system in use [15].

As a rule, above-mentioned techniques are applied not only separately, but also often integrate in a complex of the programs providing effective implementation of an objective on recognition of texts.

The unit of post-processing of results which purpose is improvement of quality of character recognition at the expense of assessment of the context information finishes recognition process. Use of methods of post-processing also allows to find the mistakes made in the text and to correct them. As a rule, two basic approaches are applied to the decision of tasks of this kind: use of dictionaries and sets of binary matrixes. The entity of the first approach consists search of the recognized words in certainly created dictionary. In the absence of coincidence to the dictionary the conclusion that the word contains an error is drawn, and the word is subject to changeover by the most similar to dictionary. The second approach is based on the N-grams representing the character strings consisting of 2-3 letters of a separate word. This method allows to obtain additional information on structure of the checked word, passing the analysis of a context and calculating the reliability level characterizing probability of absence of errors in the checked word. It is necessary to mark that data application of methods requires essential computing resources.

5. The choice of the used way of recognition of the text

The above-mentioned units executing text recognition process can be realized by different modules, programs and components. We will consider some of them.

Among software products today the leader in the field of recognition of the Cyrillic text it is possible to call the development of the Russian company "ABBYY" named "FineReader". The program allows to transfer images of documents (photos, results of scanning, the PDF files) to the electronic edited formats, such as Microsoft Word, Microsoft Excel, Microsoft Powerpoint, Rich Text Format, HTML, PDF/A, PDF, CSV and text files [16].

Also there is a set of other different software products in the field of text recognition: Office Lens (recognizes pictures of the camera and saves them in the format docx, pptx, PDF) [17], Adobe Scan (saves files only in the PDF format) [18], Free OCR to Word (Recognizes: JPG, TIF, BMP, GIF, PNG, EMF, WMF, JPE, ICO, JFIF, PCX, PSD, PCD, TGA and other formats, saves files in the DOC, DOCX, TXT formats) [19], Online OCR (recognizes: JPG, BMP, TIFF, GIF, PDF, saves: DOCX, XLSX, TXT) [20], Microsoft OneNote (recognizes: JPG, BMP, converts into files of OneNote) [21].

Unfortunately, use of above-mentioned products in the developed system is impossible because of absence of a possibility of their adaptation and implementation in system. So, the most perspective decision is use as means of recognition of text information the special modules adapted to different programming languages, in particular to Python.

The Python programming language has been chosen for development as the basic for a number of reasons:

- has the created modules for work in the field of recognition and information processing;
- has rather high speed of performance;
- it is optimized and is widely applied in the solution of problems of the required range.

As the module for recognition of text information use of the pytesseract module is offered. The kernel of the Tesseract program was developed in the Bristol laboratory Hewlett Packard in 1985 — 1994. In 1996 the considerable changes were carried out and the port for Windows is prepared [22]. The main advantage of this software package is the possibility of his adaptation under recognition of the Russian-language text.

Nevertheless, use of this program will also demand additional adaptation in view of specifics of the applied area. It is necessary to provide additional training of a product by means of medical dictionaries for increase in overall performance.

6. Conclusions

In conclusion it should be noted that the choice and setup of the block of recognition of text information is an important and responsible task. As this block is the first in the system of diagnosis of bronchopulmonary diseases, his incorrect work can lead to mistakes in all subsequent blocks of the program. As a result, at setup of the block of recognition it is necessary to pay special attention to his parameters to achieve maximum efficiency and to transfer extremely useful data to the information processing block.

References

1. Shakhmametova, G.R., Zulkarneev K. Kh., Evgrafov A.A., Analytical processing of medical data in the system of diagnosis of bronchopulmonary diseases. ITIDS'2018, International Scientific Issue, Volume 1, p.256-261 (2018)
2. Chichvarin, N.V., Recognition of images, Material from National library of N.E. Bauman, URL: https://ru.bmstu.wiki/Распознавание_образов (date of the address: 7/24/2018)
3. Zhuravlev, Yu., I., Recognition. Classification. Forecast. Mathematical methods and their application. Issue 2. M.: Science, 1989
4. Nazarov D.A., Methods of recognition of images. Basic principles, M.: Science, 2010
5. Optical character recognition, URL: https://en.wikipedia.org/wiki/Optical_character_recognition (date of the address: 7/24/2018)
6. Optical character recognition, URL: [http://wiki.technicalvision.ru/index.php/Оптическое_распознавание_символов_\(OCR\)](http://wiki.technicalvision.ru/index.php/Оптическое_распознавание_символов_(OCR)) (date of the address: 7/29/2018)
7. Gritsai A, Methods of recognition of texts, URL: <https://habr.com/post/112442> (date of the address: 8/1/2018)
8. Nagy, G., Seth, S., Viswanathan, M.: A prototype document image analysis system for technical journals. Computer 7 (1992)
9. Antonacopoulos, A., Gatos, B., Bridson, D.: ICDAR 2005 page segmentation competition. In: Proc. ICDAR, Seoul, Korea (2005)
10. Baird, H.S.: Background structure in document images. In: Document Image Analysis, World Scientific, (1994)
11. Breuel, T.M.: Two geometric algorithms for layout analysis. In: Document Analysis Systems, Princeton, NJ. (2002)
12. O Gorman, L.: The document spectrum for page layout analysis. IEEE TPAMI 15 (1993)
13. Kise, K., Sato, A., Iwata, M.: Segmentation of page images using the area Voronoi diagram. CVIU 70 (1998)
14. Shafait F., Keysers D., Breuel T., Performance Comparison of Six Algorithms for Page Segmentation (2012)
15. Kolesnikov S., Recognition of images. General information, URL: http://old.ci.ru/inform03_06/p_24.htm (date of the address: 7/26/2018)
16. Official site of the ABBYY software product, URL: <https://www.abbyy.com/> (date of the address: 8/20/2018)
17. Official site of the Office Lens software product, URL: <https://www.microsoft.com/ru-ru/p/office-lens/9wzdncrfj3t8?activetab=pivot%3aoverviewtab> (date of the address: 8/20/2018)
18. Official site of the Adobe Scan software product, URL: <https://acrobat.adobe.com/us/en/mobile/scanner-app.html> (date of the address: 8/20/2018)
19. Official site of the Free OCR to Word software product, URL: www.ocrtoword.com/ (date of the address: 8/20/2018)
20. Official site of the Online OCR software product, URL: <https://www.onlineocr.net> (date of the address: 8/20/2018)
21. Official site of the Microsoft OneNote software product, URL: <https://products.office.com/en-us/onenote/digital-note-taking-app> (date of the address: 8/20/2018)
22. Official site of the pytesseract software product, URL: <https://pypi.org/project/pytesseract/> (date of the address: 9/03/2018)

THE MULTIPROCESSOR CONTROL SYSTEM OF DYNAMICS MULTIPLY CONNECTED TECHNOLOGICAL ELECTRIC DRIVES

O. Z. Khasanov¹, Z. M. Khasanov², I.R. Akhmetzyanova³

Institute of Aviation Technologies and Materials, Ufa State Aviation Technical University, Ufa, Russian Federation¹
 Department of Computer Science and Robotics, Technology Ufa State Aviation Technical University, Ufa, Russian Federation²
 Department of Aircraft Machine Building Technology, Ufa State Aviation Technical University, Ufa, Russian Federation³
 e-mail: legnix@mail.ru, zimfirm@list.ru, iness05.17.07.93@mail.ru

Annotation: The problems of construction of practically realized adaptive control algorithms for interconnected electric drives with incomplete measurement of the state are considered. The certain results of industrial development of adaptive controllers for technological electric drives providing high efficiency in the conditions of the strengthened influence of parametric and vibration disturbances are reported. Models are developed for them, with the help of which the synthesis of a multiprocessor control system invariant to errors of both parametric and structural identification is carried out.

Design and creation of multifunctional multivariable electromechanical systems (EMS) allowing to perform diverse manufacturing operations of plasma arc spraying under different influences rendered by all holonomic and nonholonomic constraints, friction forces, elastic deformations, multivariable active forces on EMS control devices require a creation of high-speed digital adaptive automatic control systems (ACS) and control algorithms for them [1, 2].

Striving to increase an accuracy and operation speed of multivariable EMS unambiguously involves an application of fast-response noncontact DC machines with a built-in position sensor and a coordinate converter. The relevance of such machines usage is determined by their linearity, structural reliability and ability to develop high torque moments in a low-speed mode which opens wide opportunities for creation of high-speed direct-drive electromechanical systems [3, 4].

The authors of the paper together with the specialists of the E. O. Paton is Electric Welding Institute (Kiev) developed and introduced into production the high-speed mechatronic module (MM) with 5 multilinked electric drives (see Fig. 1) for technological process of plasma electric-arc spraying. The multivariable mechanisms can operate in rectangular, cylindrical systems or in combinatied coordinates system [5].

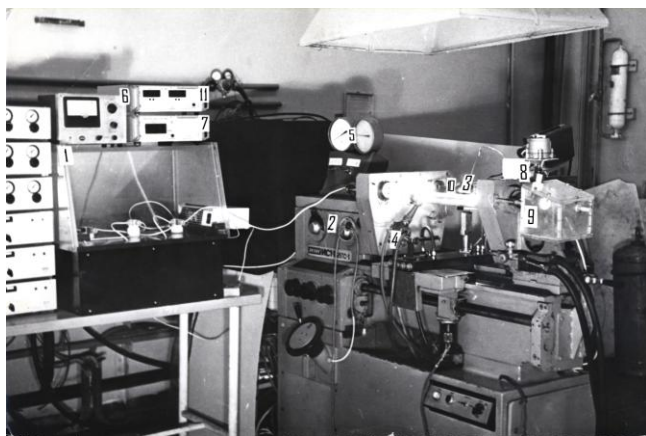


Fig. 1 Automated plasma coating device

One of the major problems encountered during the design of multivariable electric drives control systems is a restriction of dynamic parameters of the mechatronic module. This is a result of a compromise between the technical capabilities of the nodes of multivariable electric drives, physical properties of the applied materials and the performance properties of the mechatronic module (for example, minimization of the travel time or positioning time, high repeatability of the movements, etc.)

It is also necessary to consider the dynamics of multivariable electric drives actuators while constructing adaptive digital multivariable control systems. Below we consider the task of constructing and investigation the control systems of each drive of the mechatronic module with a signal adjustment according to a reference model allowing to compensate largely the dynamic effects caused by multivariable actuators arising from the joint movement of the slide assembly of plasma generator towards the movement of the work-piece.

The plasma coating equipment shown in Fig. 1 comprises the machine for plasma spraying 2, the unit of plasma-forming gas slow control 1, and the complex set of technical means of mechatronic module control system. The application of two adaptive digital control systems for rotating the work-piece and moving the slide assembly of plasma generator allows us to transfer to the next level in plasma generator control.

To implement the control laws the outer loop of each electric drive ACS contains a pulse-width position regulator selected in accordance with the requirements of a particular task. The inner high-speed loop is built as the system with the signal adjustment according to the reference model with optoelectronic speed and position sensors that transform the current position of the electric motor shaft to the appropriate voltage. The adaptive unit includes the multichannel AD converter, the adaptive interpolator and the Motorola dual-core processor MC 68302 FC20C. Having received the error signal $\Delta x(t)$ from the optoelectronic speed and position sensors concerning the model, the adaptive unit generates the adjustment signal providing the required positioning, acceleration, deceleration and stabilization algorithms.

The theory of analytical design digital controllers design using the interpolators control system can solve most of the application tasks: adaptive and modal control of electromechanical systems, optimization of actuators movement. However, in cases where it is impossible to identify coordinate and parametric perturbations in the object and to arrange a deterministic control, it is desirable to use methods of algorithmic design allowing to organize the functioning of adaptive loops based on the information accumulation about the current state of the perturbed system. One of such methods is based on the definition of Hamilton function on the movement trajectory of multivariable electric drives.

The simplest and the most widely used in simple digital controllers way to produce discrete equations using their continuous form is not suitable for digital controllers of multivariable electric drives. The known method of discretization of continuous differential equations by replacing the differentiation operation with simple difference does not provide a high accuracy of differential vector-matrix equations at low quantization frequencies. Further one of possible well developed methods of obtaining such equations based

on discretization of continuous linear equations of digital controller is considered. It represents a generalization of discretization method using a transition matrix of the control object for a nonstationary case and an apparent dependence of the vector coordinates observation on control and a range of process noise. The system of vector-matrix differential equations represented in the normal form is the initial data to obtain the necessary discrete model of the control object. Assume that the algorithm of the digital controller of multivariable electric drives ACS is described by the mathematical model in the form of linearized differential equations system in the state variables [6]

$$\Delta \mathbf{x}(t) = \mathbf{A} \cdot \Delta \mathbf{x}(t) + \mathbf{B} \cdot \Delta U(t) + \mathbf{F} \cdot \Delta \eta(t), \quad (1)$$

and the observation equation has the form of

$$\Delta y(t) = \mathbf{C} \cdot \Delta \mathbf{x}(t) + \mathbf{D} \cdot \Delta U(t) + \mathbf{W} \cdot \Delta \omega(t) \quad (2)$$

Here

$$\begin{aligned} \Delta \mathbf{x}(t) &= \mathbf{x}(t) - \mathbf{x}_3(t) \\ \Delta y(t) &= y(t) - y_3(t) \\ \Delta U(t) &= U(t) - U_3(t) \\ \Delta \eta(t) &= \eta(t) - \eta_3(t) \\ \Delta \omega(t) &= \omega(t) - \omega_3(t) \\ \Delta \mathbf{x}(t) &= [\mathbf{x}_1(t), \mathbf{x}_2(t), \dots, \mathbf{x}_n(t)]^T \\ \Delta y(t) &= [y_1(t), y_2(t), \dots, y_n(t)]^T \\ \Delta U(t) &= [U_1(t), U_2(t), \dots, U_n(t)]^T \\ \Delta \eta(t) &= [\eta_1(t), \eta_2(t), \dots, \eta_n(t)]^T \\ \Delta \omega(t) &= [\omega_1(t), \omega_2(t), \dots, \omega_n(t)]^T \end{aligned} \quad (3)$$

where $\mathbf{x}(t), y(t), U(t), \eta(t), \omega(t)$ – are the vectors of the state, observation, control, interference and noise, and $\mathbf{x}^{Pr}(t), y^{Pr}(t), u^{Pr}(t)$ – are the vectors defining the set programmed motion of the control object. The matrices $\mathbf{A}, \mathbf{B}, \mathbf{F}, \mathbf{C}, \mathbf{D}, \mathbf{W}$ have dimensions $n \times n, n \times m, p \times p, k \times n, k \times m, 1 \times 1$ respectively. The elements of these matrices are the functions of the operating mode of multivariable electric drives, the accuracy of regulator control algorithm, the metrological characteristics of used sensors and external conditions of the regulator operation. The equations (3) are given in the physical coordinates.

While developing an interpolator, the segment is considered where the following sequence is set to $t_0 < t_1 < t_2 < \dots < t_{n-1} < t_n$, where $n \in N$ and the sequence of the interpolated function values at the nodes is:

$$\begin{aligned} x_3(t_0) &= x_3(0), x_3(t_1) = x_3(1), \dots, x_3(t_{n-1}) = x_3(n-1), \\ x_3(t_n) &= x_3(n). \end{aligned}$$

Then, the following can be written to describe the dynamic motion of the object controlled by the digital controller in the state space in the vector-matrix discrete representation for the closed-loop adaptive digital controller

$$\begin{cases} x_3(n) = A(t_{n-1}, t_0) \cdot x_3(n-1) + B(t_{n-1}, t_0) \cdot U(t_{n-1}) + \\ + F(t_{n-1}, t_0) \cdot \eta_3(t_{n-1}), \\ x_3(0) = const; \end{cases} \quad (4)$$

where $x_3(n)$ – is n -dimensional state vector be the x coordinate; $A(t_{n-1}, t_0)$ – $n \times n$ matrix of the discrete system parameters at the moment t_{n-1} for the set t_0 ; $B(t_{n-1}, t_0)$ – n -dimensional control vector at the moment t_{n-1} for the set

$t_0, U(t_{n-1})$ – the control action; t_n – the discrete time, $t_n \in t_0 + jT_n, j=1,2,\dots, N$, t_0 – the time characterizing the start of the pulse width regulator; T_n – the switching period of the pulse width regulator.

The solution of equations system with the variable coefficients can be written

$$\begin{aligned} \Delta \mathbf{x}(t_n) &= \Phi(t_{n-1}, t_0) \Delta \mathbf{x}(t_0) + \int_{t_0}^{t_n} \Phi(t_{n-1}, \alpha) \mathbf{B}(\alpha) \Delta U(\alpha) d\alpha + \\ &+ \int_{t_0}^{t_n} \Phi(t_{n-1}, \beta) \mathbf{F}(\beta) \Delta \eta(\beta) d\beta, \end{aligned} \quad (5)$$

where $\Phi(t, t_0)$ – is the transition matrix.

In order to transfer the to discrete time in the continuous vector-matrix equation (5), we write the solution of equations system (1), (2), (3) for the time $t_1 = i\tau$ and $t_2 = i\tau + \tau$ (τ – is the duration of one cycle of time discretization):

$$\begin{aligned} \Delta \mathbf{x}(i\tau) &= \Phi(i\tau, t_0) \Delta \mathbf{x}(t_0) + \int_{t_0}^{i\tau} \Phi(i\tau, \alpha) \mathbf{B}(\alpha) \Delta U(\alpha) d\alpha + \\ &+ \int_{t_0}^{i\tau} \Phi(i\tau, \beta) \mathbf{F}(\beta) \Delta \eta(\beta) d\beta, \end{aligned} \quad (6)$$

$$\begin{aligned} \Delta \mathbf{x}(i\tau + \tau) &= \Phi(i\tau + \tau, t_0) \Delta \mathbf{x}(t_0) + \int_{t_0}^{i\tau + \tau} \Phi(i\tau + \tau, \alpha) \mathbf{B}(\alpha) \times \\ &\times \Delta U(\alpha) d\alpha + \int_{t_0}^{i\tau + \tau} \Phi(i\tau + \tau, \beta) \mathbf{F}(\beta) \Delta \eta(\beta) d\beta, \end{aligned} \quad (7)$$

Using the property of the transition matrix:

$$\Phi(t_2, t_0) = \Phi(t_2, t_1) \Phi(t_1, t_0) \quad (8)$$

and supposing that during one discretization interval the matrix elements $\mathbf{A}, \mathbf{B}, \mathbf{F}$ remain constant, we have:

$$\Phi(i\tau + \tau, t_0) = e^{A\tau} \Phi(i\tau, t_0) \quad (9)$$

Then equation (7) can be rewritten as:

$$\begin{aligned} \Delta \mathbf{x}(i\tau + \tau) &= e^{A\tau} \Delta \mathbf{x}(i\tau) + e^{A\tau} \int_{i\tau}^{i\tau + \tau} \Phi(i\tau, \alpha) \mathbf{B}(\alpha) \Delta U(\alpha) d\alpha + \\ &+ e^{A\tau} \int_{i\tau}^{i\tau + \tau} \Phi(i\tau, \beta) \mathbf{F}(\beta) \Delta \eta(\beta) d\beta \end{aligned} \quad (10)$$

As the matrix $\mathbf{A}$ is constant on the interval of integration, then

$$\begin{aligned} \Delta \mathbf{x}(i\tau + \tau) &= e^{A\tau} \Delta \mathbf{x}(i\tau) + \int_{i\tau}^{i\tau + \tau} e^{A(i\tau + \tau - \alpha)} \mathbf{B}(\alpha) \Delta U(\alpha) d\alpha + \\ &+ \int_{i\tau}^{i\tau + \tau} e^{A(i\tau + \tau - \alpha)} \mathbf{F}(\beta) \Delta \eta(\beta) d\beta \end{aligned} \quad (11)$$

As the control vector U and variations in interference spectrum η change in a stepwise manner at each control cycle (at this time U, η values are constant), the integral in the equation (11) can be simplified to the following form:

$$\Delta x(i\tau + \tau) = \int_0^\tau e^{A\alpha} d\alpha B \Delta U(i\tau) + \int_0^\tau e^{A\beta} d\alpha F \Delta \eta(i\tau) \quad (12)$$

where:

$$\Delta U(i\tau) = U(i\tau) - U_3(i\tau), \Delta \eta(i\tau) = \eta(i\tau) - \eta_3(i\tau)$$

Here $\Delta U(i\tau)$, $\Delta \eta(i\tau)$ – are the values of the control vector and the vector of interference spectrum at the interval $(i\tau, i\tau + \tau)$.

Proceeding to the discrete time, we obtain:

$$\Delta x(i+1) = \tilde{A}\Delta x(i) + \tilde{B}\Delta u(i) + \tilde{F}\Delta \eta(i) \quad (13)$$

where

$$\tilde{B} = \left(\int_0^\tau e^{A\alpha} d\alpha \right) B = \tau(A_0 + \frac{A_1\tau}{2!} + \frac{A_1^2\tau^2}{3!} + \dots) B \quad (14)$$

$$\tilde{F} = \left(\int_0^\tau e^{A\alpha} d\alpha \right) F = \tau(A_0 + \frac{A_1\tau}{2!} + \frac{A_1^2\tau^2}{3!} + \dots) F \quad (15)$$

Similarly (13) we will get the equation for observation vector:

$$\Delta y(i+1) = \tilde{C}\Delta x(i) + \tilde{D}\Delta u(i) + \tilde{W}\Delta \omega(i), \quad (16)$$

Thus, the equations (13) and (16) represent the system of difference vector-matrix equations in "deviations" allowing to determine the extrapolated parameters of the actuator movement at their current values and are the basis in design of digital regulator algorithm. The process of the actuators movement is smooth. Then the Kalman filter should be used in the identifier for the coordinate estimation. Here the interference and noise vectors $\eta(t)$ and $\omega(t)$ are divided into controllable $\eta_y(t), \omega_y(t)$ and non-controllable $\eta_h(t), \omega_h(t)$.

In accordance with the vectors division the corresponding separation of matrixes is conducted.

The developed adaptive control system can be easily integrated into ACS of many technological processes. The high performance of adaptive control gives an opportunity to improve the performance and reproducibility of process operations fulfillment.

Acknowledgment

This work was supported by the Russian Foundation for Basic Research grant No. 16-08-00942.

References

1. Khasanov Z. M., Khasanov O. Z., Guzairov R. M. Mathematical model of motion of multiply connected systems of electric drives for technological processes of electric arc plasma spraying // *Mechatronics, Automation, Control*, 2015, vol. 16, No. 2, pp. 116-122. (in Russian).
2. Miroshkin I. V., Nikiforov V. O., Fradkov A. A. Nonlinear and adaptive control of complex dynamic systems // Saint Petersburg: Science, 2000, 549 p. (in Russian).
3. Khasanov Z. M., Guzairov R. M. On the problem of electric drives systems design for plasma arc spraying process equipment // *USATU Vestnik*, 2013, vol. 17, No.1. pp. 146-149. (in Russian)
4. Khasanov Z. M., Guzairov R. M. On designing the wide-range electric drive for plasma electric arc spraying process equipment // *USATU Vestnik*, 2013, v.17, No.1. pp. 128-134. (in Russian)
5. Yakimovich B. A., Khasanov Z. M., Khasanov O. Z. Position-adaptive control of multiply connected electric drives of technological equipment of electric arc plasma spraying // *IZHGTU Vestnik*, 2014. No. 4. pp. 41-45. (in Russian)
6. Khasanov Z. M., Khasanov O. Z. To the synthesis of adaptive systems for automatic control of multiply connected electric drives with a reference model and with searchless algorithm of identification equipment // *USATU Vestnik*, 2017 v.21, No.4. pp. 163-172. (in Russian)

MODELLING OF THE COGNITIVE MAP FOR THE MULTIDIMENSIONAL OBJECT'S CONTROL SYSTEM

Sazonova T.V.¹, Khristodulo O.I.²

Kumertau branch of, Orenburg State University Kumertau, Russia¹

Ufa State Aviation Technical University Ufa, Russia²

e-mail: t.sazonowa@yandex.ru, o-hristodulo@mail.ru

Abstract: The instruments of synthesis in modelling of cognitive maps based on interpretation of functions characterizing the concepts and their interrelation by set of well-defined terms expressed by arguments of bivalent logic are suggested.

1. Introduction

The problem of technological processes' control in complex multiply connected, inertial, non-stationary and nonlinear objects with the distributed parameters is in that the regulation of the parameters characterizing technological process is usually made by one-dimensional proportional integral derivative (PID) controllers that are by linear control facilities [6,7,10].

Therefore, poor quality of control and as a result an increase in a share of finished goods' spoilage and unjustified energy consumption is caused. The analysis of the works devoted to automation of complex processes in a number of industries shows that potential opportunity of multidimensional PID controllers and the fuzzy logic controllers for decrease of an error regulation and mutual influence's compensation of regulation contours in the mentioned control systems and, on this basis, improvement of finished goods' quality and reduction of energy losses are almost exhausted [6,7]. Thus there is a need for development of the concept, models, methods and algorithms of the specified processes' logical control allowing to improve the quality of technological operations' control due to the experience and knowledge of subject domain experts into the correct operating algorithms and programs.

2. Theoretical basis

Cognitive modeling based on cognitive maps [1-3] allows to use incomplete, indistinct and even contradictory information of the expert in subject domain on a research's complex object for its control. Cognitive maps are used for determination of the most essential and significant interrelations, determination of the factors' value influencing the studied problem and also indications of the operating influence of necessary size, form and the sign application for the purpose of receiving the desirable results [4].

The Cognitive Computer Modelling (CCM) is the way of the analysis providing determination of force and the direction of factors influence on the transfer of a controlling object to a target state taking into account similarity and distinction of various factors influence on controlling object [5,8]. The basis of such cognitive models is usually represented by the classical cognitive map.

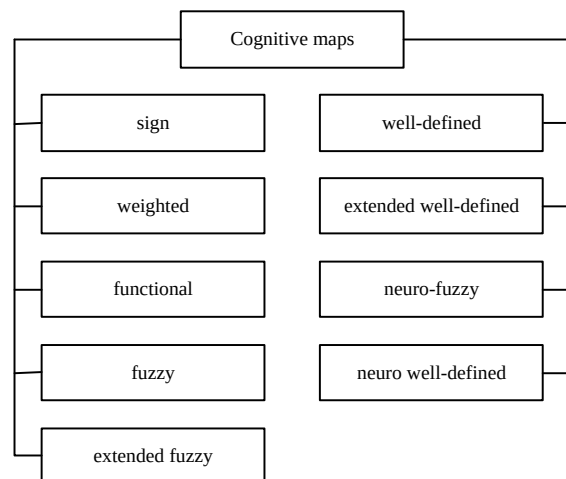


Fig.1 – Classification of cognitive maps

The analysis of literature shows that at present time the greatest distribution was gained by the following types of cognitive maps [1-3,8]: sign, weighed, functional, fuzzy, extended fuzzy and neuro-fuzzy cognitive maps.

The lack of sign, weighted and functional cognitive maps is the low accuracy of modeling which in most cases has qualitative feature. For this reason decrease in finished goods' consumer properties and also an excessive consumption of raw materials and energy in industrial technological installations which control systems are constructed on the basis of cognitive maps is observed. The advantage of fuzzy, extended fuzzy and neuro-fuzzy cognitive maps is the possibility of verbally presented algorithms modeling of the complex systems functioning comprised of various physical nature subsystems and, thus, to make use of unique professional experience of subject domain experts for creation of intellectual control systems. However fuzzy, extended fuzzy and neuro-fuzzy cognitive maps haven't found broad application in the control systems significant for modern society processes, demanding high precision (deep oil refining, nano - and chemical technologies, etc.) because of a big error of the defuzzification procedure.

The basic notion of the cognitive maps theory is the concept. A concept is the basic (indivisible) element of the considered system [1-3,5]. Let K is a set of concepts (elements) of the considered system, and a set of relations, each of which describes influence force of one concept (concept reason) on other concept (concept investigation). The orientation of this relation means that the concept source influences a concept receiver, i.e. change of values (states) of a concept source leads to change of values (states) of a concept receiver.

The activity of experts and analysts directed to a research of a situation and decision-making by means of cognitive maps represents methodology that is logic-time structure of various methods and receptions application: creation of the cognitive map, its parameterization, obtaining of situations' development forecasts, verifications, cognitive map's updating and decision-making [3,5].

The situation of the cognitive map is designed on the basis of experts' poll carried out in several phases. The first stage allows to define basic factors and to determine the value of relations. Then local cognitive maps are designed. After on the basis of expert estimates local cards brought together in one cognitive map of a situation. Accumulation of knowledge in the studied situations allows to reveal bonds character [4,5,9].

In the situation's analysis two types of tasks are solved: statistical and dynamic. The statistical analysis considers the current situation with influence of factors at each other, situational stability in general and search of structural changes for obtaining system's stability.

The dynamic analysis serves for generation and the analysis of possible changes in situation's development over time. In the dynamic analysis tasks indistinct sizes are attributed not only to relations, but also to factors. At the same time, if values of relations in the course of the analysis are considered as constants, then the size attributed to a factor changes over time. The size of a factor allows to estimate influence force on a factor and to express result of total influences as a factor's concrete value. The concept of a situation's condition gives the chance to speak about situation's development over time under the influence of various external influences which are expressed in change of factors' values i.e. to

set a forecast task (a direct task) and also to investigate possibilities of situation controlling, i.e. to look for the influences leading to the necessary (target) state (the return task).

2.2. Principles of control

Let's consider the principles of multidimensional (complex) objects' control on the basis of cognitive maps, with interpretation of concepts and relations between them by a set of arguments of two-valued logic on the example of a prepreg manufacturing with use polymeric binding and woven filler [4,5,8,9].

The device of models is necessary for designing of complex technological objects' controlling system. It enables to describe model of control in the form of rules and algorithms of the operating influences formation, and to represent this model in the course of the operation for the operating influences formation. At the heart of such device of models it is rational to use the principle of control according to situational model in which management process is considered as serial change of situations, certain controlling effects are connected with each of which. At the same time the operating system on the basis of situations' model has to monitor the current situation and form operating influences appropriate to this situation. The model can be presented in the form of the cognitive map which describes situations, transitions and corresponding control instructions, and is aprioristic information for the organization of controlling process [5,9,11].

Such cognitive map will represent hierarchy of states with the set rules of interpretation. It combines multiple levels realized by the addition to model's conditions of any level of other model considered as internal for an initial state and also the multidimensionality realized by introduction of several internal models for initial states.

Values of the parameters of the cognitive map specifying representation of model for direct control it is set in the form of the automatic machine of the special type which is carrying out model processing on the basis of its recursive bypass on active arches and also interaction of elementary models on the basis of current states checking.

Multidimensional hierarchy is more powerful modulating system in comparison with one-dimensional.

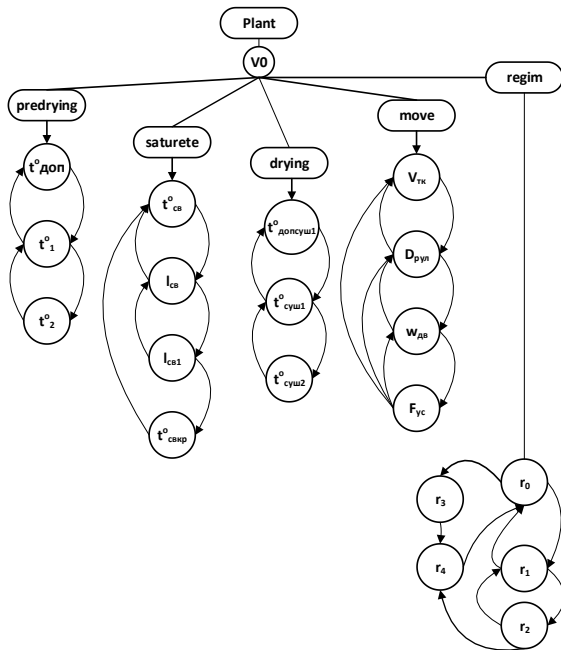


Fig. 2 – Cognitive map of multidimensional hierarchical model of a prepreg's production

The multidimensional model of a prepreg manufacturing (fig. 2) reflects four interconnected processes, each of which is substantially independent:

- subdrying of fabric for an oil agent elimination;

- the fabric saturation which is characterized by indicators of binding;
- the fabric drying which is characterized by indicators of flying agents removal out of fabric;
- fabric moving necessary for maintenance of fabric's movement speed and taking-up of a ready prepreg to a roll.

The Dryer model reflects a condition of fabric's subdrying. Conditions of this model consider: admissible temperature (t^o admis.); the precritical temperature which is characterized by emergence of temperature increase (decrease) above admissible (t^o1); the critical temperature leading to accident in technological process (t^o2).

The Saturate model reflects fabric's saturation process by binding. The condition of this model considers: temperature of binding (t^o bind.), level binding in an impregnating bathtub (l bind.); level of the binding, its characterizing reduction in an impregnating bathtub and replenishment from a tank (lbind.1); decrease in temperature binding below admissible that can cause violation of technological process (t^o bind. crit.).

The Dryer model reflects a condition of fabric's drying. Conditions of this model consider admissible temperature on zones of drying (t^o admis. drying); the precritical temperature which is characterized by emergence of temperature increase (decrease) above admissible (t^o drying1); the critical temperature leading to accident in technological process (t^o drying2).

The Move model reflects the speed of the fabric's movement taking into account change of a roll's diameter at taking-up: admissible speed of the fabric's movement (V fabric); the change of a roll's diameter (D roll) leading to change of the moment on the device of taking-up of a ready prepreg that leads to change of engine's angular speed (w eng.); the critical tension of a prepreg leading to break of a prepreg (F ang. speed) and emergency.

The Regime model reflects the generalized condition of installation in relation to critical parameters of subdrying and drying processes and emergency operation. Its state consider: normal mode ($r0$); the precritical mode – increase (decrease) in temperature is higher than admissible ($r1$); the critical mode – the mode of an emergency ($r2$); break of a prepreg ($r3$); the postcritical mode – elimination of malfunction in work of installation ($r4$).

The Regime model arches providing transitions of a state have the activity predicates determined by a condition of other models.

All processes are connected among themselves for object controlling and are characterized by independently measured parameters and can be controlled independently of each other [4].

The arch predicate ($r0, r1$) is true if the Dryer model is in a state (t^o drying1) and the Saturate model of being able (t^o bind. critical).

$P01 = \text{Path} (1, \text{Dryer}, t^o\text{drying1}) \wedge \text{Path} (1, \text{Saturate}, t^o \text{ bind. critical}).$

The arch predicate ($r1, r2$) is true at change of temperature above (below) of critical:

$P12 = \text{Path} (1, \text{Dryer}, t^o\text{drying2}) \wedge \text{Path} (1, \text{Dryer}, t^o2) \wedge \text{Path} (1, \text{Saturate}, t^o \text{ binding critical}) \wedge \text{Path} (1, \text{Move}, F \text{ ang. speed}).$

Arch predicates ($r0, r3$) it is true at a prepreg's break:

$P03 = \text{Path} (1, \text{Move}, F \text{ ang. speed}).$

Arch predicates ($r3, r4$) and ($r2, r4$) are true in the absence of a prepreg's break and change of temperature above (below critical):

$P34 = P03 = P12.$

The arch predicate ($r4, r0$) is true if processes of Dryer, Saturate, Move is normal:

$P40 = \wedge \text{Path} (1, \text{Regime}, r0).$

The analysis of the cognitive map begins with definition of factors' total influence at each other, considering, both direct influence, and mediated when one factor influences another through a chain of intermediate factors. For this purpose at first the cognitive map is presented in the form of a connectivity matrix (table 1) in which the values of direct connections between factors are placed. Then for it the matrix of transitive short circuit on a formula is defined:

$$R = V_{i=1}^{\infty} \cdot R^i \quad (1)$$

where R is a connectivity matrix.

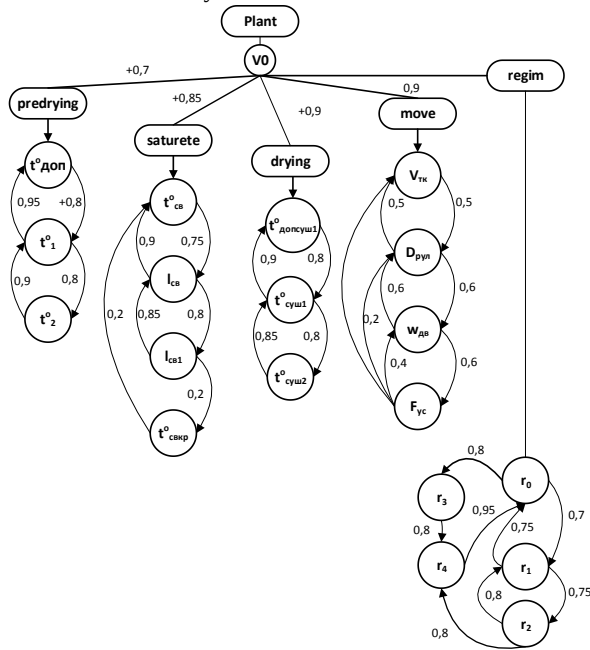


Fig. 3 - Fuzzy cognitive map of multidimensional hierarchical model of a prepreg's production

Table 1 – Matrix of processes' connectivity

subdrying	t _{admis.}	t ₁	t ₂
t _{admis.}	0	0,8	0
t ₁	0,95	0	0,8
t ₂	0	0,9	0

saturation	t _{cb}	l _{cb}	l _{cb1}	t _{cbkp}
t _{binding}	0	0,75	0	0
l _{binding}	0,9	0	0,8	0
l _{binding1}	0	0,85	0	0,2
t _{binding critical}	0,2	0	0,2	0

drying	t _{admit. drying1}	t _{drying1}	t _{drying2}
t _{admit. drying1}	0	0,8	0
t _{drying1}	0,9	0	0,8
t _{drying2}	0	0,85	0

movement	V _{fabric}	D _{roll}	W _{engine}	F _{ang. speed}
V _{fabric}	0	0,5	0	0
D _{roll}	0,5	0	0,6	0
W _{engine}	0	0,6	0	0,6
F _{angular speed}	0,2	0,2	0,4	0

Table 2 – Full matrix of interference

subdrying	t _{admis.}	t ₁	t ₂
t _{admis.}	0	0,8; 0,95	0
t ₁	0,95; 1	0	0,8; 0,85
t ₂	0	0,9; 0,95	0

saturation	t _{binding}	l _{binding}	l _{binding1}
t _{binding}	0	0,75; 0,85	0
l _{binding}	0,9; 1	0	0,8; 0,9
l _{binding1}	0	0,85; 0,95	0
t _{bind.critical}	0,2; 0,4	0	0,2; 0,4

drying	t _{admis. drying1}	t _{drying1}	t _{drying2}
t _{admis. drying1}	0	0,8; 0,95	0
t _{drying1}	0,9; 1	0	0,8; 0,95
t _{drying2}	0	0,85; 0,95	0

movement	V _{fabric}	D _{roll}	W _{engine}	F _{ang. speed}
V _{fabric}	0	0,5; 0,75	0	0
D _{roll}	0,5; 0,6	0	0,6; 0,8	0
W _{engine}	0	0,6; 0,9	0	0,6; 0,8
F _{ang. speed}	0,2; 0,4	0,2; 0,3	0,4; 0,5	0

According to the analysis of processes' physical model influence of some parameters on a prepreg manufacturing process is revealed. Factors which influence are significant for output parameters are fabric subdrying temperature; fabric drying temperature; speed of the fabric's movement.

The influence of such factors as a speed of the fabric's movement, binding's density and temperature, fabric's humidity, ambient temperature is statistically insignificant. It is connected with the fact that the range of these factors variation is very limited.

For an illustration of well-defined terms use in cognitive maps in fig. 4 the universal numerical axis of parameter p which represents with located on its n of accurate terms T1 – Tn with l width is depicted. For example, the Wij function of influence of a concept of i on j concept can be such parameter.

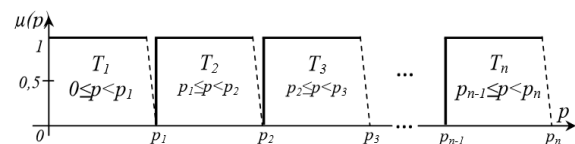


Fig. 4 – Well-defined terms of parameter p on a universal numerical axis

As T1 terms – Tn represent accurate sets, their functions accessories have a rectangular shape. And the value of these functions is equal to unit for all accurate values of physical quantity in each of the mentioned sets.

Influence are presented in well-defined cognitive maps of the Wij function by set of accurate terms (fig. 4). Here i and j – any couple of concepts entering the considered cognitive model.

Analytical expression a term set of influence function of a concept of i on a concept of j is the following:

$$T(W_{ij}) = \sum_{s=1}^n W_{ijs} ((i-1)l \leq p < il) \quad (2)$$

where n is quantity of well-defined terms with width of l used for Wij function interpretation by set of well-defined terms.

According to expression (2) influence extent of a concept of i on a concept of j will be defined by a term of W_{ijs} which in present moment is equal to logical unit. And such property in total of accurate terms only one term has. The parameters characterizing concepts well-defined cognitive maps are usually represented in the form of constants or the determined analytical expressions [10-12].

3. Conclusion

As well-defined terms by the logical nature are arguments of binary logic, it is offered to make concepts' relations and their parameters by the corresponding set of functions of Boole's algebra which arguments can be well-defined terms, output signals of timers and counters and also discrete variables from other external devices. Thereof there is an opportunity in developing cognitive maps to use such highly effective instruments of synthesis as the sequential equations and Quine-McCluskey minimization method. In addition the use of well-defined terms significantly simplifies such basic procedures relating to cognitive maps as the description of concepts' state and interference, accumulation of several concepts influence on one concept and modeling of system dynamics.

Cognitive maps with interpretation of the parameters characterizing concepts state and their influence functions, set of well-defined terms (they have a rectangular shape of accessory function and by the logical nature are arguments of bivalent logic) allow to increase the accuracy and speed of cognitive models. They simplify the description of concepts' states and interference, accumulation of several input concepts influence on one output and also dynamics modeling and the stability analysis.

Literature

1. Borisov V.V., Fedulov A. S. The generalized fuzzy cognitive maps // Neurocomputers: development, application. 2004. No. 4. P.3-20.
2. Kosko, B. Fuzzy Cognitive Maps // International Journal on Man-Machine Studies. 1986. Vol. 24. P. 65-75.
3. Kayashev A.I., Sharipov M.I., Muravyeva E.A., Bagrov K.A. Cognitive modeling of petrochemical objects' control with interpretation of concepts and relations between them by a set of well-defined terms // the Proceedings of XII All-Russian meeting on problems of management, VSPU-2014 Moscow, June 16-19, 2014.
4. Polyakova L.Yu. Technological process' control in a prepreg manufacturing for the aircraft equipment products // the Abstract of the thesis for a degree of Candidate of Technical Sciences: 05.13.06. – Ufa, USATU, 2007.
5. Kulinich A.A. Cognitive maps in support of decision-making, // Proceedings of the Congress on Intellectual systems and information technology IS&IT11. M.: Fizmatlit. 2011. T. 1. 598 p.
6. Vasilyev V.I., Ilyasov B.G. Intelligent control systems. Theory and practice: educational allowance. Moscow: Radio Engineering, 2009. 392 p.
7. Vasiliev V.I. Intellectual protection of information: Textbook. allowance. Moscow: Mashinostroy-152 p.
8. Kosko, B. Fuzzy Cognitive Maps // International Journal on Man-Machine Studies. 1986. Vol. 24. P. 65-75.
9. Stylios C.D, Groumpos P.P. A Soft Computing Approach for Modelling the Supervisor of Manufacturing Systems // Intelligent and Robotic Systems. 1999. Vol. 26. P. 389-403.
10. Sazonova T.V., Bogdanov A.V., Bondarev A.V. Parameters of the quality and power efficiency of the automatic controlling the steam boiler with the indistinct adaptation of the PID-adjuster in the fuzzyTech's environment // International technical conference: articles of the international technical conference. - Chelyabinsk: SUSU, 2017. - P. 48 - 53.
11. Sazonova T.V., Khristodulo O.I., Makhmutov A.A. Use algorithm based at hamming neural network method for natural objects classification // Procedia Computer Science 12th. Cep. "12th International Symposium Intelligent Systems, INTELS 2016". - 2017. – P. 388-395.
12. R. Axelrod. The Complexity of cooperation: Agent-Based-Models of Competition and Collaboration. Princeton: Princeton University Press, 1997

CLUSTER ANALYSIS TO IDENTIFY POTENTIAL EMPLOYERS OF UNIVERSITY GRADUATES

A.G. Karamzina

Department of Computer Science and Robotics, Ufa State Aviation Technical University, Ufa, Russia

e-mail: karamzina@tc.ugatu.ac.ru

Abstract: The results of cluster analysis are presented, the purpose of which was the verification of the base of industrial enterprises of the Republic of Bashkortostan to identify potential employers of graduates by 27.03.00 on the enlarged group of specialties and directions "Control in technical systems". The solution of the problem will allow to interact with potential graduates' consumers at the stage of developing higher education educational programs, which will ensure high-quality training of highly qualified personnel in accordance with the requests of enterprises and organizations of the region and, accordingly, further employment of graduates. The work analyzes the conditions for the formation of clusters, presents a sample of data for intellectual analysis, results of component and cluster analysis.

1. Introduction

As part of the transformation of Russian education into the Bologna system and, first of all, the adoption of federal standards, one of the most important directions of the activity of higher education institution was the formation of effective models for interaction between the university and a potential employer.

Interaction of the university with a potential employer should be conducted already at the stage of developing the basic professional educational program. According to the methodological recommendations for the development and implementation of higher education educational programs at the bachelor's level [7], when developing and implementing educational programs of applied bachelor's degree, it is recommended to use various models of interaction with potential employers, including cluster models.

The use of cluster models has already taken a key position in research [1]. The relevance of cluster models is due to the fact that the cluster approach is considered the most effective tool in the management system.

To implement high-quality interaction with employers in the process of implementing educational programs of higher education, it is necessary to solve the problem of identifying potential employers. To solve this problem, the author proposes to apply the method of cluster analysis. Clustering will allow you to break up multiple objects (potential employers of the region) into groups - clusters. Within each cluster will be those enterprises and organizations whose interaction with them will have the most positive effect in the development and implementation of undergraduate programs, depending on the needs of the region in graduates in three areas of training, within the enlarged group of specialties and directions "Management in technical systems":

- Quality control (QC);
- System analysis and management (SAM);
- Management in technical systems (MTS).

Analysis of the conditions for the formation of clusters is performed by the method of principal components and by cluster analysis.

2. Component analysis

Method of main component (MC) refers to a group of data visualization methods that allow you to identify data structures based on clarifying the relationships between objects and their characteristics. The result of the application of the method is the graphical mapping of a set of objects to a new coordinate space, and the mapping maximally reflects the features of the distribution of objects in the original multidimensional space [2, 4]. The application of linear methods of diminishing the dimension allows you to visualize multidimensional data by building a new coordinate space in which each coordinate axis is a linear combination of the original characteristics that correspond to the hidden characteristics present in the data. In addition to solving the visualization problem, linear methods of diminishing the

dimensionality allow solving the problem of classification of initial features - to determine the interrelations between the characteristics by combining significant characteristics into groups and forming new integral features.

Method of main component allows, based on the analysis of n objects, each of which is characterized by p initial characteristics $x_1, \dots, x_p$, to construct the p major components $F_1, F_2, \dots, F_p$ (the coordinate axes of the new space), the new coordinate system being a system of orthonormal linear combinations. Each i -th coordinate axis is a linear combination of the original features and is written as:

$$F_1 = \omega_{11}(x_1 - m_1) + \omega_{21}(x_2 - m_2) + \dots + \omega_{p1}(x_p - m_p),$$

where ω_{ji} – the weight coefficient determining the contribution of the i -th features to the formation of the j -th component, m_i – the mathematical expectation of the i -th features.

Advantages of the method are the least distortion of the structure of the original objects when they are projected into a space of smaller dimensions, and also the possibility of using data for analysis in combination with other methods of data exploration. The disadvantage is the possibility of a situation when the weight coefficients have values close to each other, which leads to a weak interpretation of the result obtained. This problem is solved by applying other types of analysis.

As objects n , regional industrial enterprises (potential employers) are considered in the amount of 58 pcs.

The following data are considered as feature of p :

- x_1 – demand for graduates;
- x_2 – professional growth (further training, training at the enterprise);
- x_3 – production sites (areas not occupied by main production, used directly for training students);
- x_4 – equipment (production equipment for training and honing skills in practice for students);
- x_5 – software;
- x_6 – technologies;
- x_7 – human resources (availability of competent managers of practices).

The first main component of $F_1(X)$ corresponds to the largest eigenvalue and is calculated as a linear combination of the original features, which has the largest variance. Thus, the first principal component is taken along the direction with the maximum variance. The second main component of $F_2(X)$ lies in a subspace perpendicular to where the first main component is located. Within this subspace, the second major component is taken along the direction with the maximum variance. Then the third main

component $F_3(X)$ is in the direction of the greatest dispersion in the subspace perpendicular to the first two, and so on.

The component analysis is carried out by means of the package STATGRAPHICS, which has the following advantages [7]:

- a combination of scientific methods for processing heterogeneous data with the possibility of creating modern high-quality interactive graphics;
- wide possibilities of interaction with other software products (spreadsheets, databases);
- high-quality two-dimensional and three-dimensional graphics,
- integrated graphics (all elements of graphical representations of analysis results can be converted).

The initial data for the intellectual analysis of data on seven feature (x_1-x_7) in three areas of preparation (QC, SAM, MTS), a STATGRAPHICS spreadsheet was introduced. To determine the value of each criterion for industrial enterprises, the deputy heads of the department for academic work in the relevant area of training for bachelors: 27.03.02 "Quality control" (QC), 27.03.03 "System analysis and management" (SAM), 27.03.04 "Management in technical systems" (MTS), chairman and members of the scientific and methodological council on the enlarged group of specialties and directions 27.00.00 "Management in technical systems" (50% of the size of which are representatives of employers).

The obtained data indicate that already two main components describe about 92% of the variance of the initial data. The third main component adds about 3% of the dispersion. As a result, a total of 95% of the variance.

The dispersion diagram of the entire set of objects on the plane of the two main components is shown in Fig. 1-3.

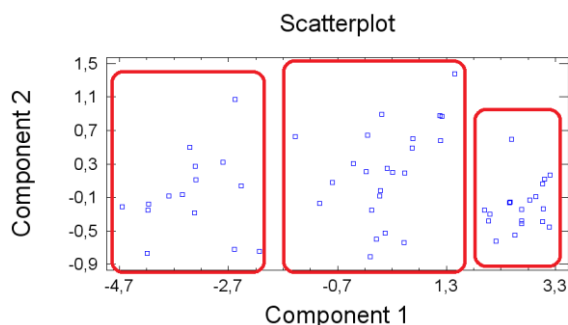


Fig. 1. Projection of the investigated objects on the space of two MC for QC

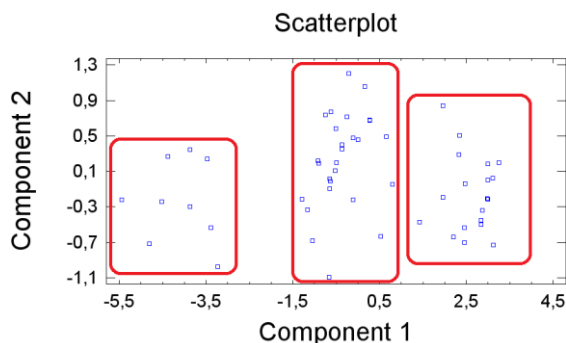


Fig. 2. Projection of the investigated objects on the space of two MC for SAM

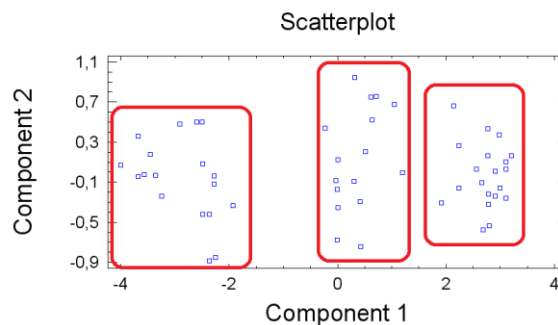


Fig. 3. Projection of the investigated objects on the space of two MC for MTS

On the presented projections, three classes can be clearly traced.

3. Cluster analysis

The next step is a cluster analysis, which is applied to the partitioning of a set of objects (industrial enterprises) into a given number of clusters (three) obtained on the basis of the method of main component.

Cluster analysis is designed to split a set of objects into a previously unknown or, in rare cases, a given number of clusters on the basis of some mathematical criterion for clustering [5].

Hierarchical methods of cluster analysis are designed to obtain a visual representation of the stratification structure of the whole set of objects under study. The rules are the criteria used to solve the question of the "similarity" of objects when they are combined into a group - agglomeration methods that are constructed on the basis of a consecutive combination of objects into groups and a corresponding decrease in the number of clusters. At the beginning of the algorithm, all objects are separate clusters, then sequentially at each step of the algorithm operation the most similar ones are combined into a cluster until all objects form one cluster. The result of hierarchical methods of clustering is a dendrogram - a tree diagram containing several levels, each of which corresponds to one of the steps in the process of sequential cluster.

The most common agglomerative algorithms are: the single-link method, the method of complete communication, the mean-coupling method, the Ward method [3]. In this paper, the Ward method was used as the clustering method, since in our case it is desirable that the clustering algorithm works well with a small number of observations.

The results of constructing dendrograms displaying the hierarchical structure of grouping of objects in three directions of preparation (QC, SAM, MTS) for three clusters are presented in Fig. 4-6.

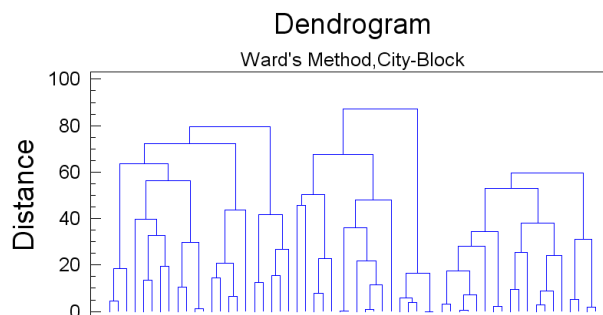


Fig. 4. Dendrogram for three clusters for QC

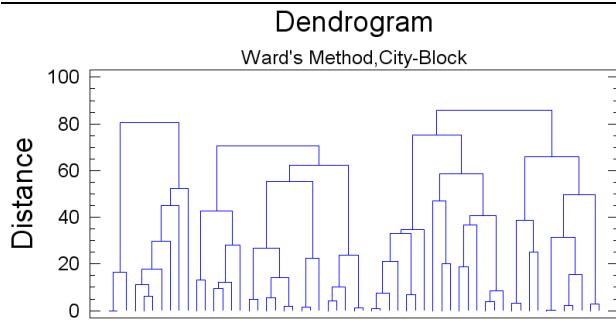


Fig. 5. Dendrogram for three clusters for SAM

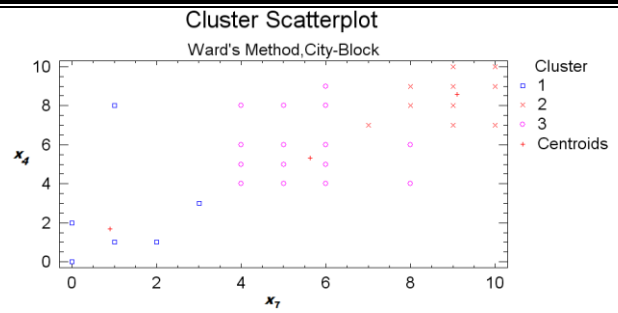


Fig. 8. The two-dimensional dispersion diagram for SAM

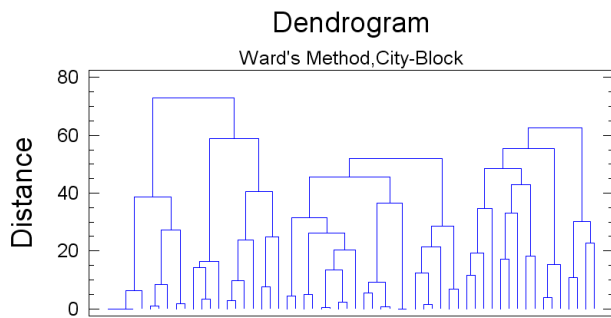


Fig. 6. Dendrogram for three clusters for MTS

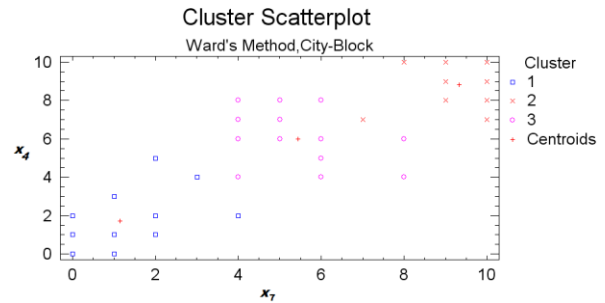


Fig. 9. The two-dimensional dispersion diagram for MTS

Based on the summary of cluster analysis, the rules for the formation of clusters.

Rule 1 for cluster 1: if the human resource = low; technology = low; PO = low; equipment = low; production site = low; professional growth = low; demand = low, then cluster 1.

Rule 2 for cluster 2: if the human resource = high; technology = high; software = high; equipment = high; production site = high; professional growth = high; demand = high, then cluster 2.

Rule 3 for cluster 3: if the human resource = average; technology = average; software = average; equipment = medium; production site = medium; professional growth = average; demand = average, then cluster 3.

The dispersion diagram, showing how the observations under study are grouped on the plane of two variables x_4 and x_7 , for each areas of preparation is presented in Fig.7-9. Each cluster is indicated in the diagram by its own symbol.

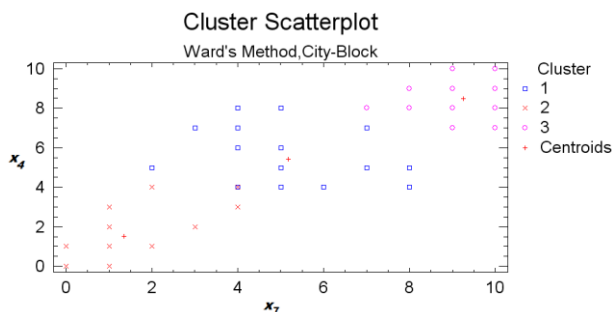


Fig. 7. The two-dimensional dispersion diagram for QC

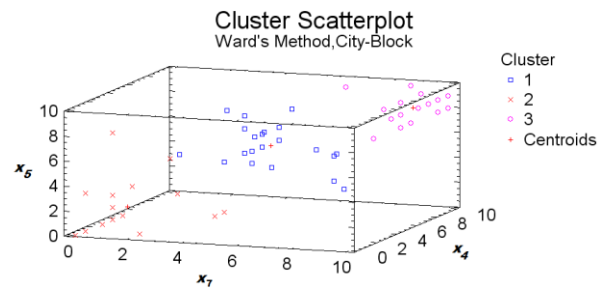


Fig. 10. The three-dimensional dispersion diagram for QC

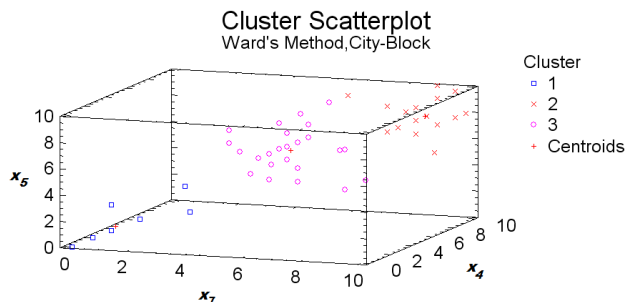


Fig. 11. The three-dimensional dispersion diagram for SAM

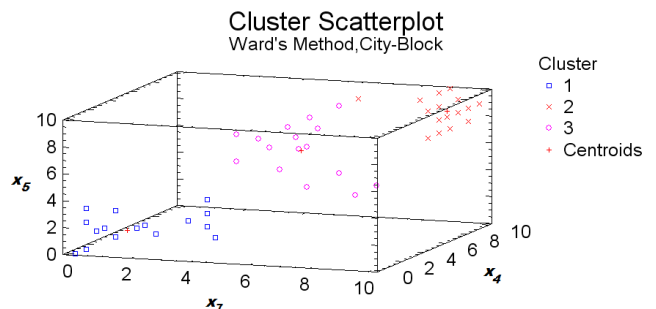


Fig. 12. The three-dimensional dispersion diagram for MTS

Based on the analysis carried out, a model of the demand for graduates at industrial enterprises of the Republic of Bashkortostan (Fig. 13) is constructed. Enterprises are selected according to the second rule of cluster analysis and display the highest indicators of the selected characteristics.

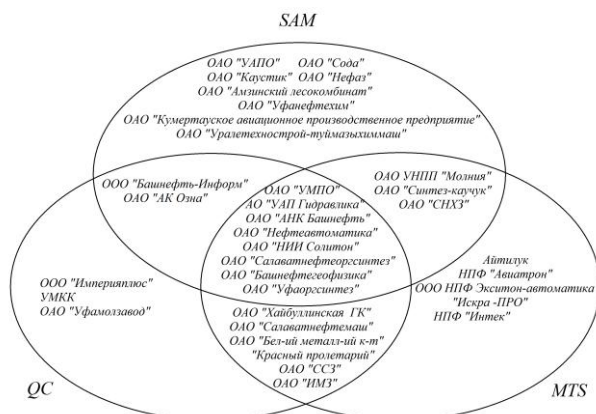


Fig. 13. Model of demand for graduates in industrial enterprises of the Republic of Bashkortostan

4. Conclusion

The paper presents the results of the component and cluster analysis, which allowed:

- identify potential employers interested in training graduates in a large group of specialties 27.03.00 "Management in technical systems" in the areas of bachelor's training 27.03.02 "Quality management",

27.03.03 "System analysis and management", 27.03.04 "Management of informatics in technical systems";

- build a model of the demand for graduates in these areas at industrial enterprises of the Republic of Bashkortostan.

References

1. Beljakin S.K., Teben'kova E.A. "Cluster model of university interaction with the professional community". *Vector of science TSU* 2013; 2 (24): 399-402.
2. Dubrov A.M., Mhitorjan V.S., Troshin L.I. "Multidimensional statistical methods", Moscow, Finance and Statistics Publ. House, 2000.
3. Djuk V., Samojlenko A. "Data Mining", StPetersburg: "Peter" Publ. House, 2001.
4. Kalinina V.N., Solov'ev V.I. "Introduction multivariate statistical analysis", Moscow, State University of Management Publ. House, 2003.
5. Korneev V.V., Garre A.F., Vasjutkin S.V., Rajh V.V. "Database. Intellectual information processing", Moscow, "Nolizh" Publ. House, 2001.
6. Methodical recommendations for the development and implementation of higher education educational programs at the undergraduate level. Type of educational program "Applied Bachelor". Approved 11.09.2014 № AK-2916/03.
7. Simchera V.M. "Methods of multivariate analysis of statistical data", Moscow, Finance and Statistics Publ. House, 2008.

EXAMPLE OF PLM-SYSTEM ADOPTION AT PJSC «UEC-UMPO» IN THE NETWORK OF INTERACTION ON THE PROJECT PD-14

Kuznecov A.A., Mavrina A.S., Sapozhnikov A.Yu.

PJSC «UEC-UMPO», Ufa, Russia

e-mail: kuznecovopkr@gmail.com, nytko_008@mail.ru, axl_mail_box@mail.ru

Abstract : The article describes the characteristics of the PLM-system, implemented on PJSC «UEC-UMPO». The concept of information support for project management is shown, the scheme of interaction of project participants is presented, the main tasks for the future are revealed.

1. Introduction

PJSC «UEC-UMPO» – one of the largest manufacturers of aircraft engines in Russia. At present, PJSC «UEC-UMPO» participates in cooperation with JSC «UEC-Aviadvigatel» in the project of creating PD-14 engine for the aircraft MS-21. The PD-14 project is notable for its branching, which has become the prerequisite for a creation of the Unified information space (UIS). As part of this task, a unique PLM (product lifecycle management) solution was developed and implemented, serving a geographically dispersed group of enterprises, which includes customers, the lead developer, a number of design bureaus and serial plants, and co-executors of the project.

Work on the project PD-14 began in 2009, initially the work in the UIS was governed by the statement of work. The document completely defined the form of the project support information system, and each participant of the cooperation had to fully ensure that its IT infrastructure met these requirements.

At the moment, the interaction of enterprises in the design of aircraft gas turbine jet engines is as follows: there is a database at the site of each participant, the data is exchanged through a distributed system for exchanging engineering data in the Teamcenter MultiSite Collaboration environment (Figure 1).

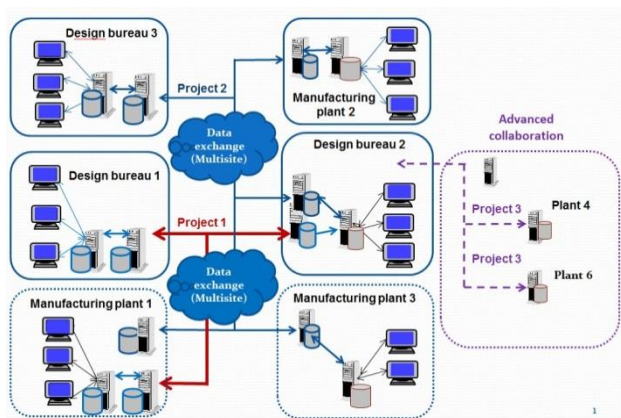


Figure 1. The scheme of interaction of the enterprises in the design of GTE

2. Project charter

At the first stage of the project, the levels of IT infrastructure, methods of computer-aided design and data management in cooperative enterprises were significantly different, and one of the first tasks was leveling of their potentials. IT-infrastructure in that enterprises has been rebuilt in short time.

The main principles of conducting this project were formulated:

1. Centralized storage of project data in a unified engineering data management system (PLM Teamcenter).
2. Providing collective interaction between project participants based on Teamcenter Multisite Collaboration technology.
3. Centralized management using the distributed project management system MS Project.
4. Use of videoconferencing in order to reduce travel expenses.
5. Applying data protection technologies for transmission between enterprises.

3. Organization of the UIS of project PD-14

To create a UIS, product lifecycle management system is needed. In the project of PD-14 creation, the PLM-system Teamcenter acts as this way. When it was put into commercial operation, the work of the whole design bureau changed, end-to-end design process has appeared.

Teamcenter PLM Software allows [3]:

- a) create and use a unified database of products and processes [4];
- b) organize a joint coordinated work of the designers on a unified electronic layout of the product;
- c) organize parallel work on design and technological departments on the product [5];
- d) use software licenses of different types (depending on the program module users are working with);
- e) formalize the processes of working with 3D-models.

The work begins with the "Structure Manager" application - one of the main applications for managing the product structure [5]. The application allows:

- configure the composition of the product;
- support the basic functions of editing a structure such as adding an element, deleting an element and change the attributes of an element's entry (count, position number and other);
- create a specification without using CAD in Teamcenter as a PDF file;
- use variant rules, which make it easier to maintain group assemblies.

The integrated CAD system NX is used for the design of 3D models.

4. NX product design software

NX product design software from Siemens PLM Software is a system for product design and calculations. NX is a 3D modeling system that allows to create products of any complexity. For designating software of this class, the abbreviation CAD/ CAM / CAE is used [2]. Prior to the introduction of NX and TCE, various CAD software (AutoCAD, Compass and even the drawing boards) could be used in one project. The work of the design engineer for different nodes was conducted on the file structure without the use of collective handling, which made it difficult to work with the current version of the frequently-modified model. It was not possible to see the complete construction of the node at the current moment, the complete node was assembled by one performer at the final stage, after the linker passed all the component parts of it. The use of the unified platform allowed to optimize data flows transmitted between specialists, to avoid unnecessary transformation processes from one system to another. The obvious advantages of implementing the NX software in the design bureau:

- a) independence from hardware and software;
- b) absence of problems in process of the exchange of geometric data;
- b) an extensive set of modules and ready-made solutions;
- r) better functionality in the development of control programs for CNC machines;

- д) The best functional capabilities in the design of parts, assemblies and for the creation of design and technological documentation;
- е) the possibility of accumulating and reusing development elaborations.

5. Information evolution of the project

The IT infrastructure of the project was based on the systems Teamcenter and NX. Gradually, the project moved to newer software versions. Transitions, as a rule, were caused by the requirements of the lead developer (JSC «UEC-Aviadvigatel»), as well as by the termination of technical support for existing software versions.

As the project progressed, the number of tasks provided by the Teamcenter PLM system increased (Figure 2).

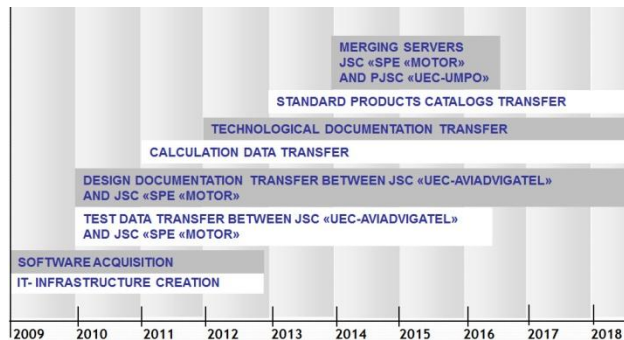


Figure 2. Bureau Informational Evolution of the project

In 2014, in connection with the restructuring of JSC «SPE «Motor» and its accession to PJSC «UEC-UMPO», a decision was made to merge the two servers.

Currently, as new versions of software are introduced, specialists are trained that allows to expand and deepen the use of NX and Teamcenter, they get necessary knowledge and tools for creating electronic layout of the product and data management. At the same time, instructions have been developed and are being implemented that define the requirements for working in NX under the management of PLM-system Teamcenter.

6. Architecture of interaction between project members

As the main architecture of interaction between the project members, "site-leader" approach was chosen. A website of the project was created (a "site-leader") on the territory of the lead developer of JSC «UEC-Aviadvigatel» in Perm. On the space of each member, the PLM server is deployed (Teamcenter). After that, communication channels were established and configured between the project server and each performer. The interaction is carried out using the technology of work through the MultiSite Collaboration implemented in a classic form. Communication channels are protected and encrypted. Initially, the "open VPN" software was used, and then the government-approved software and hardware complex for cryptographic data protection, providing network security for the corporate network of any topology with any number of tunnels. This system was described in the specifications of the main developer, and each company was obliged to purchase and to install it.

The data transfer is carried out as follows: the design documentation is developed and approved on the enterprise server, after which the data exchange operator transfers the documentation to the site-leader and notifies the recipient's enterprise. The recipient member takes the necessary data from the "site-leader".

Compared to the organization of the UIS without using the site-leader, the interaction scheme described above has several advantages:

- a) minimization of the costs of maintaining communication channels at each enterprise (compared to "each with each" network);
- б) to obtain the necessary information, each performer can search directly on the «site-leader», rather than search for it on the sites of the co-executors.

7. Project Infrastructure

To organize the described architecture on servers of all enterprises, a unified data model was deployed.

A Data Model is a way of presenting an information model of a product in a computing environment. When creating a data model in PD-14 project, a document-oriented approach was used, that is, each type of project documentation was represented with (one or another object).

One of the key tasks of the project was the organization of work with reference information. A centralized database of reference information was created, which included standard products (SP), purchased products (PP), catalogs of materials for the manufacture. The interaction scheme is shown in Figure 3.

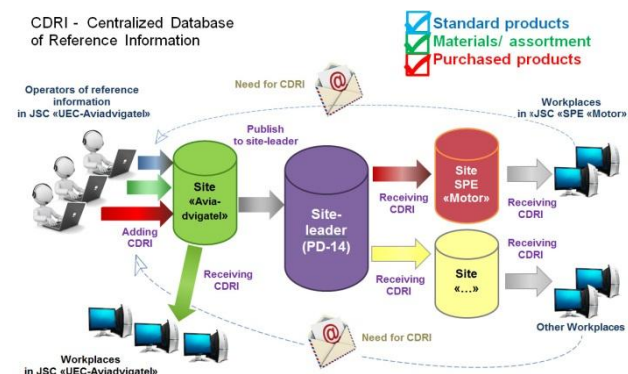


Figure 3. Managing RRI

The reference information is located in the Teamcenter database. If the co-executor had a need for SP, material or PP in the database, the co-executor formed an e-mail in which he described what he needed and sent it to the leading developer's address. Operators of reference information in JSC «UEC-Aviadvigatel» examined this problem and added the missing information to the site of their enterprise. Further, this information was published on the project site and became available to enterprises participating in the cooperation. Until 2013, each reference element was requested in this way, in 2013 the data exchange increased so much that it was not advisable to transfer separate objects. In 2013, all the material catalogs, SP's and PP's were transferred.

8. Agreement of electronic design documentation within the project

Management of design documentation in UIS is impossible without documents agreement in electronic form. Coordination of design documentation in Teamcenter is carried out using preconfigured business process templates.

Within the scope of PD-14 project, the design documentation negotiation process was implemented with the following features:

- built-in examination of design documentation;
- methods of routing changes;
- routes for returning the Design Documentation to the developer for rework.

At the end of the Design Documentation in the system, the document is assigned with the status «Approved» or «Technical studies». The «Approved» status is assigned to the Design Documentation after final approval.

If the Design Documentation is issued for particularly responsible details (assembly) or if their production is carried out at remote site, then the status «Technical studies» is assigned for this Design Documentation. After assigning the status «Technical studies», the

Design Documentation is sent out to the members of the cooperation for its further coordination. After this, the document is assigned with the status "Approved".

9. Conclusion

In conclusion, it should be noted that currently the Teamcenter PLM software within the PD-14 project continues to be actively implemented in the PJSC «UEC-UMPO», the following development steps are outlined:

- preparation for engineering infrastructure for other projects with JSC «UEC-Aviadvigatel» (expansion of computing facilities to support the work);
- integration with automated engineering process system;
- integration with the reference information management system "Semantic";
- transition to the newest up-to-date version of the NX design system;
- using of an intelligent component to solve various tasks within the project.

References

1. Goncharov P.S. NX for machine engineer/ Goncharov P.S. [Goncharov P.S and others]; etc. – M.: DMK Press, 2010. – 504 p.
2. Danilov Yu.V. Practical use of NX / Yu. V. Danikov, I.A. Artamonov. – M.: DMK Press, 2011. –332 p.
3. Nyrkov N., Osnach D. PLM-technology – new reserves in the era of competition // *SAPR I grafika* [CAD-systems and graphics]. – 2006. – no. 12. – pp. 75-77.
4. Siemens PLM Software solution to modernize production of Russian aircraft engines// *SAPR I grafika* [CAD-systems and graphics]. –2010. – no. 3. – pp. 54-57.
5. Siemens Ingenuity for life. Available at: <https://www.plm.automation.siemens.com/pub/case-studies/32082?resourceId=32082> (Accessed 27 May 2018).

JUSTIFICATION OF VOLUME OF OUTPUT DEPENDING ON PRODUCT ENVIRONMENTAL CRITERION BY NONLINEAR MATHEMATICAL PROGRAMMING

Rodionova L.N.,

Head of Finances, currency circulation and economic security department, Professor, Doctor of Economics
(Ufa State Aviation Technical University)

Ufa, Russia

e-mail: rodion@ufanet.ru

Abstract. The article provides the economical formulation, mathematical formalization and practical implementation of assortment planning of output product on the basis of financial profitability of each of the products; environmental profitability of each of the products; the period of receipt of funds for each sold product. Mathematical tools of formalization and implementation are nonlinear mathematical programming. The study was implemented on material of one of the key enterprises of the Republic of Bashkortostan.

KEY WORDS: ASSORTMENT OF OUTPUT, ENVIRONMENTAL FRIENDLINESS OF PRODUCT, NONLINEAR MATHEMATICAL PROGRAMMING.

1. Introduction

Among many practical challenges at enterprises there is always the problem of planning of goods and services. At the stage of promotion of goods or services into the market any it is difficult to propose any mathematical models allowing to justify a quantity of products and its assortment. In marketing it is proposed to use test-market sales, market research, participation in exhibitions and fairs in order to justify the volume of products.

In the period of strong sales there is a sufficient number of sales statistics that is why the construction of mathematical models of planning of goods assortment is already quite possible. Assortment planning is considered as periodically solved problem and the results of its solution determine the assortment strategy of the firm for a certain period.

The comparison of various strategies of assortment planning (maximum difference of prices between the price of wholesale supply and sale price of goods excluding penalty payments; maximum demand for previous sales, regardless of penalty payments; equal distribution of assortment; replenishment of previous sales; adaptive strategy focused on future demand taking into account environmental factor and amount of penalty payments and etc.) allows to propose ultimate environmental effect of assortment planning and percent decrease of expected profit.

2. Publication

The mathematical model applied in this study allows finding better solutions on planning of product assortment taking into account the identified above factors and limits in relation to the conditions of the existing market. The approach to this problem is based on the method of nonlinear mathematical programming implemented in the software program - Microsoft Excel.

Consider the enterprise producing and distributing i -type products ($i = 1, \dots, N$). The manufactured types of products are still in demand at the market that is why it is necessary to focus on promotion of already producing types of products before launching a new product line.

Besides, the following indicators should be defined:

j – investigated period ($j = 1, \dots, J$)

P_{ij} – sale price of a unit of i -type product in the j period;

C_{ij} – direct variable costs on production of a unit of i -type product in the j period;

G_{ij} – penalties for production of a unit of i -type product in the j period;

t_{ij} – period of return of funds for i -type product since the start of its production for the j period, measured in periods;

k_{ij} – quantity of i -type product manufactured and distributed in the j

period.

Taking into account the fact that in practice there are limits on volumes of output and sales of products of each type:

$$k_{ij \min} \leq k_{ij} \leq k_{ij \max}, \quad i = 1, \dots, N, \quad j = 1, \dots, J \quad (1)$$

$k_{ij \min}$ – minimum volume of output of i -type product in the j period (volume of output which the enterprise is to produce in order not to lose its market share),

$k_{ij \max}$ – maximum volume of output of i -type product in the j period (this indicator is defined by the firm's opportunities to sale the end product and also production capacity).

Maximum profit can be achieved by distribution of the available current assets by types of products in accordance with financial profitability but due to environmental profitability.

Thus, the profit per unit of product (without penalties) is

$$S_{Mij} = P_{ij} - C_{ij} \quad (2)$$

Financial profitability per unit of product is calculated as follows

$$M_{ij} = \frac{P_{ij} - C_{ij}}{C_{ij}} \quad (3)$$

Unit financial profitability of product unit is

$$R_{ij} = \frac{P_{ij} - C_{ij}}{C_{ij} t_{ij}} \quad (4)$$

The term of unit financial profitability is introduced due to the need to compare products with various values of profitability and period of return of funds for the manufactured products.

The term of environmental profitability, like financial profitability, is introduced. But in the latter case instead of profitability at cost, profitability of environmental payments, penalties for harmful products is calculated.

Unit environmental profitability per unit of products is

$$E_{ij} = \frac{P_{ij} - G_{ij}}{G_{ij} t_{ij}} \quad (5)$$

The effectiveness of environmental activities is defined by aggregate indicator of total profitability, combining various proportions of financial and environmental profitability.

$$RE_{ij} = (1-a) \cdot R_{ij} + a \cdot E_{ij} \quad (6)$$

where a – a coefficient of environmental friendliness of the production set depending on the selected production strategy, $0 \leq a \leq 1$.

The effective functioning of the enterprise requires to maximize accumulated financial profit obtained from the enterprise's principal activities during the period which is calculated as accumulated receipts in this period.

$$Z^{0j} = \sum_{j=j-g}^j \sum_{i=1}^N (P_{ij} - C_{ij} - G_{ij}) k_{ij} \rightarrow \max_{j: t_{i(j-g=g)}} \quad (7)$$

where g – variation limits t_{ij} , $g = \{1, 2, 3, 4\}$

While the problem is considered as dynamic (calculated for j periods), the profit obtained in the j period ($j = 1, \dots, J$) is reinvested. Minimum volume of output W_{min}^j for next period is initially provided:

$$W_{min}^j = \sum_{i=1}^N (C_{ij} + G_{ij}) * k_{ij min} \quad (8)$$

Costs for producing minimum volume of output in the coming period are deducted from the receipts at the beginning of the period Z^{0j} . Then, the funds remaining at the enterprise's disposal are

$$Z^j = Z^{0j} - W_{min}^j \quad (9)$$

The remaining Z^j is invested to produce more cost-effective types of products (first, the output with the highest unit aggregate profitability is manufactured, then, output with less profitability and etc.) for the current period

$$Z^j = \sum_{i=1}^N (C_{ij} + G_{ij}) k_{ij} \quad (10)$$

$$i: RE_{ij} \rightarrow \max$$

At this stage the objectives are identified.

At the beginning of the planning time interval $[0, T]$ divided into j periods ($j = 1, \dots, J$) the management of the enterprise provide financial resources in the amount of Z^{01} for production of i types products ($i = 1, \dots, N$).

At the beginning of each period it is necessary to distribute the available financial funds by i types of products ($i = 1, \dots, N$) due to the following criteria:

1) limits on output are defined by the formula 1.

Cost of minimum volume of output of i -type products in the j period is calculated by the formula 8;

2) the funds remaining at the enterprise's disposal at the beginning of the j period (besides the first period when they are set) are determined by the formula 2.9 and the accumulated receipts in the j period – by the formula 7;

3) distribution of the available funds is conducted in accordance with unit financial profitability of each product in each particular period and is calculated by the formula 10. Total profitability of product unit is determined by the formula 6, unit financial profitability of product unit – by the formula 4, unit environmental profitability of product unit is defined by the formulas 6, 4 and 5, consequently.

The result of funds distribution in the j period is the determination of the set of values $\{k_1^j, \dots, k_N^j\}$ which ensures that the enterprise will obtain maximum profit due to unit financial profitability and unit environmental profitability;

4) fixed costs which the enterprise pays at the beginning of each month (except for the beginning of the planning) are covered due to profit and are financial values;

5) the enterprise follows the strategy of reinvestment of profit. Such policy means that at the end of the period the net balance should be minimum (the latter is limited by that volumes of production are integer values)

$$Z^j - \sum_{i=1}^N (C_{ij} + G_{ij}) k_{ij} \rightarrow 0 \quad (11)$$

6) the indicators P_{ij} , C_{ij} , G_{ij} , t_{ij} , $k_{ij min}$, $k_{ij max}$ change from period to period and are provided by Marketing Department (for a month ahead with greater accuracy, for a longer period there are predictive estimates which are corrected as production and sales plans are implemented).

Now, it is necessary to solve the problem of selection of output depending on environmental friendliness of products.

OJSC "Ufa varnish and paint plant" manufacturing eight

types of products: enamels ПФ-115, ПФ-226, ЭП-751, XB-785, paints MA-15 and red lead MA-15, prime coating ГФ-021, varnish БТ-577 is the object of the investigation. The plant's objective is to distribute the funds optimally during the set time interval. The main criterion is to maximize profits taking into account the factor of environmental friendliness of the products in the long term. The problem is solved in stages using the following procedure.

1st stage. The table with indicators for each type of products in each period needed during analysis should be prepared (Table 1).

The indicator Z^{0j} – the enterprise's financial funds at the beginning of the period ($Z^{01} = 7000$ roubles) is calculated.

Table 1 – Calculation of profitability indicators

Item	Product sale price P_{ij} , RUB in thousands/tonness	Cost price, C_{ij} , RUB in thousands/tonne	Penalties payments, G_{ij} , RUB in thousands/tonne	Return period of funds t_{ij} , week	Unit financial profitability of product unit R_i , %	Unit environmental profitability of product unit E_i , %	Total profitability of product unit RE_i , %
Prime coating ГФ-021	26.67	20.34	1.53	1	31.12	16.39	23.75
Enamel XB-785	36.70	21.74	2.67	2	34.41	6.36	20.38
Enamel ЭП-751	34.00	22.17	2.52	2	26.68	6.25	16.47
Varnish БТ-577	13.54	8.46	0.65	3	20.02	6.65	13.33

Enamel ПФ-115	32.99	23.56	2.46	2	20.01	6.22	13.11
Enamel ПФ-226	31.24	22.46	1.25	3	13.03	8.02	10.53
Paint MA-15	28.44	22.14	1.91	3	9.49	4.62	7.05
Red lead MA-15	12.05	8.59	0.86	4	10.07	3.26	6.66

$$v = Z^j / (C_{ij} + G_{ij}) \quad (12)$$

2nd stage. The values of unit financial profitability R_i and unit environmental profitability E_i for each type of products are calculated. Products become ordered when total profitability RE_{ij} ($a = 0.5$) decreases. Priorities of types of products are identified while total profitability decreases.

3rd stage. The value of accumulated variable costs needed for production of minimum quantity of each type of products is calculated. All accumulated variable costs of each type of products corresponding to minimum volume of output are added up and the indicator W_{min}^j is obtained. Thus, the value of the funds available at the enterprise's disposal Z^j is calculated. The value of the index i is 1.

4th stage. The remaining funds for each type of products having higher priority in accordance to limits on volumes of output are distributed iteratively, for example, the product i . In this case the number of units k_{ij} which can be produced is calculated if there is an identified earlier

Then, the volume of output of i -type product is defined in accordance with the following equation. If

$$k_{ij \min} + v < k_{ij \max}, \quad (13)$$

the value k_{ij} has the value $k_{ij \min} + v$. Otherwise, the value k_{ij} has the value $k_{ij \max}$.

The value is calculated as follows $(C_{ij} + G_{ij}) * k_{ij}$.

The value Z' is defined by the formula

$$Z' = Z^j - (C_{ij} + G_{ij}) * (k_i - k_{i \min}) \quad (14)$$

If Z' is so low that this amount cannot be used for production of any product, or if it is zero, the 6th stage starts, otherwise, Z^j has the value Z' .

5th stage. The index i has the value $i + 1$. Then, it is necessary to return to the 4th stage.

6th stage. As a result of implementation of 4th and 5th stages all financial resources are distributed by types of products due to the criterion of total profitability (Table 2 shows the obtained data for the first period).

amount of financial resources Z^j .

The following formula is used for calculation

Table 2 – Calculation of indicators in the first period

1st period						Z^{01}	7 000.00
Priority	Item	RE_{it}	k_{it}	W_{min}^1	W^1	S^1	Z^1
1	Prime coating ГФ-021	23.75%	39	196.87	853.09	1 280.16	5 365.69
2	Enamel XB-785	20.38%	30	390.62	732.42	1 688.20	4 512.60
3	Enamel ЭП-751	16.47%	19	172.81	469.05	884.00	3 780.18
4	Varnish БТ-577	13.33%	32	72.85	291.39	541.60	3 311.13
5	Enamel ПФ-115	13.11%	54	624.38	1 404.86	2 573.22	3 019.74
6	Enamel ПФ-226	10.53%	9	71.12	213.35	374.88	1 614.87
7	Paint MA-15	7.05%	24	96.22	577.30	796.32	1 401.52
8	Red lead MA-15	6.66%	5	9.45	47.25	72.30	824.22
	Total:		212	1634.31	4588.71		

RE_{it} - total profitability by products, W_{min}^1 - costs on minimum volume of output, W^1 - costs above minimum volume of output (the amount of distributed financial resources due to the criterion RE_{it} and limits on volumes of production), S^1 - sales revenue by each type of products, Z^{01} - amount of financial resources at the beginning of the period, Z^1 - funds gradually distributed by types of products, the last line shows the demand balance after distribution in the current period, Z^{02} - amount of financial resources obtained in the second period is defined as an amount of revenue by types of products and demand balance as of end of previous period.

7th stage. If $j \leq J$, the index j has the value $j + 1$. It is necessary to return to the 1st stage, otherwise, the procedure is completed.

The problem has been solved under the following conditions – fixed costs of 1000 roubles per month. Three strategies have been applied:

- 1) distribution of funds and production not including environmental factor, profit maximization;
- 2) distribution of funds due to environmental factor but excluding financial profitability;
- 3) usage of 50% proportion of financial and environmental profitability.

Interpretation of results of solving the problem of output product depending on environmental friendliness of product.

While maintaining the strategy of profit maximization without other limits the enterprise receives higher profits in the long term as it invests the funds obtained from sales of products into production of types of products with the highest financial profitability.

If the enterprise follows the strategy of maximum environmental effect under the current limits, total sales revenue and profits decrease.

The strategy proves the need to manage output of products depending on the criterion of environmental friendliness when the enterprise sets an average profitability (total as a combination of financial and environmental in equal proportions of 50% and 50%) – as of end of investigated period the profit is 4 089 roubles. This amount is rather lower than the enterprise would have received if the first strategy was implemented. But in comparison with the second strategy it is much more (Fig. 1). It would be good to use the criterion of environmental friendliness as financial losses are slightly less and the production of greener products allows to define

that the enterprise follows the selected strategy of consistent improvement of the manufactured products and aims for maximum environmental friendliness under the current economical limits.

Conclusion. The advantages of this problem are that its solution allows automating the process of decision making in the sphere where there is difficulty of accurate forecasting of future prices for products, sales volume and etc. That is why projected values calculated, for example, for a quarter ahead, can be corrected every day

and the data for a month ahead are more likely to be calculated.

Thus, the solution of the problem allows managers of enterprises to use figures of profits, costs, work load of production facilities calculated for a week, a month ahead (for a quarter and a year ahead), to have the data on profitability of each type of products and its compliance with quality standards in order to distribute funds for advertising and promotion of products by types more effectively.

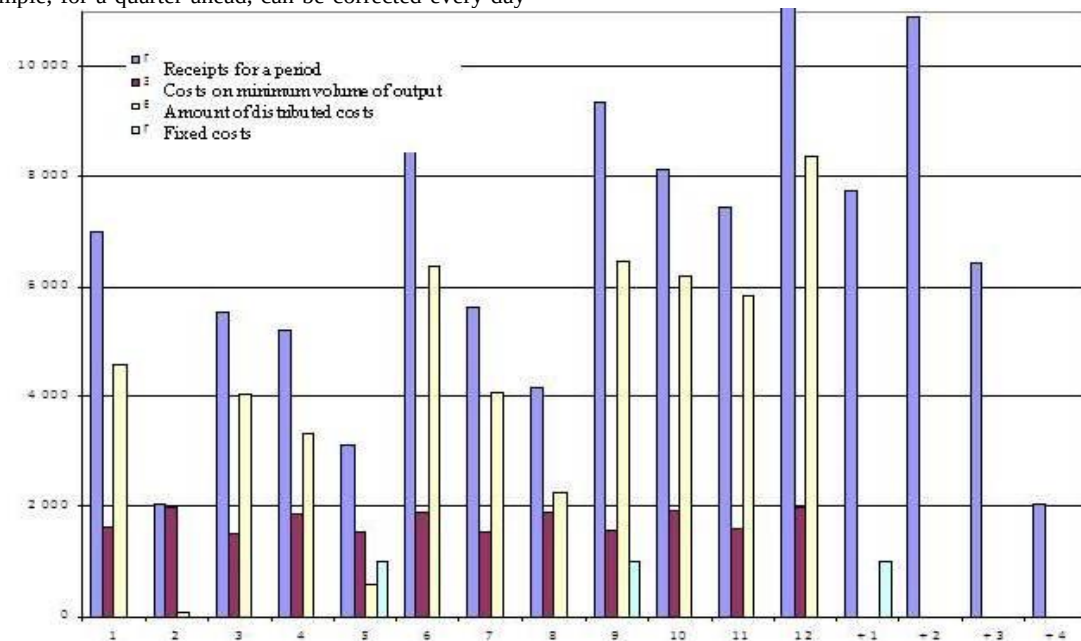


Figure 1 – Dynamics of indicators of solving the problem of output of products depending on environmental friendliness of the product while using strategy of compromise, rubles in thousands.

LITERATURE:

1. Bronshteyn A.M., Litvin V.A., Rusin I.I. Greening the economy: methods of regional governance. M.: Nauka, 1990. 120 p.
2. On environmental expertise: federal law of 23.11.1995 No. 174-Φ3 (edit. of 18.12.2006) // Russian Federation Code. 1995. No. 48. Art. 4556.
3. Gizzatullin R. Kh. Theoretical aspects of environmental function of modern Russian government / R.Kh. Gizzatullin // Legal policy and legal life. 2005. No. 1. Pp. 121-129.
4. Dorzhieva R. Ts. Environmental requirements: concepts and correlation with other requirements / R. Ts. Dorshievs // Environmental law. 2006. No. 5. Pp. 7-10.
5. Petrova T.V. Technical regulation as a part of legislative regulation of environment protection / T.V. Petrova // Environmental law. 2005. No. 1. Pp. 77-81.
6. Reymers N. F. Environmental management: glossary. – N.F. Reymers. M.: Mysl, 1990. 637 p.
7. Rodionova L.N., Mukhtarova E.G. Market orientation of methods of environmental management in the region. Economical vestnik UGNTU. Collection of scientific papers/ Edit. Rodionava L.N. and others. Ufa: pub. house of Fund of assistance in development of scientific investigations, 2001. 168p. Pp.105-110.

SCENARIO APPROACH FOR ANALYZING EXTREME SITUATIONS IN ENERGY FROM A CYBERSECURITY PERSPECTIVE

Aleksei Massel, Gaskova Daria

Information Technologies in the Energy Sector, Melentiev Energy Systems Institute of SB RAS, Irkutsk, Russia

e-mail: amassel@gmail.com, gaskovada@gmail.com

Abstract: The article describes the approach based on the Bayesian networks for construction of probabilistic scenarios to simulate extreme situations in the energy sector. The paper proves the feasibility of using Bayesian networks to simulate energy security threats caused by the implementation of cyber threats. The main components of the scenario and their interrelations are described. The main stages of modeling extreme situations in the energy sector are determined. The position of the scenario approach in the main stages of identifying critical objects in the case of energy is recognized.

1. Introduction

Critical infrastructures are regarded as large complex systems of strategic scale on the one part, and as cyber-physical systems on the other. In research of critical infrastructures, a large number of works are based on Simulation modeling, including Agent-based Modeling, Discrete-event Modeling, Dynamic Modeling, Structural Modeling, and Semantic Modeling [1], which involve Bayesian networks, Markov processes, Petri nets and Ontology-based Modeling [2].

Critical infrastructures are deemed to be part of the civil infrastructure, which is a collection of physical or virtual systems and means that are important to the state insofar as their failure or destruction can lead to disastrous consequences comparable to impact on military establishment [3]. Energy security protection is the national security part of the country [4], equally as energy sector is a one of the main critical infrastructure [5].

In the energy sector, the digitalization is primarily reflected in the Smart Grid, which is designed to expand the existing capabilities of generation, transmission and distribution systems, and to provide an infrastructure capable of meeting future needs for distributed generation, renewable energy sources and demand management [6]. The introduction of digital technologies increases the risks of security violation given the current move introducing new more complex solutions into the technological process of the energy facility, coupled with advent of new object business models, which are often undeveloped in terms of security. The above factors are contributing to receive incomplete and unreliable information for decision making, that are exacerbated by the dynamic nature of energy systems and their large physical infrastructure.

Under the "Digital Economy of the Russian Federation" programme approved by Government Order No. 1632-p of 28 July 2017 the main concepts of digital economy is highlighted. The authors consider ones as the whole new level of the Smart Grid in Russia.

The introduction of new information and communication technologies provokes the creation of new risks in the cyber environment of energy facilities.

There are many groups of international safety standards applicable to the energy industry such as CIP, NISTIR, etc. [7]. Since 2018, Federal Law No. 187-fz of 26 July 2017 "On the Security of the Critical Information Infrastructure in the Russian Federation" entered into force in Russia. However, a unified regulatory framework in the field of ensuring the cybersecurity of energy facilities in Russia has not been formed at the moment.

Risk management is traditionally applied for cybersecurity protection [8]. The transition from traditional methods and systems for providing strength, resource and reliability, to methods of risk assessment and management is the results of fundamental and applied research on the problems of technogenic security and risks [9].

The energy security threats are traditionally grouped as: (1) economic, (2) social-political, (3) technogenous, (4) natural and (5) managerial-legal [10]. This threat list was supplemented with the cybersecurity threats [11], their implementation possibly provoking serious emergency situations in the energy fraught with a drastic reduction of energy resources to be provided to consumers. The cyber threats are the least studied of these threats groups.

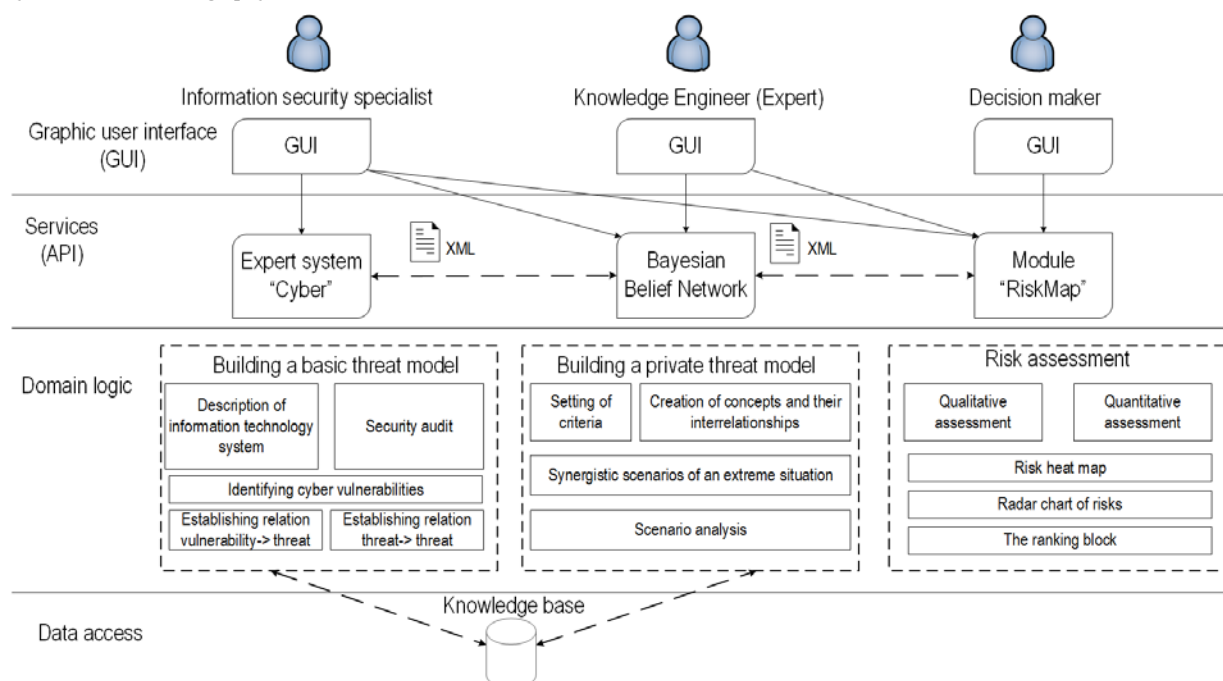


Fig. 1. Structure of intelligent system

2. Intelligent system for threat analysis and risk assessment of cybersecurity violations in energy facilities

Cyber threats could lead to realization of other energy security threats. Since information technologies are dynamic in their development the tendency of constant appearance of new vulnerabilities and ways of their use is undeniable.

The intelligent system is now under development [14] with a view to investigate possible circumstances of energy security violations caused by the implementation of cyber threats. Figure 1 shows the general structure of the intelligent system.

The development objective of the intelligent system for threat analysis and risk assessment of cybersecurity violations in energy facilities focuses on:

- auditing code for security vulnerabilities and the most probable cyber threats (trivial attacks);
- building of probability scenario of extreme situation resulting from cyber threats activities on site;
- extreme situation risk assessing;
- detecting critical facilities using the risk ranking of extreme situation.

The intelligent system includes a set of three main components: 1) the expert system for conducting security audits in an enterprise and identifying critical cyber vulnerability of the information technology system of the facility; 2) the block of Bayesian networks for the analysis of vulnerabilities, threats (cybernetic, threats to

energy security and external threats to the environment and objects external to the object under consideration) and the consequences from their implementation, which together lead to an extreme situation; 3) and the block of risks assessment of extreme situation.

The component of the Bayesian networks outlined in this paper is a tool for analyzing possible extreme situations in the energy sector and cybersecurity-wise. The next section of the article presents a scenario approach for analyzing threats leading to an extreme situation in the energy sector and justifying its applicability.

3. Opportunity analysis of using the scenario approach for research critical infrastructures

The original scenario approach arose from social systems studying. The specificity of scenario planning lies in consideration of simultaneously by several alternative option, for each of which internal and external factors are determined, criteria and indicators are established, and risks are estimated [15].

Usually, scenarios represent a probabilistic description of a possible or desirable the development variant of the phenomena and processes under consideration. The reaching of preferred states and situations in the form of course of action requires building development plans and making managerial decisions based on the scenarios analysis [15].

The development scenario of a complex system includes models, which contain significant factors that can be formalized with an acceptable degree of accuracy [16]. In case of extreme situations in the energy sector, the scenario is a pessimistic assessment of the main

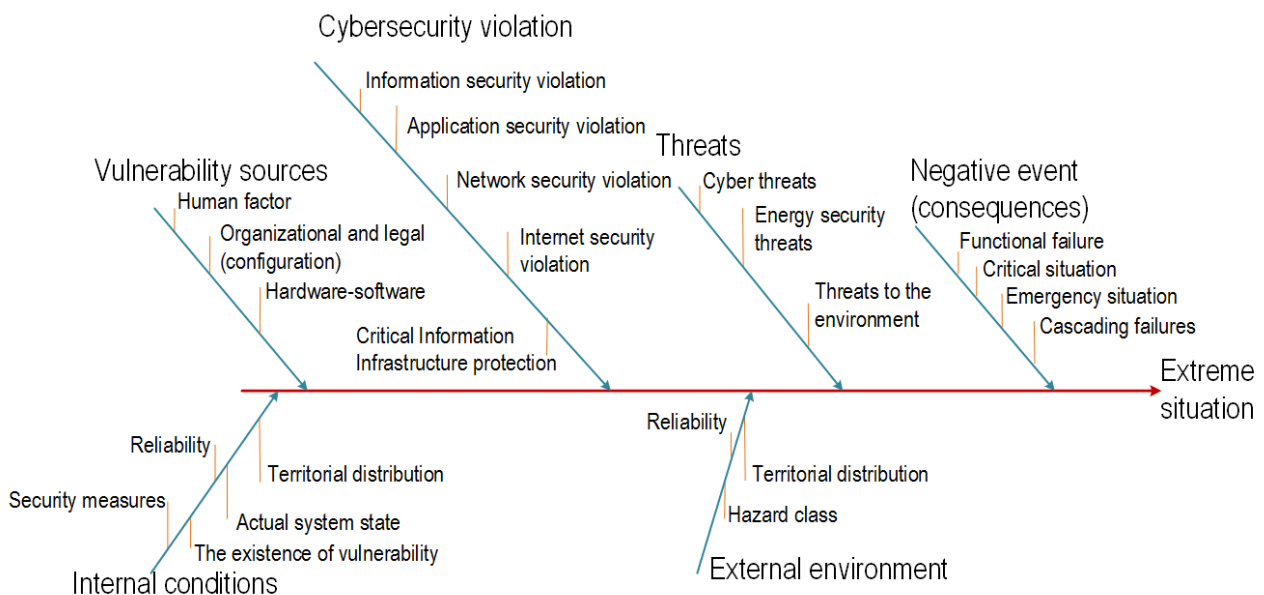


Fig. 2. Extreme situation caused by the implementation of the cyber threat at the energy facility (in Cause and Effect Diagram notation)

quantitative characteristics of energy facilities under implementing the threats chain: "cyber threats-energy security threats-threat to the external environment (and / or cascade accident)".

The advisability of using sign-oriented graphs to generate scenarios of complex systems is justified in [16]. One also describes the requirements for the mathematical device for modelling activities under the scenario approach:

- the possibility of constructing computationally effective analysis algorithms;
- low requirements (sensitivity) to the completeness and accuracy of the initial data;

- the possibility of a relatively simple software-system implementation.

The Bayesian network is proposed to construct such models. The Bayesian network is a graphical model of probabilistic and cause-effect relations between sets of variables, which is a directed acyclic graph whose vertices represent variables, and the edges show conditional dependencies between variables [17].

The Bayesian networks were formerly used to analyze threats to energy security [18]. The Bayesian network is a suitable tool for implementing a scenario approach in the framework of strategic planning. The application of Bayesian networks allows building models of extreme situations using either Bayesian probabilities in case of numerical values based on the expert's knowledge and

experience or frequency probabilities in the presence of statistical data, which often absent from free access.

Scenario analysis of risks using for investigation of complex systems is characterized by the absence of a "mass" nature of events in the scenario, and is aimed at studying unique situations and relationships [16].

4. Scenario approach

The proposed scenario approach promotes the research of extreme situations that might emerge on complex technological objects of power engineering. This approach also allows analyzing and generating scenarios of negative events leading to extreme situations and consequences from them. The results of scenario analyzing use to further justify decision-making and choose measures to ensure both the cybersecurity of the local facility and the energy security of the territory under consideration in the context of targeted (cyberattack) or accidental (cyber negligence) impacts on the information technology system of the energy facility.

The extreme situations are considered as both emergency situations and critical situations in studies on energy security. The definition of a particular situation is based on indicative analysis and assumes assessment of the system state or objects through the scale: "norm", "pre-crisis" - a critical situation, "crisis" - an emergency situation. With this in mind, the critical situations are referred to the situations when something threatens an uninterrupted functioning of the technical objects and the objects of life support and / or the life or health threats of individuals or social (professional) groups. These threats can be eliminated by adopting appropriate preventive and operational measures that will not allow the critical situation to develop into an emergency situation, in which operational and liquidation measures are needed [19].

Figure 2 presents the structure of the extreme situation and the main cause-effect concepts.

The scenario represents the conjunction of the chain of energy security threats that with a certain probability might flow from implementing cyber threats. This model also takes into account the conditions for the onset of negative events and the consequences that lead to damage.

Modeling the extreme situation begins with the identification of cyber vulnerabilities at the energy facility with the involvement of specialists (information security specialist). The objective of this stage is to create "vulnerability-threat" and "threat-threat" chains, that reflecting the trivial (predictable) attacks and behavior of the cybersecurity offender. An expert (a knowledge engineer in energy security) further continues to work with the chains by adding "threats-consequences" part. Subsequently, the concepts are concretized and scenarios are formed using the Bayesian network. Scenario includes the following groups of concepts:

- factors for negative consequences or conditions softening them;
- vulnerabilities of the information technology system of the energy facility;
- cyber threats posed by successful use of cyber vulnerabilities by the attacker;
- energy security threats caused by the implementation of cyber threats;
- threats to the environment caused by the implementation of cyber threats to the investigated facility;
- consequences using for further definition of damage and risk assessment of extreme situation.

Figure 3 illustrates the example of exploiting vulnerability scenario named "Remote code execution", which is considered on an industrial server.

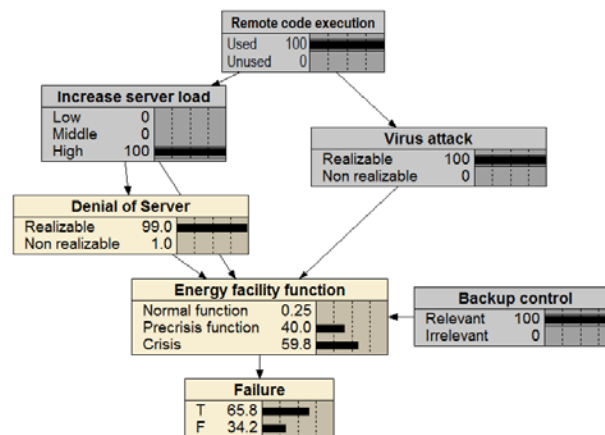


Fig. 3. "The remote code execution" vulnerability scenario using the Bayesian network in the Netica

This vulnerability can be exploited in several ways such as conducting a virus attack, embedding web shells, or conducting a Dos attack.

A virus attack can be aimed at either stealing information and disrupting the operation of an industrial server or using its resources for instance by including it in the Botnet network. This usually occurs for the purpose of sending SPAM (so-called black SEO), or to implement DDos attacks on other resources or Mining.

Scenario analysis will enable the expert to identify vulnerabilities that are most likely to be used by an attacker, a chain of threats that will have the greatest impact on the occurrence of negative consequences, and the consequences themselves and their probabilistic assessments. The above factors influence the risk of an extreme situation and the amount of damage from it. The information obtained in the scenario analysis is further suggested to manage risks and to develop protective measure.

5. Conclusions

The article describes some aspects of the energy sector as a critical infrastructure the security of which lies in the national security of the country. The scenario approach is proposed for predicting and analyzing possible failures, accidents and catastrophes as a result of the implementation of cyber threats. The scenario approach is aimed at the analysis of extreme situations in the energy sector as part of the development of intelligent system. The paper considers the main aspects of the extreme situation in the energy sector taking into account cybersecurity. In addition, an example of exploiting vulnerabilities is developed in the "Netica" software environment.

Acknowledgement

This work was partially supported by RFBR grants № 18-37-00271, №16-07-00474, №18-07-00714, №17-07- 01341. The authors are grateful to this organization.

References

1. L.V. Massel, A.G. Massel. "Semantic technologies based on the integration of Ontological, Cognitive and Event modelling", III International Science and Technology Conference "OSTIS-2013" Minsk, 2013, pp. 247-250 (in Russian).
2. M. Rybnicek, R. Poisel, M. Ruzicka and S. Tjoa, "A Generic Approach to Critical Infrastructure Modeling and Simulation", International Conference on Cyber Security "CyberSecurity" Alexandria, VA, USA, 2012, pp. 144-151.

3. L. Barannik, S. Klementev, "Organization of critical infrastructure security in the U.S.". No. 8. Foreign Military Review, 200, p. 3 (in Russian).
4. V.I. Rabchuk, S.M. Senderov, G.B. Slavin, "Energy Security in Russia: Problems and Solutions", Novosibirsk: ESI SB RAS, p. 197, 2011 (in Russian).
5. A. Kondratev "The current trends in research of Critical Infrastructure in foreign countries". No. 1. Foreign Military Review, 2012, pp 19-30 (in Russian).
6. S. Sridhar, A. Hanh, M. Govindarasu, "Cyber-physical system security for the electric power grid". Vol. 100. No. 1. Proc. IEEE, 2012, pp. 210-224.
7. L.V. Massel, A.G. Massel, "Cyber security of Russia's energy infrastructure as a component of national security," 6th International Conference on Liberalization and Modernization of Power Systems, 2015, pp. 66-72.
8. V.V. Mohor, A.M. Bogdanov, A.S. Kilevoj, "Information Technology. Methods of security. Cybersecurity manual (ISO/IES 27032:2012)," Three-K Kiev, 2013, p. 129, (in Russian).
9. N.A. Mahutov, N.V. Abrosimov, M.M. Gadenin, "Provision of safety - the priority in the sphere of fundamental and applied research". No. 3. "Economic and Social Changes: Facts, Trends, Forecast", 2013, pp. 39-61 (in Russian).
10. N.I. Pyatkova, V.I. Rabchuk, S.M. Senderov, M.B. Cheltsov, "Energy Security in Russia: Problems and Solutions", Novosibirsk: SB RAS, 2011, p. 211 (in Russian).
11. L.V. Massel, N.I. Voropai, S.M. Senderov, A.G. Massel, "Cyber Danger as one of the strategic threats to energy security," Cybersecurity issues, No. 4 (17), 2016, pp. 2-10 (in Russian).
12. Positive Research 2017. Collection of research on practical safety. 2017. Available at: <https://www.ptsecurity.com/upload/corporate/ru-ru/analytics/Positive-Research-2017-rus.pdf> (accessed 07.07.2018). (in Russian)
13. "WannaCry in industrial networks: work on mistakes", Report of Kaspersky Lab, 2017. Available at: <https://ics-cert.kaspersky.ru/reports/2017/06/08/wannacry-in-industrial-networks/> (accessed 07.07.2018). (in Russian).
14. A.G. Massel, D.A. Gaskova, "Application of risk-based approach to identify critical facilities in the energy sector with regard to cyber threats", Proceedings of the 19th International Workshop on Computer Science and Information Technologies. Germany, Baden-Baden. Publisher Ufa: USATU, Vol. 1, 2017, pp. 159-163.
15. K.A. Feofanov, "Scenario capabilities of modern forecasting and management", Vestnik MSTU "Stankin", No. 4, 2009, pp. 126-132 (in Russian).
16. V.V. Kulba, V.L. Schultz, A.B. Shelkov, "Information management. Part 2: Scenario approach", National Security / Nota Bene, No. 4, 2009, pp. 4-15 (in Russian).
17. D. Heckerman, "A Tutorial on Learning with Bayesian Networks", Technical Report MSR-TR-95-06, Microsoft Research, March, 1995, p. 57.
18. L.V. Massel, E.V. Pyatkova, "Application of Bayesian Networks for the intelligent support of Energy Security problem researches", Proceedings of Irkutsk State Technical University, No. 2, 2012, pp. 8-13 (in Russian).
19. L.V. Massel, A.G. Massel, "Technologies and tools for intelligent support of decision-making in extreme situations in energy", Computational Technologies, Vol. 18, No. S1, 2013, pp. 37-44.

METHODS AND TOOLS FOR PROCESSING SEMI-STRUCTURED DATA ON THE EXAMPLE OF ACCOUNTING FOR EDUCATION IN THE SELECTION OF PERSONNEL IN THE IT INDUSTRY

I N. I. Yusupova, O. N. Smetanina, M. M. Gayanova

Department of Computer Science and Robotics, Ufa State Aviation Technical University, Ufa, Russia

e-mail: yussupova@ugatu.ac.ru, smoljushka@mail.ru, maya.gayanova@gmail.com

Abstract: *This article is devoted to the analysis of documents containing requirements and underlying the development of educational programs, their structures, methods and tools that facilitate the automated processing of data contained in documents and the ability to make decisions using the results obtained. In the content plan, the task is to identify patterns between a given list of educational standards with possible types of professional activity in the IT field and key technologies using automation of data processing tools.*

1. Introduction

The volume of digital data by 2020 in the world will be 40,000 billion Gb [1]. And, as a rule, these data can be used to make decisions in various spheres of human activity.

Due to the interest in the results of the use of heterogeneous data and processing technologies, it became necessary to train specialists who possess technologies and methods for developing applications based on digital data. Therefore, questions concerning the training of personnel in the field of IT, as well as the development of tools for analyzing and processing data, identifying new knowledge and making decisions with their use, are undoubtedly relevant.

Education, specializing in the reproduction of skilled workers for all branches of material and non-material production, ideally should take into account the needs of the relevant industry.

The issue of training IT staff is directly related to the improvement of university educational programs, since this field of activity is very dynamic. In addition, it should be noted that one of the sections approved in July of 2017, the program "Digital Economy of the Russian Federation" is dedicated to training.

In this regard, the issues of the educational program analysis, taking into account the requirements of original documents and educational standards become an urgent task. With the questions of the study of educational programs, many works are connected [2,3,4,5].

This article is devoted to the analysis of documents containing requirements and underlying the development of educational programs, their structures, methods and tools that facilitate the automated processing of data contained in documents and the ability to make decisions using the results obtained. In the content plan, the task is to identify the patterns between a given list of educational standards with possible types of professional activity in the IT field and key technologies using automation of data processing tools.

2. State of the analysis of educational programs and IT sphere development

Professional staff in the field of IT is always in demand on the labor market. The education system in this sphere has been developing since the 60s of the 20th century, but due to the rapid development of information technology around the world, the number of specialists with university education is much less than the needs of the industry.

In 2017, the Government of the Russian Federation developed and approved a program to create conditions for the country's transition to a digital economy [6]. This Program defines goals and objectives within the framework of five basic directions for the development of the digital economy in the Russian Federation for the period up to 2024, among which are areas such as personnel and education, formation of research competencies, and technical reserves.

Many scientists are engaged in the training of personnel for the IT sphere, such as: N.A. Polyanskaya, A.E. Shamin "Actual problems of training personnel for the IT sector of the Nizhny Novgorod region", D.I. Monakhov "Model of training of personnel in the field of information technology", V.P. Gergel, R.G. Strongin

"Experience of Nizhny Novgorod University in training specialists in the field of supercomputer technologies", etc. [2].

The said authors are engaged in the analysis of educational programs for more than 20 years. Many issues related to the problems formulated in this article were previously considered in such articles as: "Models and methods of information processing in the analysis of educational programs" [3], "The state and prospects of the development of the information technology industry in the context of training" [4], "Decision-making in the selection of personnel for enterprises in the IT industry" [5].

The labor market for IT graduates is associated with three activities [7]: IT services, software development and equipment supply.

At each stage of the development of the digital society, new information technologies are emerging. At the moment, within the digital economy, six key technologies have been identified that will guide the development of the IT industry for many years to come. These are cloud computing, large data, Internet of things, digital production, mobility and information security.

Analysis of staff requirements for enterprises in the IT industry. The IT industry today is not only large enterprises, but now also small private business; for example, only in Belarus today there are more than 1000 companies in the IT market. The analysis showed that despite the different activities of these enterprises, three main activities of these enterprises can be distinguished: IT services, software development, supply and maintenance of equipment. IT services are projects for the integration of IT systems, development of applications for the customer, outsourcing, support and maintenance, consulting, training and education in the field of IT. The software development area includes infrastructure applications, user applications, business software and consumers, and so on. The scope of supply includes the supply of PCs, servers, peripherals, equipment, etc. [7].

In connection with the lack of IT professionals in companies, it is natural to seek and retain high-quality specialists. Research in this article should show how the specificity of basic education can be taken into account in the selection of personnel.

Analysis of Russian educational programs in the field of computer science and IT. The system of training bachelors, specialists and masters in the Russian Federation is based on educational standards [8]. In the Russian Federation, there are two enlarged groups of specialties and areas (EGSA) for training personnel for the IT industry: 02.00.00 Computer and Information Science and 09.00.00 Informatics and Computer Science.

Educational programs are developed taking into account educational standards, often, professional standards.

In terms of the characteristics of the document, the educational standard (FSSES), despite being presented as a text, has some structure / scheme. Consider the structure on the example of the educational standard developed for bachelors (Figure 1). Each element can include hierarchy associated with the basic concepts, and / or have attributes. In more detail, the elements used for the task are "defined".

GEF (level, direction of training: cipher, name))		
	- Application area	
	- Abbreviations used	
	- Characteristics of the direction of training	
	- Characteristics of professional activities of graduates who have mastered the bachelor's program	- Area of professional activity
		- Objects of professional activity
		- Types of professional activity
	- Requirements for the results of the bachelor's program	
	- Requirements for the structure of the bachelor's program	
	- Requirements for the conditions of the bachelor's program	
Figure 1 - Structure of the educational standard		

So for the "professional activities of graduates", characteristics / attributes are defined in the field of professional activity, objects of professional activity, types of professional activity. Below in the article are their "values" for a number of areas of training in the field of IT.

Such a document as a professional standard has the following structure (Figure 2).

Professional standard (name, date of approval, number, who approved)				
	- General information			
		Name of the type of professional activity		
		The main purpose of the type of professional activity		
		Occupation group	OKZ code	
			Name	
		Assignment to economic activities	OKVED code	
			Name	
	- Functional map of the type of professional activity	Generalized labor functions	Code	
			Name	
			Qualification level	
		Labor functions	Code	
			Name	
			Qualification level	
	- Characterization of generalized labor functions	Generalized labor function	Code	
			Name	
			Qualification level	
			Origin of labor function	
			Possible job titles	
			Requirements for education and training	
			Requirements for practical work experience	
			Special conditions for admission to work	
			additional characteristics	
			Work function	Name
				Code
				Level (sublevel) of qualification
				Origin of labor function
				Labor actions
				Required skills
				Required knowledge
				Other Features
	- Information on professional standards organizations			

Figure 2 – Structural scheme of the professional standard

Italicized elements of the document, which "have" several meanings. Thus, in the professional standard "programmer", the following generalized labor functions are indicated: development and debugging of the program code (labor functions: formalization and algorithmizing of tasks, writing code using programming languages, defining and manipulating data, creating code in accordance with established requirements, work with the version control system, checking and debugging the program code); validation and refactoring of the software code (development of procedures for verifying the health and measuring software characteristics, developing test data sets, checking the operation of the software, refactoring and optimizing the program code, correcting defects recorded in the database of defects); integration of program modules and components and validation of software product releases (development of procedures for the integration of software modules, integration of software modules and components and verification of software product releases), etc.

The considered schemes will allow selecting a tool for automated processing of data contained in documents.

Returning to the educational standard, we note that the field of professional activity of graduates who have mastered the bachelor's program includes the development, implementation and operation of software for various purposes, etc. (Table 1).

Features of training specialists 02.00.00 EGSA is their focus on mathematical aspects, and graduates 09.00.00 EGSA - applied aspects of tasks to be solved, such as software development,

automated information processing and management systems; research, development, implementation and maintenance of information technologies and systems; system analysis of the application domain, etc.

When hiring a decision-maker in a company, it will be necessary to analyze the capabilities of graduates of larger groups of specialties and areas (EGSA) [8]: 02.00.00 "Computer and Information Science", 09.00.00 "Computer Science and Computer Science" (Table 1, * - repeated for bachelor's and master's training areas and objects of professional activity).

When developing the basic educational programs, the types of professional activity provided for by the educational standard in each direction are taken into account. Figure 3 is a diagram showing the types of professional activity that are selected for the preparation of bachelors in the direction of 09.03.01 Computer science and computer technology by a number of Russian universities. Considering that only a part of the universities chose all six types of professional activity recommended by the educational standard, it can be concluded that the graduates of higher educational institutions are more narrowly trained.

Another aspect that is taken into account in the development of the educational program is the so-called profile (Fig. 4).

Table 1. The areas and objects of professional activity of graduates of training directions 02.00.00, 09.00.00 (bachelor's and master's)

Level of training	Area of professional activity	Objects of professional activity
Mathematical support and administration of information systems		
02.03.03	Development, implementation and operation of software;	<i>Mathematical and algorithmic models, programs, software systems and complexes, methods for their design and implementation, methods of production, maintenance, operation and administration in various fields, including interdisciplinary, simulation models of complex control processes, software, administration of computing, information processes *</i>
02.04.03	Development, implementation and operation of software for various purposes	
Informatics and computer facilities		
09.03.01	Software for computer systems and networks, automated information processing and control systems	<i>Computers, systems, complexes and networks, automated information processing and control systems, computer-aided design and information support systems for the life cycle of industrial products, software for computer facilities and automated systems, mathematical, information, technical, linguistic, ergonomic, organizational and legal support for listed systems *</i>
09.04.01	Theoretical and experimental research of scientific and technical problems and solving problems in the development of hardware and software for computer systems and networks, automated (including distributed) information processing and control systems, as well as computer-aided design and information support systems for products	Computer systems, complexes, systems and networks; on means of computer facilities and automated systems (programs, software complexes and systems)
Information systems and technologies		
09.03.02	<i>Research, development, implementation and maintenance of information technology and systems *</i>	<i>Information processes, technologies, systems and networks, their instrumental (software, technical, organizational) support, methods and methods of designing, debugging, production and operation of information technologies and systems in the areas of: engineering, instrumentation, technology, education, medicine, jurisprudence, business, etc., as well as enterprises of various profiles and all kinds of activities in the economy of the information society *</i>

Level of training	Area of professional activity	Objects of professional activity
09.04.02		
Applied Informatics		
09.03.03	System analysis of the applied field; formalization of the solution of applied problems and IP processes, development of projects for automation and informatization of applied processes and the creation of IP in applied fields	Applied and information processes, information technologies, information systems *
09.04.03	Study of patterns information and development of the information society, properties of information and characteristics of information processes; research and development of effective methods for implementing information processes and building IP in applied areas based on the use of modern ICT; organization and implementation of system analysis and reengineering of applied and information processes, formulation and solution of applied problems; modeling of applied and information processes, development of requirements for the creation and development of IP and its components; organization and implementation of works on feasibility study of design solutions, development of projects for automation and informatization of applied processes and creation of IP in applied fields; project management for informatization of enterprises and organizations, decision-making on the implementation of these projects, organization and management of the implementation of IP projects in the applied field; automation quality management of applied problems, IP creation processes; organization and management of IP operation; Training and consulting on automation and informatization of applied processes and the introduction of IP in the application areas	
Software Engineering		
09.03.04	Industrial production of software for information and computing systems for various purposes *	A software project (a software product development project), a software product (the created software), software product lifecycle processes, methods and tools for developing a software product, personnel participating in life-cycle processes
09.04.04		Methods and algorithms for data processing in information and computing systems; parallel, high-performance and distributed information and computing systems; processes of industrial software testing; programming languages and their translators; network protocols and network services; OS

The analysis shows that in general, the objects of activity of the graduates of the second group are mathematical and algorithmic models, simulation models of complex management processes, etc., and the graduates of group 9: computers, complexes, systems and networks, information processes, technologies, methods and algorithms for data processing, etc. The authors believe that on the basis of this analysis, it is possible to establish a link between educational programs and core activities in the IT field and give practical recommendations to decision-makers in the field of recruitment at IT enterprises.

Different types of training can be associated with different sectors in the IT field, with activities in each sector. For example: the sector associated with the provision of services as one of the activities considers training and education in the field of IT. Among the types of professional activity in the educational standards of the training

areas under consideration are scientific-pedagogical or pedagogical ones. A custom application development can be associated with such types of professional activities as design, design and technology, etc. These are simple examples of correspondences. More complex relationships, considering the areas and objects of professional actions, training profiles, as well as tasks to be solved, will require the automation of the analysis process.

In Russia, more than 350 universities conduct training in the direction of 09.03.01. "Computer Science and Computer Engineering", more than 90 universities in the direction of 09.03.04. "Software Engineering", more than 50 universities in the direction of 02.03.03. "Mathematical support and administration of information systems". Such a variety of educational programs leads to the fact that it is not easy to perform the "manual" analysis. Therefore, the authors proposed to develop tools for analysis.

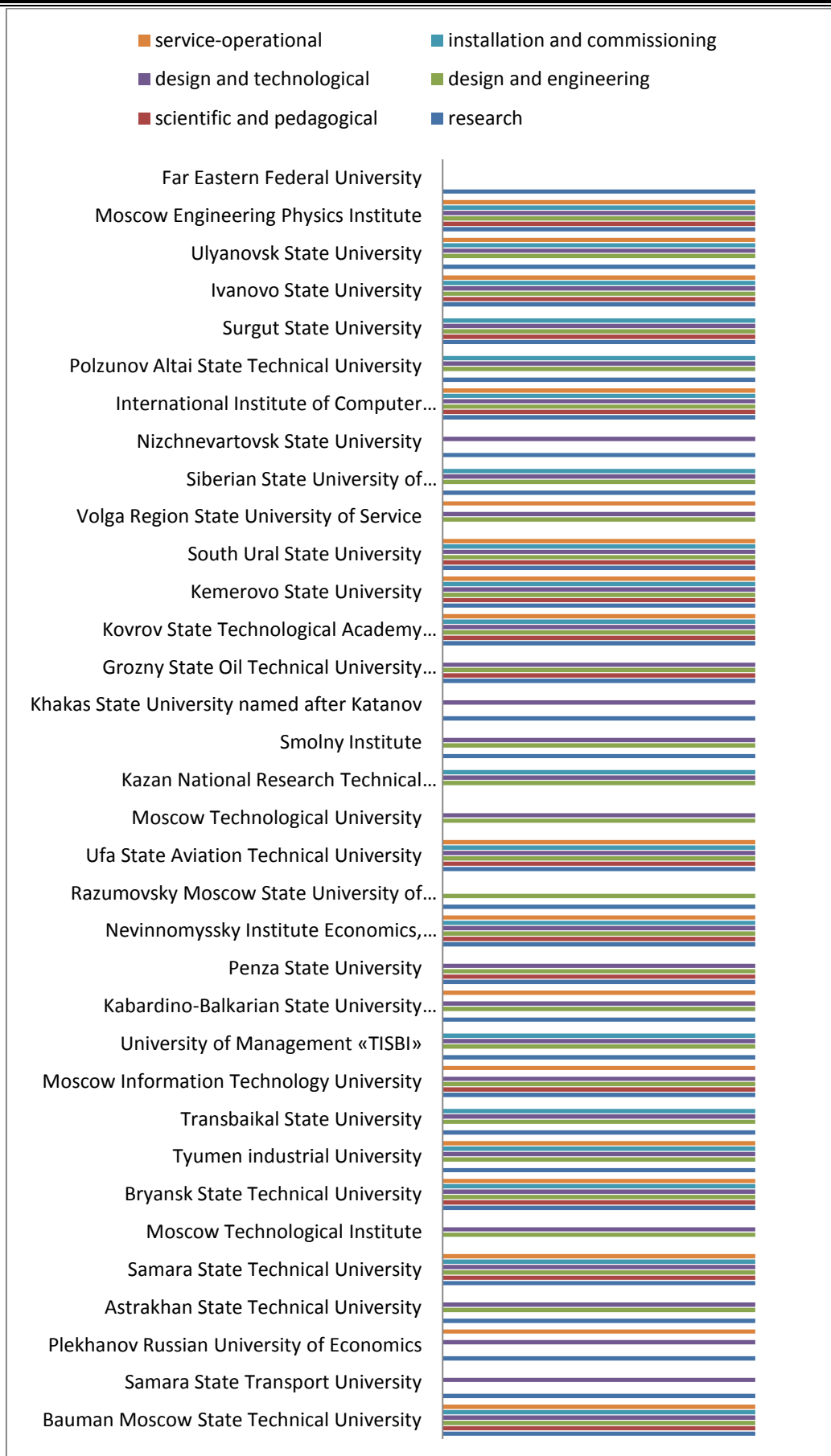


Figure 3. Types of professional activities identified for educational programs in the area of training 09.03.01 (Computer science and computer technology) in a number of Russian universities

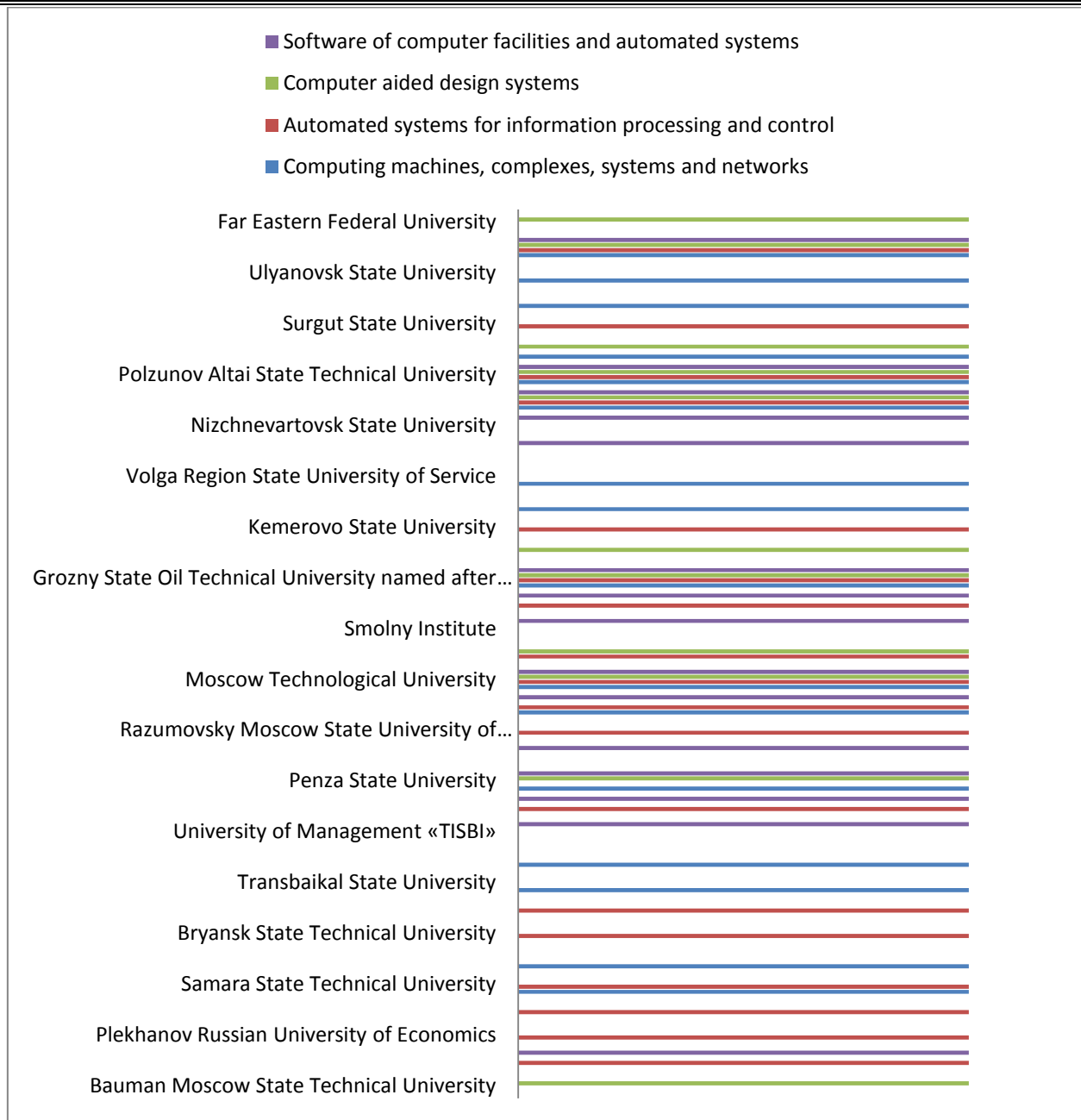


Fig. 4. Profiles of educational programs for training directions 09.03.01 (Computer science and computer technology) at a number of Russian universities

3. Formulation of the problem

The task of the authors within the framework of this article is to investigate and establish the patterns between the given list of educational standards and the educational programs developed within them with possible activities in the IT field and key technologies using automation of data processing tools. In terms of set theory, the establishment of regularities can be represented as the identification of functions $f^1: S_o \times S_{op} \rightarrow SecIT$ and $f^2: S_o \times S_{op} \rightarrow KIT_{IT}$, where S_o – educational standard, S_{op} – educational program; $SecIT$ – IT sector (services, software development, equipment supply), KIT_{IT} – key technologies in IT.

The analysis will allow to correct the list of directions considering modern requirements, improve educational standards and give recommendations to the IT industry on the accounting of education in the selection of personnel.

As it was shown earlier, all the documents submitted are inferior structured data. When working with such data, as a rule, there are problems of search in large data sets and their immediate processing, due to the absence of rigid structure, tabular

representation and presentation of metadata. One way to present data for processing them is an ontology that can be defined as a set of the domain description schema and the rules for assigning data to this domain. The ontology must contain concept-entities, entity attributes and loaded connections.

The author [9] notes the ways of processing such data: ontology allocation - a description of the schema for a domain characterized by a certain logical structure (the content of the selected ontology is filled using the semantic analysis of the text, a set of written linguistic rules); search for mentions, categorization and retrieval of facts (search by keywords, selection of facts related to objects of search); emotional coloring, evaluation of interest, attitude (semantic analysis based on linguistic rules applied after the ontology is singled out); the identification of patterns - the dynamics and course of the change in the relationship, the allocation of general borrowing (based on the concept-entities identified in the framework of ontology, their attributes and connections).

The developed ontology will allow to solve the following problem more effectively.

4. Models and methods for data analysis

The authors have repeatedly dealt with the problems of data analysis in the educational sphere using artificial intelligence technologies, including ontological models [10]. In particular, it included the analysis of educational programs for improving the quality of training, for the formation of educational routes, etc.

The analysis of previously developed ontological models in the field of education made it possible to single out both general classes regardless of the tasks to be solved (the main educational program, discipline, educational standard, competence, etc.) and classes directly related to the purpose of the application, such as provision (hardware, software), literary sources, educational process, etc. In addition, we can talk about the use of basic relations in ontology (for example, subsumption, "part-whole", situational (when in one set combined knowledge relevant to a typical situation)), and so special. The improved ontology in the part of the educational system in combination with the ontology being developed in the IT sphere can be used for the task posed in the article. The formal representation of an ontology using the set-theoretic description can look as follows: $KB = \langle C, In, Rel, D, Rules \rangle$, where C is the set of classes (concepts) of subject ontology ("educational standard", "enlarged group of training directions", "basic educational program", "IT sector", etc.); In - a set of instances of classes (for example, instances of the class "enlarged group of training directions" can be "09.00.00 Computer Science and Computer Science", "02.00.00 Computer and Information Sciences"); Rel - a set of ontology relations, D - data (property values); $Rules$ - developed rules. Let us consider relations and rules in more detail.

Ontologies include different types of connections: taxonomic ("is a species", "part / whole"), composite ("is part", "includes"), topological ("object is its property"), problem-specific ("the object participates in process", "the object manages the process"), etc. Separately, one can note the relations set by descriptive logic: "is", "has". Each link is presented in a specific ontology by a variety of options. Some examples of relationships in the proposed ontology can serve as the following: class-subclass (the class "enlarged group of training areas" may contain a subclass of "training direction"); class-instance (copies of the class "training direction" can be "09.03.04 Software Engineering" (bachelor's degree), "09.04.04 Software Engineering" (Master's program)); the properties of classes / individuals (the class "professional activity" can "possess" the properties "view", "area", "object", "view" "defines activity" "research" or "design"); problem-specific communication ("student is studying" or "university is preparing").

Let us consider in more detail the relationship of the hierarchy on the example of the concept of "characteristics of the professional activities of the graduate "PAG" in the framework of the problem being solved (Figure 5), and an apparatus that can be used for automatic language processing.

To model the semantics of free word combinations with subsumption / hierarchical relationships, a mathematical apparatus for describing deterministic, discrete and finite information objects - the algebra of finite predicates [11] can be used.

For the example in question, you can enter variables and areas of their allowed values. Let $x_1 \in \{1\}$, $x_2 \in \{2, 3\}$, $x_3 \in \{4, 5, 6, 7, 8\}$, $x_4 \in \{9, 10, 11, \dots, 31\}$. The equation describing the tree will look like this.

$$\begin{aligned} & (x_1^1 \sim x_2^2 \vee x_2^3) (x_2^2 \sim x_3^4 \vee x_3^5 \vee x_3^6) (x_2^3 \sim x_3^7 \vee x_3^8) \\ & (x_3^4 \sim x_4^9 \vee x_4^{10} \vee x_4^{11} \vee x_4^{12} \vee x_4^{13} \vee x_4^{14}) \\ & (x_3^5 \sim x_4^{15} \vee x_4^{16} \vee x_4^{17} \vee x_4^{18} \vee x_4^{19}) (x_3^6 \sim x_4^{20} \vee x_4^{21}) \\ & (x_3^7 \sim x_4^{22} \vee x_4^{23} \vee x_4^{24} \vee x_4^{25} \vee x_4^{26}) (x_3^8 \sim x_4^{27} \vee x_4^{28} \vee x_4^{29} \vee x_4^{30} \vee x_4^{31}) = 1. \end{aligned}$$

Having determined the maximum possible number of variables at each level, we can create a universal equation for a given fragment of the ontology. The solution of this type of equation corresponds to the nodes of the genus-species tree.

We describe an example in terms of the algebra of finite predicates. Syntax: the alphabet of symbols $A = \{a_1, a_2, \dots, a_k\}$, the alphabet of variables $B = \{b_1, b_2, \dots, b_n\}$, the predicate $P(x_1, x_2, \dots, x_n)$. The formulas are based on the symbols of the alphabet A , variables, signs (basic for the algebra of predicates), logical constants 0 and 1. The predicate specified on D_n (full-text documents) is a function $P(x_1, x_2, \dots, x_n)$ mapping the set D_n into the set $K = \{0, 1\}$. The variables x_j are subject variables, their meanings are objects. Since we are talking about a bounded n , then the predicate is finite. The basic predicates for the algebra of predicates are predicates of the form:

$$x_j^{a_i} = \begin{cases} 1, & x_j = a_i \quad (1 \leq j \leq n), \\ 0, & x_j \neq a_i \quad (1 \leq j \leq n), \end{cases}$$

where $i = \{1, 2, \dots, n\}$; a_i - is an element of D_n . A predicate $x_j^{a_i}$ is called the predicate of recognition of the object a_i with respect to the variable x_j . The hierarchy level j can be represented as follows:

$$x_j^1 \sim (x_{j+1}^{a_1} \vee x_{j+1}^{a_2} \vee \dots \vee x_{j+1}^{a_l}) = 1, \text{ where } a_i \text{ are the elements of the set } A.$$

Simulation of semantic relations is used in the task of automatic language processing.

Returning to a formal description of the ontology, let us consider an example of a rule constructed in the form of Horn clauses:

$$\text{Rule } i: C_1(?x_1) \wedge C_3(?x_2) \wedge P_1(?x_1, ?x_2) \wedge C_2(?x_1, ?z) \rightarrow C_4(?z, ?x_2).$$

Here: C_i - classes (university, student, training direction, IT sector, respectively); P_i - predicates (learns); x_i - instances or variables, z - variables or values.

Thus, a recommendation for a decision can be formulated. In general, the ontology being worked on is a solution to the problems posed, namely, the definition of the functions $f^1: S_o \times S_{op} \rightarrow Sec_{IT}$ and $f^2: S_o \times S_{op} \rightarrow KIT_{IT}$ using rules and additional data.

5. Personnel selection taking into account the peculiarities of educational training programs

The analysis of the content of educational programs for bachelors and major activities in the field of IT allowed the authors to establish some types of communication [3], which are presented in Figures 6 and 7.

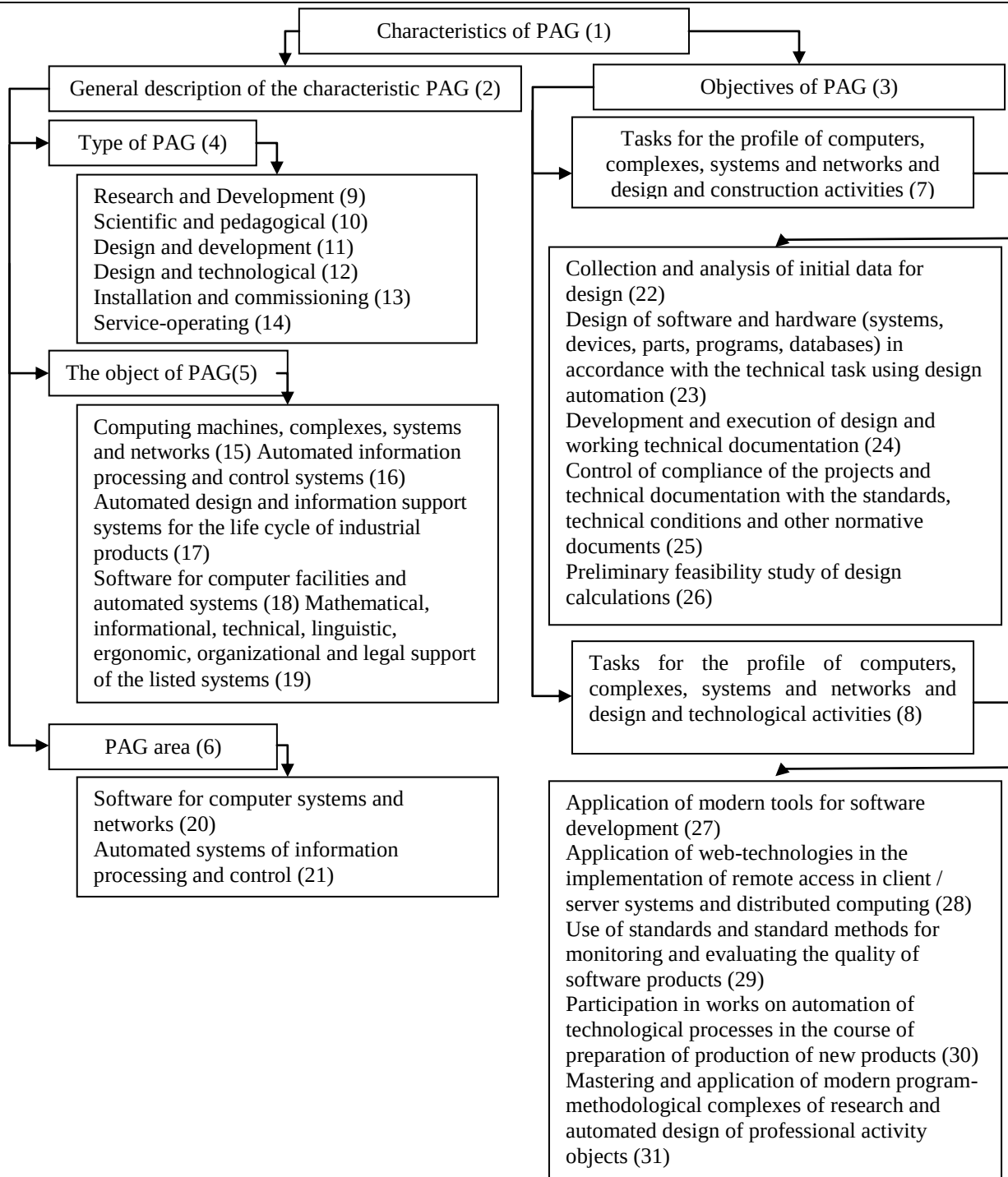


Fig. 5. The term "characteristic of the professional activity of a graduate"

To improve the educational process, it is necessary to understand the extent to which the educational process ensures the formation of competences in the main types of IT activities and key technologies.

The authors are convinced that this information is useful for persons engaged in the selection of personnel in IT enterprises; there is a link between the activity of the enterprise and the direction of training bachelors. In many cases, representatives of different areas

(02 and 09 EGSA) can be selected to form an effective project team. For example, a graduate of the 02.03.03 direction "Mathematical support and administration of information systems" will be able to build a mathematical model, a release of the direction 09.03.03 "Applied Information" – to compile system requirements, graduate of the direction 09.03.04 "Software Engineering" – write the software

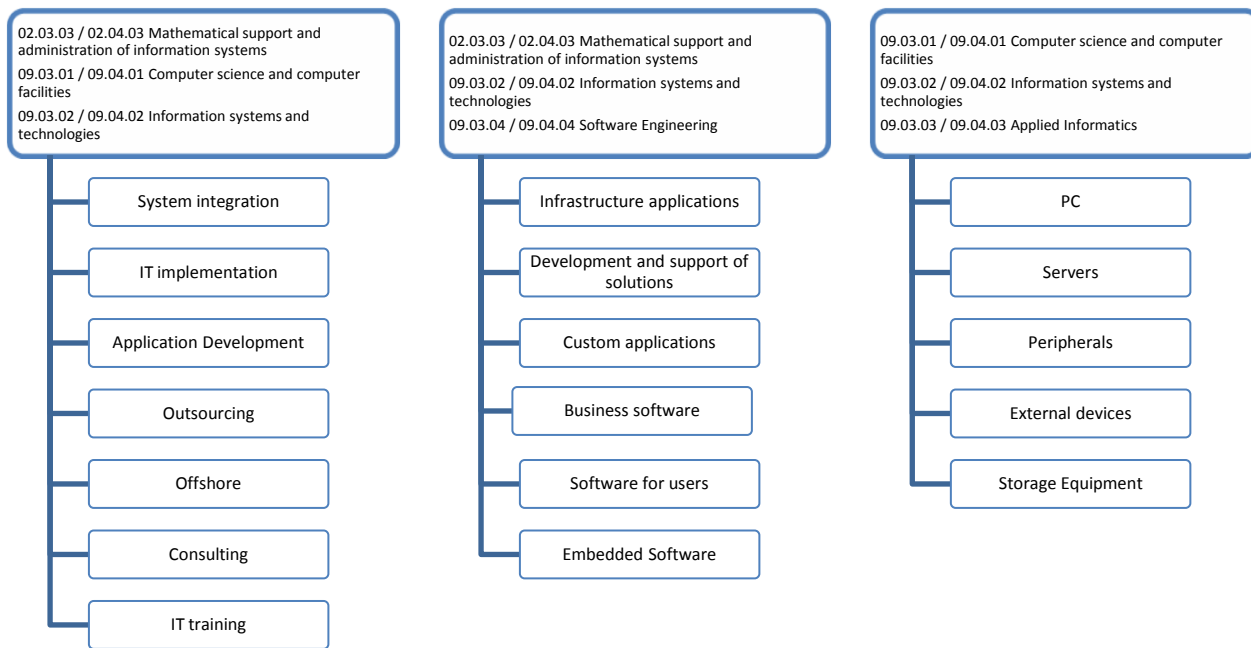


Fig. 6. Linking the areas of training with IT activities (Bachelor and Master)

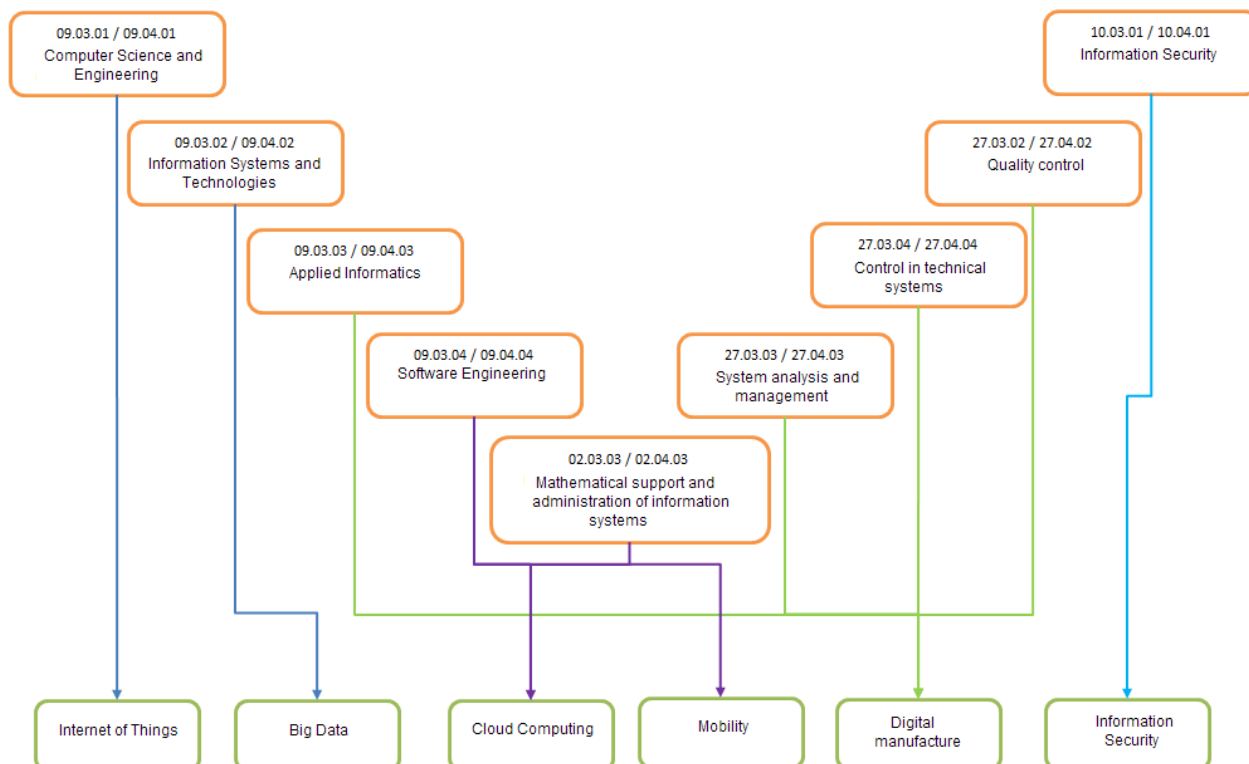


Fig. 7. Linking the areas of training with key technologies (Bachelor and Master)

6. Conclusion

The analysis of the documents used in the development of educational programs showed that they represent poorly structured data. To effectively use this kind of data, an ontology can be built. To simulate the semantics of free word combinations with subsumption / hierarchical relations, a mathematical device can be used to describe deterministic, discrete and finite information objects - the algebra of finite predicates.

The rules developed in ontology in the form of Horn clauses will make it possible to formulate recommendations for decision-making.

Analysis of existing educational programs and core activities in the IT field allows us to identify regular links. The conducted analysis can be useful to HR managers, developers of educational programs to take into account the requirements of the IT industry, etc.

References

1. Saraev, V. When the data became large / V. Saraev // Expert. Series Science and Technology. IT, 2013, №19 (850) <http://expert.ru/expert/2013/19/kogda-dannyye-stali-bolshimi/>
2. Scientific electronic library "Ciberleninka" [Electronic resource] // <https://cyberleninka.ru/> (available at 03.08.2018);
3. Yusupova N.I., Smetanina O.N., Gayanova M.M. Models and methods of information processing in the analysis of educational programs (on the example of training informatics) / Journal of Information Technologies №6. Tom 23. 2017
4. Yusupova N.I., Smetanina O.N., Gayanova M.M. On the state and prospects of the development of the information technology industry in the context of training – Materials of the VI International Scientific Conference "Information Technologies and Systems", Bannoe, Russia, Chelyabinsk: Chelyab. state. university, 2017. pp. 348-353.
5. Yusupova N.I., Smetanina O.N., Gayanova M.M. Decision-making in the selection of personnel for the IT industry, taking into account the peculiarities of the bachelor's degree – Materials of the V All-Russian Conference ITIDS-2017, Ufa, USATU, 2018. T.2., pp. 225-228
6. Russian Government [Electronic resource] // <http://static.government.ru/media/files/9gFM4FHj4PsB79I5v7yLVuPgu4bvR7M0.pdf> (available at 07.08.2018);
7. Association of Computer and Information Technology Enterprises [Electronic resource] // <http://www.apkit.ru/> (available at: 02.08.2018);
8. Portal of Federal State Educational Standards [Electronic resource] // <http://fgosvo.ru/> (available at: 03.08.2018 г.)
9. Grishkovsky A. Integrated processing of unstructured data / Grishkovsky A. // Open systems. DBMS. 2013. №06. <https://www.osp.ru/os/2013/06/13036849>
10. Methodical aspects of artificial intelligence. Guzairov M.B., Yusupova N.I., Kudryashev A.F., Yelkhova O.I., Popov D.V., Rizvanov D.A., Yennikeyeva K.R., Dolomatov M.Yu., Yartsev R.A., Smetanina O.N., Verkhuturov M.A., Verkhuturova G.N., Makeyev G.A., Gayanova M.M., Boyko M.V., Bogdanova D.R., Klimova A.V., Galyamov A.F., Orekhov YU.V., Orekhov E.U. and others. Monograph. Moscow: Mashinostroyeniye, 2014.
11. Modeling semantic ontology relationships / N.V. Babkova, O.V. Kanishcheva, Z.A. Kochueva // BionicaIntellecta: Sci. Mag. 2013. № 1(80). P. 1-4.

INTELLIGENT DECISION-MAKING SUPPORT IN ENERGY AND ECOLOGY IN VIEW OF THE QUALITY OF LIFE

L.V. Massel

Laboratory of Information Systems in energetics, Melentiev Energy Systems Institute of SB RAS, Irkutsk, Russia
e-mail: massel@isem.irk.ru

Abstract: *The article examines the results of an international project carried out with the support of the EAPI Fund together with researchers from Belarus and Armenia. The project aims to develop methods and technologies for assessing the impact of energy on the geoecology of the region. The article is devoted to the development of tools for intelligent support of decision-making in this field. (style CSIT- abstract).*

1. Introduction

Studies of the impact assessment of energy on the geoecology of the region [1, 2] are conducted within the framework of an international project supported by the EASR-RFFI funds, in cooperation with the teams of scientists of Belarus and Armenia. The article examines the main provisions and results of the project carried out by the Russian side.

The fundamental scientific problem addressed by the Project is the development of methods and geoinformation technologies for assessing the impact of energy on the geoecology of the region [3]. The object of research from the Russian side is the Baikal natural territory, comparable in size to Belarus and Armenia.

The article is devoted mainly to the development and integration of modern information technologies for intelligent decision-making support within the framework of the problem. To implement the project, a Web-based information system (WIS) is being developed that integrates mathematical and semantic methods [4, 5] and tools for assessing the impact of energy on the geoecology of the region, a database, a knowledge base and a geographic information system. Individual WIS components will be implemented as agents-services [6]. The ontology of the energy impact on the environment, ontology of pollutants and WIS architecture is presented in the article.

The aim of the project is to develop methods and technologies for assessing the impact of energy on the geoecology of the region. Geoecology is understood as an interdisciplinary scientific field that unites research into the composition, structure, properties, processes, physical and geochemical fields of Earth's geospheres as a habitat for humans and other organisms. The main task of geoecology is to study the changes in the life-supporting resources of the geosphere shells under the influence of natural and anthropogenic factors, their protection, rational use and control in order to preserve productive and natural environment for present and future generations [1, 2].

The purpose of this project fits into the main task of geoecology. The relevance of the project is determined, on the one hand, by the importance of the problem of assessing the impact of energy on the geoecology of the region, and on the other, its insufficient research and the need to attract modern geoinformation and intelligent technologies for its solution.

In the course of the project, the formulation of the main specific tasks to be solved in the project was clarified and expanded: 1) to conduct an analysis of existing methods for estimating of pollutant emissions from energy facilities and existing models for the spread of pollution caused by emissions from energy objects (taking into account the wind rose, transfer, etc.); 2) to make a choice and justify the methods recommended for use in the project, their modification and adaptation, the development, if necessary, of the original methods; 3) to identify critical objects that affect the life support and natural environment of the region (in the energy, water supply, etc. sectors), the connection of critical facilities with the quality of life of the population [7]; 4) to analyze the approaches to the construction of geoinformation systems, design a

geoinformation system based on 3D geo-visualization, determine the types of interfaces for displaying and analyzing information; 5) to determine the composition of information necessary for the use of recommended methods, to identify sources of information, assess their availability and financial costs of acquiring information; collection and structuring of necessary information; design and implement a database; 6) to develop the architecture of a Web-based information system (WIS) that integrates mathematical and semantic methods and tools for assessing the impact of energy on the geoecology of the region; database, knowledge base and geoinformation system; to develop a knowledge base structure within WIS; 7) to develop a system of ontologies [8] for describing the domain, to adapt and develop tools for semantic modeling, to construct semantic models for assessing the impact of energy on the geoecology of the region; 8) carry out approbation of WIS and apply the developed methods and technologies to decision making support on the justification and development of recommendations for the development of energy taking into account the requirements of geoecology

2. The impact of energy on the environment.

In recent years, the problem of the impact of energy on the environment has become widespread in the scientific world. Various scientists are trying to investigate the negative consequences of the functioning of energy enterprises for geoecology and identify areas of harmful influence. Below the results of Russian and foreign scientists are considered.

According to Vorobyov V.I. (Russia), the analysis of existing principles for the design and development of large thermal power plants (TPPs), as well as optimization models described by different authors, shows that they do not take into account the actual effects of air pollution, since the specific placement in the area of settlements (especially in built-up areas) that fall into the contaminated zone. On the base of a full-scale instrumental survey of the urban area, Vorobyov V.I. determined the concentrations of harmful impurities at various distances from the pollution source - the territory of the TPPs, the sanitary protection zone, the residential development, confirming the exceeding of the maximum permissible concentration at a distance of up to 18 km. Arslanbekova F.F. (Russia), who investigated the damaging effect on the environment of thermal power plants (TPPs) and motor vehicles, believes that the zone of the most intense air pollution by harmful impurities under torches of TPPs reaches a radius of 3-5 km. Nikiyenko Yu.V. (Russia) investigated the main points of the influence of thermal pollution on the microclimate of the adjacent territories. Based on the calculations, she concludes that the presence of a cooling pond in the area where NPPs and TPPs are located will inevitably lead to negative environmental consequences, including maximum temperature, precipitation and relative importance anomalies. Kozhanov A.A. (Russia) offers methods of geoecological assessment of the influence of the fuel and energy system, based on establishing the relationship between natural conditions and antropogen impact. Studies on the interconnection of Energy, Environment, Climate Change are also conducted abroad, for example, [9, 10, 11, 12], the use of GIS-technologies and 3D-geovisualization are considered in [13, 14].

Nevertheless, in studies related to the assessment of the impact of energy on the environment, the quality of life has not been considered so far.

3. The quality of life

This concept was first introduced into the scientific revolution in the 60s of the last century in connection with attempts by foreign researchers to model the trajectories of industrial development. There are many different definitions of the quality of life, but this concept is most fully disclosed in the context of health care. Quality of life is understood as a set of objective and subjective parameters that characterize the maximum number of sides of a person's life, his position in society and his satisfaction with him. Among the factors determining the quality of life according to the definition of the World Health Organization [15], not only financial well-being, but the state of security, health, human position in society, ecology and, most importantly, its own assessment of all these factors is included (Fig. 1) The integral indicator of the quality of life summarizes the indicators of health, social-personal well-being and financial well-being. In the framework of our project, it is important that ecology is included in the quality of life indicators.

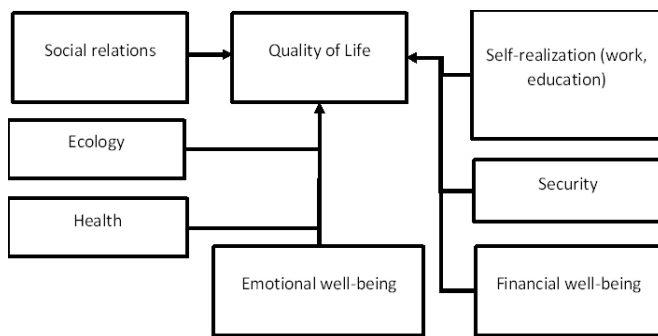


Fig. 1. Quality of life as defined by World Health Organization

Under the guidance of the author, work has been done in which it is justified that it is difficult to obtain an integral index using rigorous mathematical methods. It is suggested to involve artificial intelligence methods for this purpose, namely, cognitive modeling is one of the directions of semantic modeling [5, 16]. In addition, quality of life research in the integrated quality of life indicator has not taken into account the impact of energy supply until recently, while a shortage of energy resources can have a significant impact on both the level and quality of life. It is suggested to include external factors in the quality of life indicators, in particular, the degree of provision with energy resources (Fig. 2) [17].

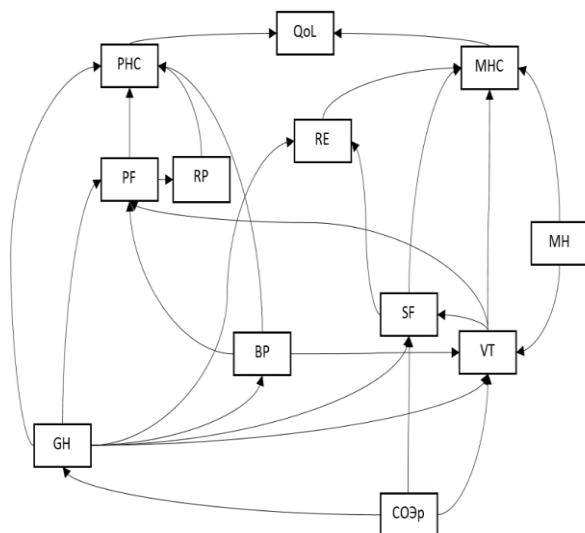


Fig. 2. Cognitive map of indicators of life quality using procedure SF-36 (all links are positive)

Comments on Fig. 2:

- PF (Physical Functioning) – physical functioning.
- RP (Role-Physical Functioning) – role functioning conditioned by physical state.
- BP (Bodily pain) – pain intensity.
- GH (General Health) – general health.
- VT (Vitality) – life activity.
- SF (Social Functioning) – social functioning.
- RE (Role-Emotional) – role functioning preconditioned by emotional state.
- MH (Mental health) – Mental health.
- PHC (Physical health) – general component of physical health.
- MHC (Mental health) – general component of mental health.
- QoL (Quality of Life) – integral indicator of quality of life.
- CO3p (DSEr) – degree of supply with energy resources.

4. Proposed methods and approaches for solving the problem

In the development of tools for intelligent support of energy and environmental decision-making, the proposed project is based on the application of methods of geoinformation technologies based on 3D geo-visualization [18], critical infrastructure research methods [7], decision support methods, knowledge engineering methods, object approach methods (analysis, design, programming), system and application programming methods, design methods of database ,of information and expert systems, as well as author methods of situational management, semantic modeling (primarily ontological and cognitive) and intelligent technologies for decision-making support [5]. It's proposed to develop and adaptate to the project theme the author's methods of constructing an ontological space of knowledge in the field of energy; methods of semantic (ontological and cognitive) modeling in power engineering, methods of 3D-geovisualization and methods of visual analytics with elements of cognitive graphics, as well as methods of intelligent systems development for supporting the adoption of strategic decisions in energy [19].

As an illustration of the approach to constructing a ontologies system of applied domain , Fig. 3-4 present the ontology of the impact of energy on the environment (Fig. 3) and the ontology of pollutants from energy facilities (Fig. 4).

From Fig. 4 it can be seen that energy companies (enterprises of electric power industry, heat power engineering and fuel and enterprises for energy resources extraction and transportation) can pollute water, air and soil (first of all, the chain associated with air pollution is considered in the project, see Fig. 3) Negative impact on the quality of life of a person can be either direct or indirect (through the plant and animal world, i.e. through the food chain).

In Fig. 4 shows the first version of the ontology, of pollutants from energy facilities, built on the basis of [20], which rather illustrates the way of structuring knowledge about the subject area. In the following, the ontology can be extended or used as a hybrid: some of its concepts can be considered as meta-concepts, which will be described by detailed ontologies: for example, the concept of "Purification" can be represented by detailed ontology of the methods and levels of purification; the concept "Combustion method" can be represented by extended ontologies describing different combustion methods for small and large boiler plants and thermal power plants, etc

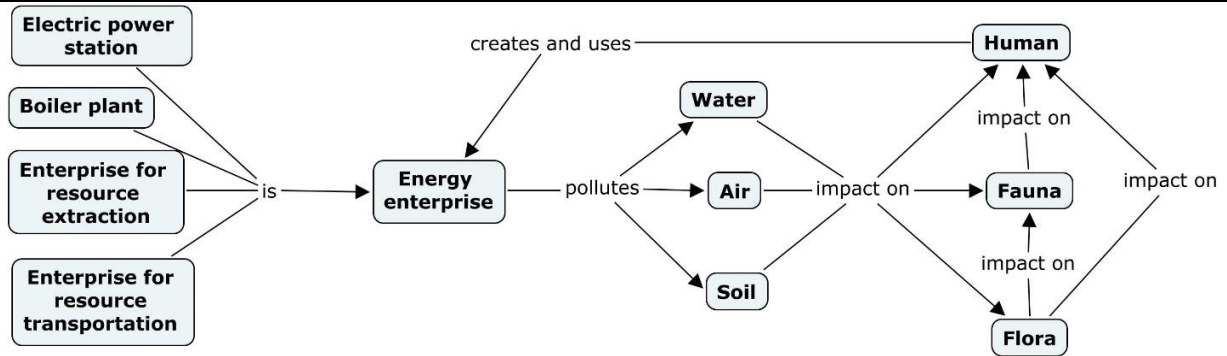


Fig. 3. Ontology of the energy sector impact on the environment

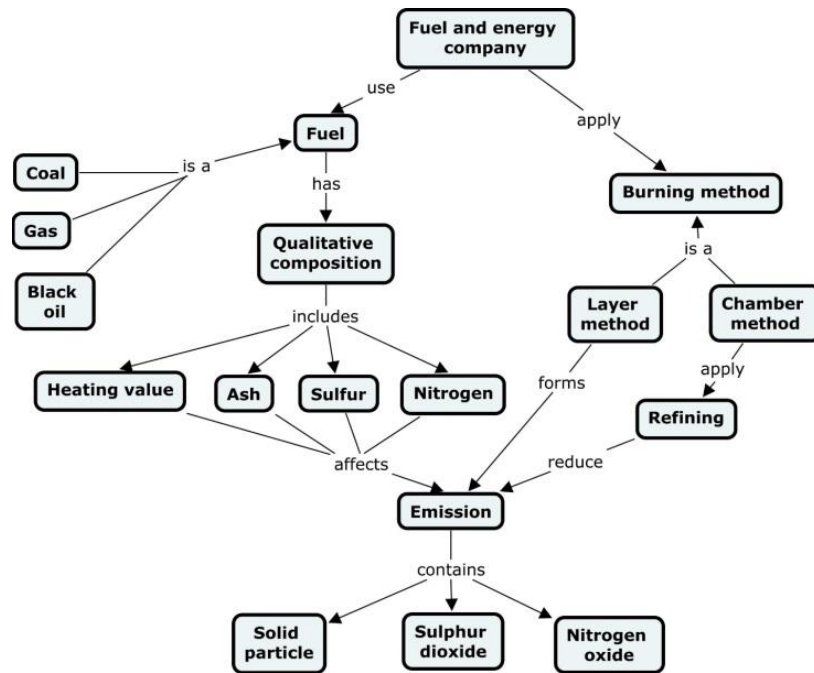


Fig. 4. Ontology of pollutants from energy facilities

As mentioned in the introduction, a Web-oriented information system (WIS) is being developed to implement the project, integrating mathematical and semantic methods, tools for impact assessment of energy on geoecology in the region, a database, a knowledge base and a geographic information system. It is supposed to use the authors' results done earlier to study energy security problems: semantic modeling tools, Geocomponent, knowledge base tools and the Repository. Individual WIS components (Geocomponent, semantic modeling support tools, individual computational modules, database access components), can be implemented as agents-services [6].

The WIS architecture is shown in Fig. 5. There are four levels in the architecture: 1) the level of mathematical methods, models and software, which includes developed on the basis of selected methods and models the software for calculating the volumes of pollutants and their impact on the quality of life of the population, taking into account the capacity of energy facilities (energy supply) and population density in the territory under consideration; 2) the level of semantic modeling, including semantic (primarily cognitive) models for describing the interrelationship of factors that determine the quality of life, taking into account anthropic-technical factors: the provision by energy resources and the influence of pollutants from energy facilities on the environment; 3) the level of knowledge representation - unites the knowledge base storing descriptions of knowledge for constructing semantic models, and an ontology system for describing knowledge of the subject domain; the latter can be used both for building a knowledge base and for

database designing; 4) level of data representation - integrates the geographic information system (GIS) and database, including geographic coordinates of energy facilities. GIS can be used both to illustrate the results of calculations, and for visual interpretation of semantic models.

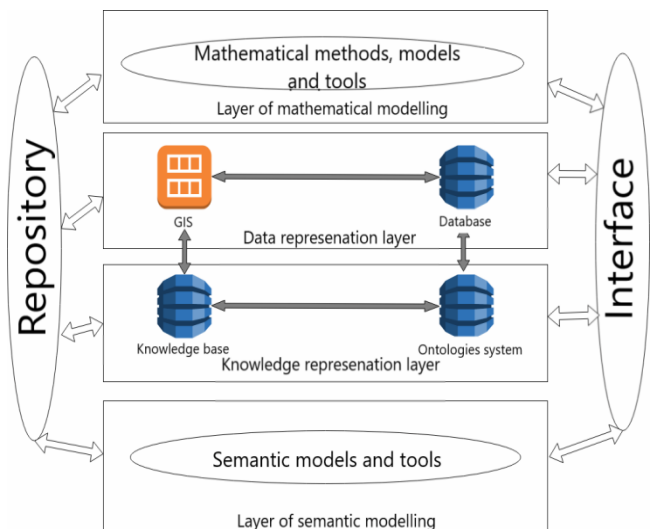


Fig. 5. The architecture of the Web-based information system (WIS) to assess the impact of energy on the geoecology of the region.

The meta descriptions of information presented at all four levels are stored in the Repository (its scientific prototype and tools for working with it were developed by Kopaygorodsky A.N., the co-executor of the project). When implementing the user interface, it is intended to apply the components of the situational management language CML [21].

5. Results and discussion

The article is devoted to the actual problem of assessing the impact of energy on the geoecology of the region. As a rule, emissions from energy facilities are not considered separately in environmental studies, so it is difficult to separate them from general environmental pollution. This makes it difficult to plan and implement measures to reduce pollution by individual energy facilities. The article proposes an approach to solving the problem of monitoring the impact of energy sector on the geoecology of the region, based on the integration of mathematical models, GIS technologies and modern intellectual technologies in the framework of Web-oriented information system. The novelty of the project is also determined by the fact that when assessing the impact of energy on geoecology for the first time it is proposed to take into account the quality of life of the population. Cognitive modeling is seen as a tool for implementing this idea. To organize information support, it is proposed to use the ontology system as a basis for designing databases and knowledge bases. The proposed Web-based system is considered as a prototype of the intelligent decision-making support system for improving the quality of life, taking into account the requirements of geoecology, including an improved analytical tool for estimating emissions of energy facilities and the spread of pollution.

6. Conclusion

- The article considers the International Project, carried out under the guidance of the author with the support of the EASR-RFFI funds.
- The statement of the problem is formulated (the fundamental scientific task and the project goal), the urgency and expected results of the project are determined, and proposed methods and approaches to its implementation.
- The main attention in the article is given to the information and technological part of the project carried out by the Russian side.
- The illustrations of proposed approaches are presented: cognitive map of indicators for assessing the quality of life, taking into account the availability of energy resources, ontology of the energy sector impact on the environment and ontology of pollutants from energy facilities, as well as the developed architecture of the Web-based information system (WIS), which, together with the technology of its use, will be the final result of the project.

Acknowledgments

The results were obtained with partial financial support of RFBR grants № 18-57-81001, № 16-07-00474, № 16-07-00569, № 18-07-00714.

References

1. "Geoecology". In: O.V. Petrov (ed) *Geological Dictionary / 3rd ed., Revised, and additional*. St. VSEGEI, Petersburg, 2010, p. 244. (T. 1. A-I)
2. "Geoecology". In: V.I. Danilov-Danilyan. (ed) *Ecological Encyclopedia*, OOO "Izdatelstvo Encyclopedia", Moscow, 2010, p. 22 (V. 2. G-I)
3. Massel L.V. "The problem of the impact assesment of energy on the geoecology of the region: setting and solutions". *Information and mathematical technologies in science and management* 2018; 2(10):5-21. DOI: 10.25729/2413-0133-2018-2-01
4. Khoroshevsky V.F. "Semantic Technologies: Expectations and Trends". In: *Proceedings of the II International Scientific and Technical Conference "Open Semantic Technologies for the Design of Intelligent Systems"*. BGUIR, Minsk, Republic of Belarus, 2012, pp. 143-158
5. Massel L.V., Massel A.G. "Semantic technologies based on the integration of ontological, cognitive and event modelling". In: *Proceedings of the III International Scientific and Technical Conference "Open Semantic Technologies for the Design of Intelligent Systems"*. BGUIR, Minsk, Republic of Belarus, 2012, pp. 247-250
6. Massel L.V., Galperov V.I. "Development of multi-agent systems for the distributed solution of energy problems using agent scenarios". *Izvestiya of Tomsk Polytechnic University*, 2015; V. 326, 5:45-53
7. Massel L.V. "Convergence of research on critical infrastructures, quality of life and safety". In: *Information Technologies and Systems: Proceedings of the 6th International Scientific Conference*. ChelSU, Chelyabinsk, 2017, pp. 170-175.
8. Gavrilova T.A. "Ontological engineering". *Electronic resource. Technologies of knowledge management*. Access mode: http://www.kmtc.ru/publications/library/authors/ontolog_engeneering.shtml
9. "Energy, Climate Change, Environment". *International Energy Agency*, 2016.
10. K. Pavlickova, A. Miklosovicova, M. Vyskupova. "Effects of Sustainable Energy Facilities on Landscape: A Case Study of Slovakia". In: N. Nakagoshi and J.A. Mabuhay (eds) *Designing Low Carbon Societies in Landscapes, Ecological Research Monographs, Chapter 7*, Springer Tokyo, 2014, pp. 109-127. DOI 10.1007/978-4-431-54819-5_7
11. Hussey, K., and J. Pittcock "The Energy-Water Nexus: Managing the Links between Energy and Water for a Sustainable Future". *Ecology and Society*, 2012; 1(17): 31. DOI: 10.5751/ES-04641-170131
12. Pfeiffer A., Millar R., Hepburn C., Beinhocker E. "The '2°C capital stock' for electricity generation: Committed cumulative carbon emissions from the electricity generation sector and the transition to a green economy". *Applied Energy*, 2016.; 179:1395-1408.
13. Alaa Alhamwi, Wided Medjroubi, Thomas Vogt and Carsten Agert "GIS-Based Urban Energy Systems Models and Tools: Introducing a Model for the Optimisation of Flexibilisation Technologies in Urban Areas". *Applied Energy*, 2017; 191:1-9. DOI: 10.1016/j.apenergy.2017.01.048
14. Liang, J., Gong, J., Li, W., Ibrahim, A. N. "A visualization-oriented 3d method for efficient computation of urban solar radiation based on 3d-2d surface mapping". *International Journal of Geographical Information Science*, 2014; 4(28): 780-798.
15. "WHO Health Impact Assessment Toolkit - Guidance on health impact assessment". *World Health Organization*, 2005
16. Trakhtengerts E.A. "Computer support for decision-making". SINTEG, Moscow, 1998.
17. Massel L.V., Blokhin A.A. "Method of cognitive modeling of quality of life indicators taking into account external factors". *Science and education. Scientific publication of the MSTU. Bauman*, 2016; 4:65-75. DOI: 10.7463 / 0416.0839061 (13).
18. Ivanova I.Yu., Ivanov R.A. "The use of geovisualization in the analysis of the location of energy infrastructure objects of the central ecological zone of the Baikal natural territory". *Information and Mathematical Technologies in Science and Management. Science Magazine*. 2016; 2(4): 80-89.
19. Massel L.V. "Problems of the creation of semiotic-type intelligent systems for strategic situational management in critical infrastructures". *Information and Mathematical Technologies in Science and Management. Science Magazine*, 2016; 1:7-27.
20. Maysyuk E.P., Ivanova I.Yu., Ivanov R.A. "Assessment of the volume and composition of emissions of pollutants into the atmosphere from energy objects of the central ecological zone". In: *Proceedings of the International Scientific and Practical Conference "Current state and prospects for improving the ecology and safety of vital activity of the Baikal region. White Nights - 2016"*. Irkutsk: Publishing House of the IRSNU. - 2016. - Pp. 475-482 (in Russian).
21. Massel L.V., Massel A.G. "The language for knowledge description and knowledge management in intelligent system of semiotic type". In: *Proceedings of 20th All-Russian Baikal Conference "Informational and Mathematical Technologies in Science and Management" (IMT'2015)*, Vol.3. ISEM SB RAS, Irkutsk, Russia, 2015, pp. 112-124.

FEATURE EXTRACTION FROM IMAGES WITH USE OF CONVOLUTIONAL NEURAL NETWORKS: APPLICATION TO SECURING PERSONAL DATA

T. R. Sokhin, K. V. Mironov

Department of Computer Science and Robotics

Ufa State Aviation Technical University

Ufa, Russia

e-mail: qwinpin@gmail.com, : mironovconst@gmail.com

Abstract: *The task of recognizing the bank cards on images is considered in this paper. This task is motivated by premoderation of images in social networks in order to avoid leakage of personal data. An example of such data is the number of bank card. Detection algorithm based on convolutional neural networks is applied for this task. A number of experiments were conducted to develop the most optimal neural network architecture, which took into account the speed and accuracy of recognition. In the course of the experiments, the recognition accuracy of 91% and the processing speed of 25 images per second were achieved.*

1. Introduction s

With the development of network technologies in recent years, the task of monitoring content has become increasingly important. Large companies often use traffic monitoring systems to avoid sending sensitive corporate information over open communication channels. In private life, it is also important to keep track of the content that users put on publicly available resources, such as social networks. The main threat in this case is the unwanted publication of personal data. In this paper, the threat of accidental publication of a bank card image in a social network is considered. In 2016, a study was conducted on the Visa payment system at Newcastle University, which concluded that the application of a distributed brute-force attack allows to receive a CVV card in a matter of seconds may be sufficient to steal from it all available funds.

Thus, the ability to verify photos for having bank cards on them could improve the security of personal data in social networks. Such a check can be built into the content premoderation system. In this paper, we propose an algorithm for recognizing bank cards in photographs.

2. Related works

One of the simplest methods, which allows to significantly reduce the complexity of the image recognition problem is the method of contour analysis. A contour is a curve that describes the boundary of an object in an image. Using this approach assumes that the contour contains enough information about the shape of the object, while internal points are not taken into account. Examining only the contours of objects allows you to transition from the image space to the contour space, which significantly reduces the complexity of algorithms and calculations. The main advantage of contour analysis is the invariance with respect to rotation, scale and contour displacement on the image under test. It is best suited for finding an object of some given shape [2,3].

The next one is a relatively simple method of a comparison with a template (template matching). This method is used to find areas of images that are most similar to some given pattern. The size of the template should be smaller than the size of the image being scanned. Search for a template is made by sequentially moving it along the tested image and assessing the match of each new area with the template. Based on the results of such check, the area that has the highest matching coefficient is selected. In fact, this is the percentage of matches between the area of the picture and the template.

Searching by the template does not allow to say with certainty whether the original object was found, since this is a probabilistic characteristic, depending on the scale, viewing angles, picture rotations, and the presence of physical interference. Also, false triggering of the algorithm is possible, when the object is not actually found, but there are some common details between the template and the area in the image under test. Of course, such a situation can be avoided by checking the value of the matching

coefficient (so that it is not less than some boundary limit), but this does not always work properly due to the reasons described above [2]. The disadvantages of this method include the time of operation, which increases depending on the size of the incoming image.

The idea of searching by key points is that the most significant points are highlighted in the image, which will be preserved even if the size, noise, or lighting would change. This method also has several disadvantages: if one of the elements in the image to be examined is absent or will be greatly changed in the list of key points, the image will not be recognized. In this regard, try to introduce as many similar points as possible, which greatly increases the algorithmic complexity of the method, and, consequently, the recognition time in general.

In the method of Viola-Jones [4], a scanning window approach is used: the image is scanned by the search window (the so-called scan window), and then the classifier is applied to each position. The system of training and selection of the most significant features is fully automated and does not require human intervention, but, despite the considerable accuracy and ease of use, the method has a number of disadvantages that are significant for the algorithm being developed: a relatively long image processing, because it requires a sliding window method across the image many times, with different window scales, and a strong dependence on the size of the image fed to the input.

Neural networks can also be used to recognize objects in images. Now there are many tasks solved with neural networks. It often has no reason – like many young technologies, NN are becoming mainstream. But it is indispensable in area of computer vision. However, not all kinds of NN are useful here; actually, we need to use the principles of visual system of mammals: the reaction to certain details of so-called features that help in object recognition. That NN are called convolutional.

Convolutional neural networks are one of the most recently popular models of deep learning, used primarily for image recognition. The concept of convolutional networks is built on three main ideas [6]:

1) local receptive fields - the recognition of an item in an image should primarily be affected by its immediate environment, while pixels located in another part of the image are most likely not related to this element and not contain information that would help to identify it correctly;

2) shared weights - the same object can be found in any part of the image, and for its search in all parts of the image, the same pattern (set of weights) would be applied;

3) subsampling - when compared with a pattern, not the exact value for a given pixel or area of pixels is taken into account, but its aggregation in some neighborhood, for example, the average or maximum value.

From the mathematical point of view, the basis of convolutional neural networks is the operation of matrix convolution, which

consists in the elementwise multiplication of the matrix, which is a small portion of the original image (for example, 7×7 pixels) with a matrix of the same size, called the convolution kernel, and the subsequent summation of the obtained values. In this case, the core of the reconciliation is essentially a certain pattern, and the number resulting from the summation characterizes the degree of similarity of the given area of the image to this template. Accordingly, each layer of the convolutional network consists of a number of patterns, and the task of learning the network is to select the correct values in these templates - so that they reflect the most significant characteristics of the original images. In this case, each pattern is matched sequentially with all parts of the image - it is in this that the idea of dividing the weights finds expression. Layers of this type in a convolutional network are called convolution layers. In addition to convolution layers, there are sub-sampling layers in convolutional networks or sub-sampling, which replace small areas of the image with one number, thereby simultaneously reducing the size of the sample for the next layer and making the network more resistant to small changes in the data. In the last layers of a convolutional network, one or several fully connected layers are usually used, trained to perform a direct classification of objects.

Before 2012, at which CNN demonstrated their abilities at the ImageNet competition [5], this technology was not in demand, because of weak database and expensive hardware. But now, we have high computing power per dollar and millions of databases in free access. The methods of contour analysis [2] (the object presented in the form of the exterior outlines), template matching (finding small parts of an image which a template image) in general and Haar-like features in particular are far behind [3].

3. Architecture

In order to study the technologies and to design information security system, a model of CNN capable to recognize image of a bankcard has been developed. In cases of publication of those in social networks owner is deprived of support even when his money is stolen. Detection of such images allow to take measures to prevent event like that or another threats. To work with NN used TensorFlow with TFLearn API [7], CPU first-generation i7 (only because OpenCL still are not supported enough) and not data of card from 2000 images.

There are first-priority principles for building NN:

- Avoid representational bottlenecks early in network. The size of image was chosen 200×200 .

- Increasing the activations per tile allows for more disentangled features. Higher dimensional representations are easier to process locally within a network.

- Spatial aggregation can be done over lower dimensional embeddings without much or any loss in representational power. That promotes faster training without loss in quality, but practice shows decrease in efficiency with dimension reduction on early layers.

- Balance the width and depth of the network. It is more important with high-performance system.

Consideration of several typical architectures led to the ideas of the Inception v3 network: the use of convolutions of various sizes on one source data and their subsequent integration into a single feature map. In the early layers, the correlated neurons concentrate in local areas, which means that if several neurons in the same coordinate can learn about the same thing, then in the tensor after the first layer, their activation will be in a small region near some point. The greatest number of such correlations can be obtained with convolutions 1×1 . Slightly less with 3×3 and even smaller with convolutions 5×5 . This allows to extract features of the image at various levels of detail while preserving the original image.

Also used the idea of using 1×1 filter isolate the core of feature on large area (a convolution of 5×5 or 3×3 pixels) in the image.

We know convolution of 5×5 can be replaced by two convolutions of 3×3 , which reduces the number of network parameters, and hence the learning and networking time. For example, convolution 5×5 contains 25 parameters. If we replace it with a stack of two layers with the contractions 3×3 , we get the same mapping, but the number of parameters will be less - 18, which is 22% less. Next, we can proceed to asymmetric convolution $1 \times n$ and $n \times 1$ type, but this factorization does not work well on early layers, it gives very good results on medium grid-sizes (On $m \times m$ feature maps, where m ranges between 12 and 20), so it is useless in this case. For reducing overfitting in neural networks in final layers we used dropout [8] - at each stage of training, some nodes are ignored with a probability of 0.5.

3. Experimental results

Two different structures of neural networks were developed. One of them used only successive convolutional layers, terminating in two fully connected layers in fig. 1. The second network used layers from several parallel convolutions and merged the results in fig. 2.

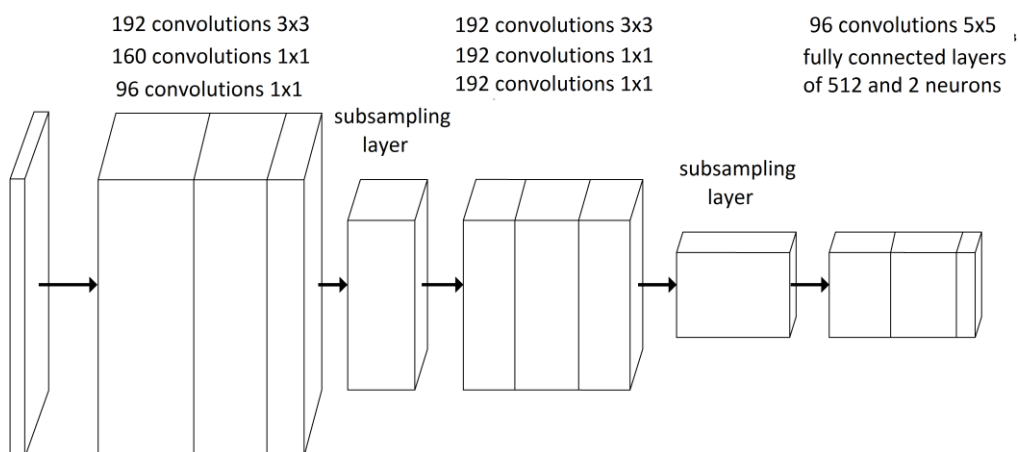


Fig. 1. "Classical" architecture of the CNN applied for the recognition of bank cards.

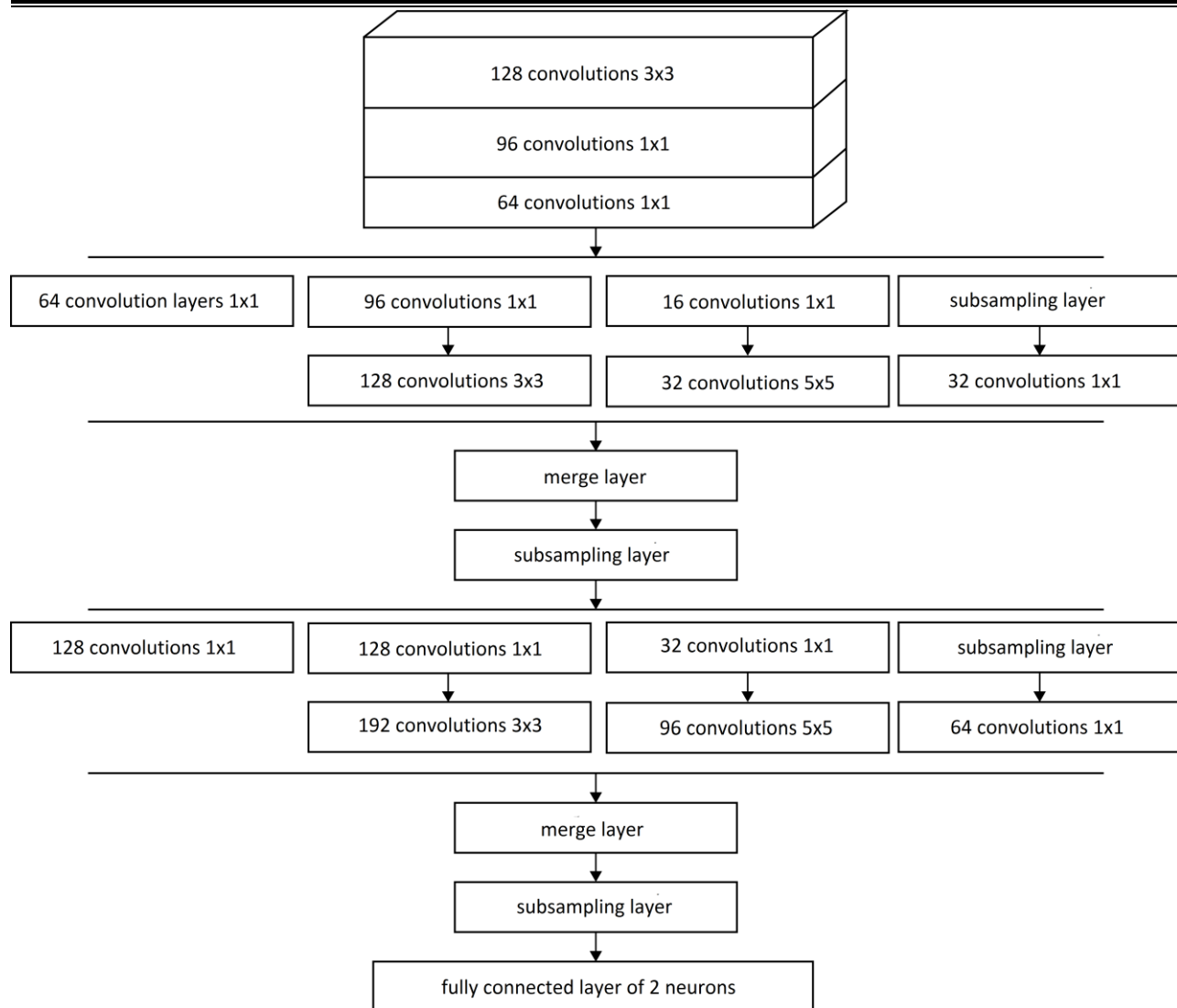


Fig. 2. CNN based on "Inception" model

The training sample of images includes 2000 photos in jpg format. The test sample images has 1000 photos in jpg format. The input of the network is supplied with images of 32 by 32 pixels in 3 color RGB channels. At the output of CNN, matrices containing the characteristics of the original image are obtained, then these matrices are transferred to the fully connected layers - perceptrons - from 512 and 2 neurons, whose output indicates the neural network's confidence in the object's belonging to the image of one of the classes. In this case it is "Is a bank card" and "Is not a bank card".

In order to achieve maximum efficiency, the training was conducted with the following parameters:

- Learning rate;
- Dropout percentage - the percentage of training images that will not be used to train the weights of the neural network.

Turning to the work of Jan LeCun in 1988 [6], which proposes the idea of convolution and pulling using the method of learning the backward propagation of an error, we obtained a classical "flat" convolutional neural network, depicted in Figure 1. The final accuracy of the neural network is 70.89 %. This neural network demonstrates good results, however, there are signs of retraining during testing - low accuracy of detection of bank cards with designs that differ from given ones.

The best result is 79,64% with recognize speed about 21 image per second with structure in fig. 2. The decision to use neural networks, which consist of elements whose functionality is analogous to most functions of a biological neuron, first originated in the 30s of the last century (Pitts, Rashevsky, Turing). However, only recently they have found practical application, including in the field of information security. The developed model includes the most successful solutions obtained from the existing experimental database of researchers from around the world. Based on these techniques, the best architecture of the convolutional neural network and training parameters were empirically obtained, allowing to achieve a high result, taking into account the existing limitations on processing speed and processing power. The already existing practice of introducing neural networks into social networks, other resources on the Internet demonstrates that the ratio of implementation costs and support to the benefits obtained is extremely positive.

The best result of training obtained with the Inception model, is 91.35% with an average processing time of 1000 images in 47 seconds, which is almost 1.5 times better than the results of the previous model. This model can be used to process a large number of images with high accuracy.

7. Conclusion

The task of protecting personal data has one of the highest values in the field of information security at the current time. Bank cards, which are so common among all segments of the population, have many vulnerabilities that allow them to gain access to funds through them. Together with the possibility of an easy access to the Internet and social networks in particular, even for those who do not have sufficient knowledge of the simplest precautions, in order to ensure the security of their personal data, this becomes a wide-spread problem.

It is possible to protect this vector of the distribution of personal data through the participation of the social networks themselves, which is why the possibility of premoderating the content that the user publishes is offered. To enable fast enough moderation of images with minimal increase in equipment costs and corresponding costs, it was necessary to select an image recognition algorithm that combines high speed with accuracy of recognition. As a result, it was decided to use neural networks. Convolutional neural networks are used to work with images.

Due to the fact that there is no full theory of architecture development and training of neural networks, a significant part of the research work relied on the method of experimental study, during which the most optimal architecture of a convolutional neural network was developed, which ensures the accuracy of bank cards recognition on the image in 91% and speed recognition of 1000 images in 47 seconds.

Acknowledgements

The research work is supported by the Russian Fund for Basic Research, grant #16-07-00243.

References

1. M. A. Ali, B. Arief, M. Emms, A. van Moorsel. "Does the Online Card Payment Landscape Unwittingly Facilitate Fraud?" IEEE Security & Privacy, vol. 15, No. 2, pp.78 to 86, 2016.
2. ArealIdea "Analysis of Computer Vision Algorithms" Web-page. <https://arealidea.ru/articles/stati-i-publikatsii/analiz-algoritmov-kompyuternogo-zreniya-poiska-obektov-i-sravneniya-izobrazheniy/> visited on 30.06.2018 (in Russian).
3. I. O. Sakovich, Yu. S. Belov. "Survey on basic methods of contour analysis for highlighting the contours of moving objects" Engineering Journal: Science and Innovation, No. 12, 2014 URL: <http://engjournal.ru/catalog/it/hidden/1280.html> visited on 30.06.2018 (in Russian).
4. P. Viola, M. Jones "Rapid object detection using a boosted cascade of simple features" Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001.
5. ImageNet Large Scale Visual Recognition Competition, <http://www.image-net.org/challenges/LSVRC/2017/> visited on 30.06.2018.
6. Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard and L. D. Jackel: Backpropagation Applied to Handwritten Zip Code Recognition, Neural Computation, 1(4):541-551, Winter 1989.
7. Tensor Flow: An open source machine learning framework for everyone, <https://www.tensorflow.org/> visited on 30.06.2018.
8. Cireşan D.C. "Deep big multilayer perceptrons for digit recognition" – Springer Berlin Heidelberg, 2012. – c. 581-589.