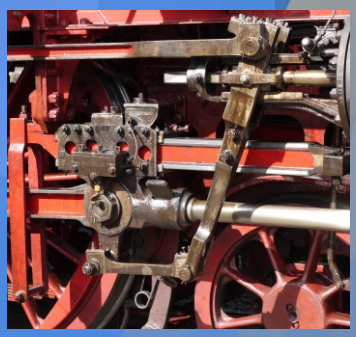
INTERNATIONAL SCIENTIFIC JOURNAL

INDUSTRY 4.0

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

YEAR III ISSUE 4/2018






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 4 / 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. Gerard Lyons, IE
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 Velev, 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, DE

C O N T E N T S

TECHNOLOGICAL BASIS OF "INDUSTRY 4.0"

MONITORING AND CONTROL OF CYBER-PHYSICAL SYSTEMS – THE BACKBONE OF INDUSTRY 4.0
Prof. Dr. Batchkova I. A., Prof. DCs. Popov G. T., Eng. Belev Y. A., Eng. Tzakova D. D. 157

ANALYZING EMOTIONS FROM TEXT CORPUS USING WORD SPACE
Mohsin Manshad Abbasi, Professor Anatoly P. Beltiukov 161

METHOD OF STRUCTURING THE SELF-ORGANIZED INTELLECTUAL SYSTEM ON THE BASIS OF REQUIREMENTS OF THE ISO/IEC 15288 STANDARD IN THE FORM OF THE CARTESIAN CLOSED CATEGORY. (ON THE EXAMPLE OF DESIGN OF INFORMATION AND ANALYTICAL SYSTEM)
G.G. Kulikov, V. V. Antonov, A.R . Fakhrullina, L.E. Rodionova 165

CONCEPTUAL MODEL "STAGE" AND INTELLECTUAL SUPPORT OF THE SOLUTION PROCESS OF "TACTICAL PROBLEMS" BY CREWS OF AIRCRAFTS
B.E. Fedunov, A.A. Plyatcovoi 168

IMPROVED STOCHASTIC CONTROL FLOW MODEL FOR LLVM-BASED SOFTWARE RELIABILITY ANALYSIS
V. Vidineev, N. Yusupova, K. Ding, A. Morozov, K. Janschek 172

DOMINANT TECHNOLOGIES IN "INDUSTRY 4.0"

METHODOLOGY FOR CONFIGURING AN INTELLIGENT NETWORK EXPERT SYSTEM [AINES] USING THE EXAMPLE OF AN ELECTROCHEMICAL DIMENSIONAL PROCESSING MACHINE
Davletbaev A.S., Kulikov G.G., Startcev Y.V. 175

GEOINFORMATION ANALYSIS WITHIN THE OBJECTIVES OF WASTE MANAGEMENT SYSTEMS CONSTRUCTION
V.E Gvozdev, O. Khristodulo 180

MACHINE LEARNING APPLICATIONS FOR DESIGN OF NEW MATERIALS: A REVIEW
Sr. Lecturer Kumar S. PhD., Prof. Nezhurina I. M. PhD. 186

BUSINESS & "INDUSTRY 4.0"

DYNAMIC COGNITIVE MAPS FOR THE SUBSTANTIATION OF STRATEGIC DECISIONS ON MANAGEMENT OF ENERGY SECTOR DEVELOPMENT
A.G. Massel 190

SOLVING THE POINT-PLANE PROBLEM FOR THE CLASS OF AFFINE TRANSFORMATIONS AND DEVELOPMENT OF A FAST ITERATIVE ALGORITHM FOR REGISTERING OF 3D POINT CLOUDS
A. Vokhmintsev, A. Melnikov, T. Botova 194

USING DATA MINING TECHNIQUES TO CREATE AN AUTOMATED MODEL THAT MAKES COMPARISONS BETWEEN MARKET DEMANDS AND UNIVERSITY CURRICULA
M.Sc. Ylber Januzaj PhD (C)., Prof. Assoc. Artan Luma PhD., Prof. Assoc. Azir Aliu PhD., Prof. Assoc. Besnik Selimi PhD., Prof. Assoc. Bujar Raufi PhD., Prof. Assoc. Halil Snopce PhD., Prof. Ass. Vehbi Ramaj PhD. 199

ANALYSIS OF DATA TO DETERMINE CUSTOMER LOYALTY IN THE EDUCATIONAL SPHERE USING KOHONEN MAPS
I. R. Gallyamova, N. I. Yusupova, O. N. Smetanina, L. Y. Uzbekova 203

SOCIETY & „INDUSTRY 4.0"

THE STRUCTURE OF SECURE SYSTEM FOR COLLECTION, STORAGE AND PROCESSING OF TELEMETRIC INFORMATION ON THE SLATE OF AIRCRAFT SUBSYSTEMS
V.V. Berkholts, A.M. Vulfin, M.B. Guzairov, A.I. Frid, K.V. Mironov 209

WEB-BASED INFORMATION SYSTEM FOR THE IMPACT ASSESSMENT OF ENERGY ON THE GEOECOLOGY OF THE REGION
Massel L.V., Kuzmin V.R 213

THE METHOD OF MATHEMATICAL MODELING OF PROCESSES TO IDENTIFY, ESTABLISH CHARACTERISTICS AND RELATIONSHIPS OF THE CONNECTIONS OF SOCIALLY SIGNIFICANT PHENOMENA
V.V. Antonov, Ya.S. Mikhailova, V.A. Kolesnikov, L.V. Chernyshova 216

DEEP LEARNING BASED PERSON BIOMETRIC IDENTIFICATION
M.R. Bogdanov, I.N. Dumchikova, I.S. Dokuchaev, A. Dumchikov 219

MONITORING AND CONTROL OF CYBER-PHYSICAL SYSTEMS – THE BACKBONE OF INDUSTRY 4.0

Prof. Dr. Batchkova I. A.[1], Prof. DCs. Popov G. T.[2], Eng. Belev Y. A.[1], Eng. Tzakova D. D.[1]
[1]University of Chemical Technology and Metallurgy, Sofia, Bulgaria
[2]Technical University of Sofia, Sofia, Bulgaria

idilia@uctm.edu, gepop@tu-sofia.bg

Abstract: *Monitoring and control of Cyber-Physical Systems (CPS) have many challenges related to the heterogeneous environment, the high degree of interaction between the components and the high requirements for functionality and scale. The paper presents an analysis of the state of the art in this area and proposes an approach for monitoring and control of CPS through the integration of the IEC-61499 based models for distributed control systems and the IEC-61512 standard providing domain specific models for design and control of batch production processes as well as the Signal- Interpreted Petri Nets used for the purpose of formal specification and verification of unit procedures. The approach is illustrated with an example of CPS for loading a batch tank and mixing of liquid flow.*

Keywords: CYBER-PHYSICAL SYSTEMS, INDUSTRY 4.0, MONITORING, CONTROL, IEC-61499, IEC-61512

1. Introduction

The rapid development and broad penetration of information and communication technologies in the industry has led to the emergence of new industry development strategies with a view to enhancing its competitiveness, such as: the German Initiative "Industry 4.0" [1], the American "Smart Manufacturing" [2], the Chinese "Made in China 2025" [3], the Japanese IVI (Industrial Value Chain Initiative) [4], the Italian "Industrial National Plan 4.0" [5], etc. All of them aim "The complete transformation of the whole sphere of industrial production through merging of digital technology that and Internet with conventional industry". Recent studies have shown that digitization of products and services can add more than 110 billion Euros a year to Europe in the next five years [6]. Successful implementation of all these initiatives is possible with the use of Cyber-Physical Systems (CPS) technologies, which are considered to be new types of systems that expand the capabilities of the physical world through computing, communications and control, and upgrading the electronic automation.

Cyber-physical systems (CPS) represent integration of computational and physical processes. They use embedded computers and networks for the purpose of monitoring and feedback control of physical processes [7]. Through cyber-physical systems, the physical world connects to the virtual world to form Internet of Things, Data, and Services. One of the most promising and challenging applications of CPS is the cyber-physical production system. Research in the field began in 2006 in the United States, when the President's Council of Advisers for Science and Technology (PCAST) identified research in the field of CPS as a national priority and hundreds of millions of dollars have been invested since then [8]. The European Industry Association ARTEMIS-IA [9] also focuses its interest on embedded systems and their evolution towards the CPS, with around € 2.7 billion invested so far, mainly in more than 50 projects under the European Framework Programs (7 Framework Program and Horizon 2020). It is commonly believed that the development and implementation of CPS should be based on a new scientific basis in order to build a bridge between sequential semantics and parallel physical world.

To meet the challenges of modern enterprises, CPS must meet certain requirements, which can be summarized as follows:

- The architecture of these systems must be decentralized, based on the knowledge product/resource;
- Interactions between the elements of these systems must be abstract, generalized and flexible;
- Control must be reactive and proactive;
- Control must be self-organizing.

The above requirements can be achieved using a range of approaches and methods from different scientific fields. Some of the key topics in research and development are:

- Modeling, Simulation and Verification of CPS;
- Development and application of software process models in the development of CPS (Modeling of the development life cycle of CPS;
- Reference Frameworks and Architecture of CPS;
- Development of semantic service oriented architectures of CPS;
- Distributed control of CPS using advanced methods and algorithms;
- Interoperability in CPS and between them;
- Data analysis (big data) and decision making.

The main objective of the research is to develop and present an approach for monitoring and control of cyber-physical systems, based on defined requirements to the cyber-physical control systems. The paper is organized in 4 parts. After the introduction, in part 2 a short review and analysis of different approaches and methods for monitoring and control of cyber-physical systems is proposed. Part 3 presents the suggested approach based on the reference architecture and models of the IEC-61499 standard for distributed measurement and control systems [10], the standard IEC-61512 providing domain specific models for design and control of batch production processes [11] and the Signal Interpreted Petri Nets [12] as a tool for specification and verification of unit procedures. In Part 4 of the paper the approach is illustrated with an example from the domain of batch CPS. Finally some conclusions are made.

2. State of the art in monitoring and control of CPS

2.1. Basic requirements to the CPS

Cyber-physical systems can be seen as successors of the steam engine governor, emerging during the Industrial Revolution in the eighteenth century and closely related to the emergence and development of control theory. The subsequent advance in the field of computer, computing and communications technologies causes the Information Technology revolution, and their convergence with control and automation has led to the emergence of so-called cyber-physical systems. The use of key enabling technologies such as: ubiquitous embedded computing, sensing, wireless networking technologies, multi-objective optimization, decision-making tools, specification and verification technologies are paramount to the development of smarter, more flexible, reliable and efficient cyber-physical systems. Some of the most important features that distinguish today's cyber-systems from older generation control systems are defined in [13] as:

- Much larger scale of CPS;
- CPS components operate in a heterogeneous environment;
- CPS components interact with each other in a very complex manner;
- CPSs should be highly extensible for new functionalities, and flexible for runtime adaptation.

2.2. Reference architecture of CPS

NIST and in particular the created CPS Public Working Group offers a reference architecture of CPS, shown in Fig.1, consisting of 6 layers and even more transversals capabilities, each of which is included in each layer [14]. The first layer "physical systems" includes all the engineered devices. The sensors on the second layer acquire data from the physical systems, and transmit that information to the next layer and actuators receive data from the control components and act on the physical systems in an adequate way. Monitor and control systems at the third layer consist of hardware and software components, which acquire data from sensors, perform local processing and control the actuators in order to reach the wanted state in the physical systems. The forth layer "Data analytics" includes software for processing, filtering and storing the information from different control components. It is also possible to include software for pattern recognition, decision-making, predictive analysis and machine-learning. The fifth layer (Modeling, Optimization and Simulation) develops and maintains dynamic computing models. The sixth layer (business and consumer goals) refers to the measurable objectives defined by the users to be achieved. In this paper the focus is on the third layer, as in the next section the current state of the art in the field of CPS monitoring and control will be analyzed.

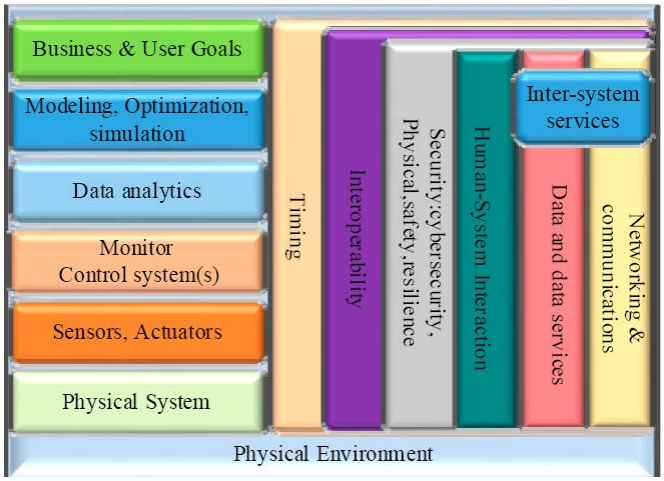


Fig.1: *NIST Reference architecture of CPS [14]*

2.3. Monitoring and control of CPS

The structural and behavioral complexity of cyber-physical systems poses great challenges in terms of the methods and environments for their design and analysis. It is necessary to develop the theoretical foundations of CPS, as well as to create software platforms with appropriately defined levels of abstraction, architecture, languages for modeling different aspects of CPS and transformations between these models. Especially important are methods that have to integrate the discrete dynamics of the computing part with the continuous dynamics of the physical part and the stochastic nature of communications, which must be expanded to cover a wider context. There are three main approaches to designing the CPS:

- Networked control systems - The emphasis on these approaches is on the communication between the different components of the system. There are two main approaches for data acquisition from the physical part of the CPS: timed driven and event driven sampling. The latter approach requires continuous monitoring of the physical system. Compromise approach, the so-called self-triggering approach, is related to the determination of safe intervals during which the physical system is not observed and the time when the data is collected. Particular attention is also paid to methods of addressing the effect of delayed network signals, such as scheduling or stability analysis methods. Other important tasks that seek a solution, especially from the field of control theory, are to determine the optimal placement of computations and to deal with the availability of channels with very low data rates. An overview of networked CPS is given in [15].
- Hybrid control systems that emphasize their continuous and discrete dynamics. Various approaches are known, among which the most popular are hybrid automata using different mathematical formalizations to reflect the transition between discrete states and the evolution of the continuous states over the time. Among the most frequently used are finite state machines and timed automata, bisimulation, transition systems, linear hybrid automata (LHA), rectangular hybrid automata (RHA), temporal logic and others. It should be noted that there are a lot of software tools that allow the synthesis of control systems and their formal verification, such as UPPAAL, HyTech, etc. Some analysis of these approaches are presented in [13, 16];
- Distributed hybrid systems – One of the most often applied approaches uses software platform, called middleware, which uses an appropriate abstraction of complex systems and offers architecture for the rapid deployment of CPS applications [17]. The successfully applied architectures include: component-based, service-oriented, agent-based, and the CPS 5C (connection, conversion, cyber, cognition and configuration) architectures [18]. Another successfully applied approach is based on the Embedded Virtual Machine (EVM) and uses a modular architecture that separates the tasks from the unreliable physical part allowing the integration of system components and their run-time reconfiguration [19]. Other successful approaches use specialized programming languages such as Giatto, Esterel, Signal, etc. or the Model Driven Development (MDD) approaches. MDD [20] are some of the most promising and challenging approaches for development and maintenance of highly distributed control systems such as CPS. Here the systems are presented as models that conform to meta-models, and the model transformations are used to manipulate the various representations. The main difference from other development methods based on models is that MDD uses models as inputs to parameterized implementation generators, i.e. implementation is (semi)automatically generated from the models.

In the next section of the paper an analysis of the IEC 61499 standard in respect to its application for development of distributed monitoring and control for CPS is presented and an approach for these purposes is proposed.

3. An approach for monitoring and control of CPS based on IEC-61499 standard

3.1. Analysis of IEC-61499 standard in respect to CPS

Contemporary CPSs are characterized by their high dimension and complexity, including a variety of decision-making capabilities and control logic. The degree of communication of these systems with physical processes, based on algorithms with increased efficiency and robustness, also increases. The amount of logic-based programming code far exceeds traditional control algorithms.

The IEC 61499 standard defines different reference models supporting the design of distributed control systems at physical, logical and conceptual levels and from different viewpoints. The key models of an IEC 61499 based distributed control system are system, device, resource and application models as shown in Fig.2. All these models are based on the Function Block (FB) concept. Three different kinds of FBs are defined: basic (BFB), composite (CFB) for encapsulation of complex functionality through networks of BFB and service interface function block (SIFB) for providing interfaces for unidirectional (publish/subscribe) and bi-directional (client/server) communications as well for resource or device management. An application model is a network of FBs and may be executed by one or more devices, including one or more resources.

Resource models provide support for program execution by their scheduling part, communication and process interfaces. A device is a control unit having one or more processors that defines specified function for the purposes of automaton and has two types of interfaces - process and communication in order to communicate with the process and other devices on the network respectively. The system model consists of a number of devices, with global or local communication links between them. From the above said it is clear that combining encapsulation of functionality, component-based design and event-driven execution and distribution accelerates and facilitates significantly the dynamic reconfiguration of control systems.

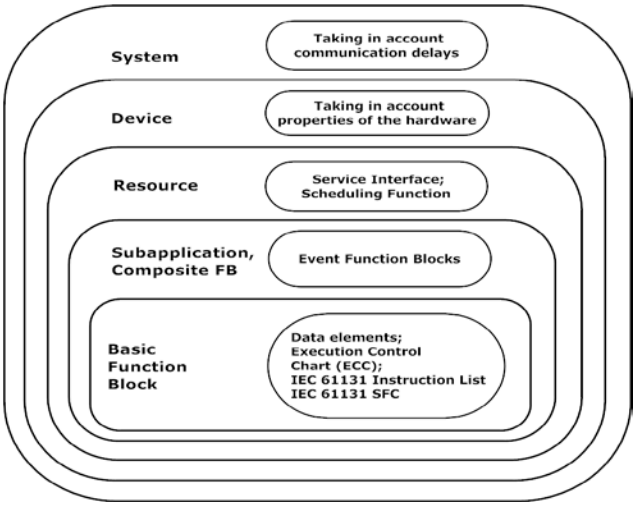


Fig.2. *Basic models in IEC61499*

Extra for the configuration purposes the so-called Management SIFBs are also considered in the standard. They enable the management of devices, resources and applications as it is possible to define and use 6 operations - create, initialize, start, stop, delete and query and to provide notification of changes in availability and status of data types, function block types and instances, connections among function block instances (IEC TC65/WG6, 2005). Thereby IEC 61499 provides the basic interfaces to support dynamic reconfiguration, but the mechanisms to do it are still under development.

There are several engineering environments supporting the development of control systems based on the IEC-61499 standard. The most popular and most commonly used open source environment is the FBDK (Function Block Development Kit) [21]. The product is developed by Rockwell Automation and allows the definition of basic and composite functional blocks, resources, devices, and system configurations. XML format and Java interface are used to support the visual testing of the developed models. In the FBDK environment there is a library of different functional blocks, resources and devices, as well as graphical interfaces and simulation capabilities. Other known environments are 4DIAC [22], FBench [23], ISaGRAF [24], and nxtStudio [25]. Although the industry is aware of the benefits of using IEC 61499, this standard has not yet been widely applied in the industry. The main reasons for this are the immateriality of development tools and the lack of a sophisticated integrated development environment (IDE) and integrated design methodologies to facilitate the component-based development of automation systems throughout the development life cycle.

The main advantages of the standard with regard to CPS can be summarized as follows:

- IEC-61499 allows the integration of time-triggered mechanism with event-triggered systems, based on SIFB;
- The IEC-61499 based reference architecture and models supports the reusability of developed components of CPS and shortens the time for their development and configuration;
- The achieved distributed control structure facilitates the reconfigurability of the system through inserting, deleting and replacing IEC-61499 based functional components;
- The suggested models allow the creation of library with IEC-61499 based reusable components for the different application domains.

The main shortcomings in the implementation of the standard with regard to CPS are the following:

- IEC-61499 does not provide opportunities for verification of control logic, both in the design phase and after online reconfiguration;
- It does not allow modeling of the physical part of the system;
- There is no possibility of performing parallel control logic;
- There is a need to improve the execution semantics, in order to independent from deployment platforms (deterministic) and to address the real-time constraints [26];
- Lack of modern interfaces for communication with the upper management and control levels;
- Middleware for Real-Time Communications enabling distributed elaboration of signals and cognitive functions of CPS.

3.2. Short description of the suggested approach

The proposed approach for monitoring and control of CPS supports the development of reusable software components based on the combined use of three different formalisms: the IEC-61499 standard for distributed process measurement and control systems, the IEC-61512 standard for batch control and the Signal Interpreted Petri Nets (SIPN), which are used as a tool for formal verification of the correctness in the behavior of the developed components or CPS. IEC-61512 provides domain specific models for design and control of batch production processes. The models allow the description of continuous production of finite quantities of materials (batches) from two distinct views – physical and control (cyber). The approach includes the following steps:

- Functional component development - each physical component as for example pressure sensor, temperature sensor, level sensor, valve and pump, has corresponding cyber component (i.e. functional component). If several components are employed for an equipment or unit module, their software components are then compounded as a composite component. The functional component can be instantiated several times.
- Control recipe modeling using SIPN - the procedural control of each unit procedure
- Mapping of model to an IEC 61499 based application - The SIPN model with regard to the functional component is mapped into IEC-61499 by using some rules.

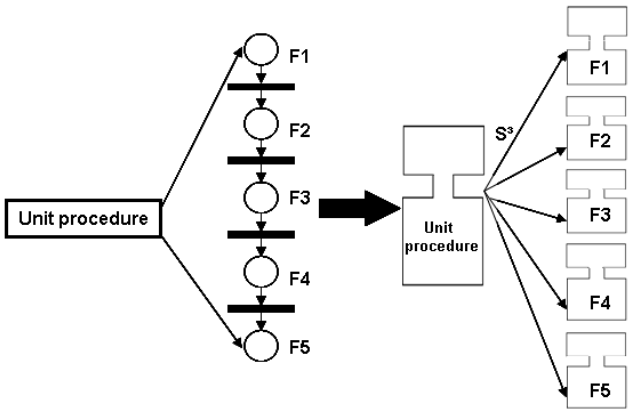


Fig.3: *Basic models in IEC-61499*

4. Case study: Functional component of batch CPS

4.1. Short description of the CPS

The goal of the described case study is to provide highly reusable components concerning batch process and an easy way to

reconfigure their executions. As a result, the control strategy can be flexibly managed. The physical system, called "Loading" is shown in Fig.5-1 and includes equipment for pumping a batch tank with liquid feed and its optimal mixing. The informal specification of unit procedure is shown in Fig.4.

1. VP-01(1) || VA-06(1) || VA-07(1) || VA-09(1) || PV-01 (1)– *"for vacuum creation purpose"*
2. **IF** PI2.04 = 280 mbar **THEN** VA-08(1) || PIC01 is "automatic mode"
3. **IF** dPIC-01= 25 mbar **THEN** dPIC01 is "automatic mode"
4. **IF** LS2.01 = Lmin (300 l) **THEN** dPIC-01 - "manual mode"
5. **IF** PIC.01 ≥ 200 mbar **THEN** VP-01(0) || PIC01 – "manual mode" || PV-01 (0)
6. **IF** PIC1.01 = 0 mbar **THEN** VA-06 (0) || VA-07 (0) || VA-08 (0)
7. **IF** PI2.03=0mbar **THEN** VA-09 (0)

Fig.4: *Informal specification of unit procedure*

The SIPN model of the unit procedure is shown in Fig.5-2. It is modelled by six places and five transitions. Places with a circle shape show the related component, which is valid if it has a token. For valve, 1 means opened and 0 closed. For pump, 1 means turned-on and 0 turned-off. Besides, level sensor is mapped into transition (i.e. bar), while the time condition is represented at the arc (i.e. directed arrow). The unit procedure will be reused for a new batch process.

Mixture Control is a component which is responsible to control the route of liquid feed to distillation cube. The FB composes four valves and two pumps. It is responsible to open and to close the route to the distillation cube. The reuse of FB valve is three times, of FB pump is one time. The implementation of Mixture Control in a composite FB is shown in Fig.5-3 and Fig.5-4. As a starting point, the elements of SIPN model represent the schedule model of the "Loading" unit control are interpreted by FB's identity. S³ technique is applied with corresponding FBs (Scheduler, Selector, and Synchronizer) to control the component based on the given component schedule. The component schedule representation uses the following SOP (Scheduler of Operations), SOP_Setup = [21, 1, 2]; SOP_Stop = [12, 26].

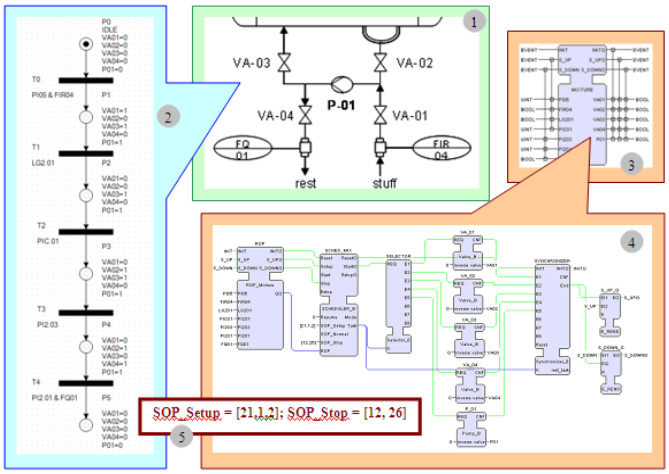


Fig.5: *Illustration of the case study*

5. Conclusions

One of the main benefits of the proposed approach is that the development of cyber components for the related physical components is influenced by some common properties in Batch Control in order to improve flexibility. It is achieved by using generic re-usable control components that can be reused to control different equipments and processes with similar functionality. The suggested formal specification and verification using SIPN satisfies the following three important for the CPS requirements: (1) find conflicting outputs especially in parallel executions, (2) find undesired output combinations (e.g. referring to a combination of opened valves that is forbidden in the current setting), (3) verification of specified flow.

6. References

1. Kagermann H., Wahlster W., Helbig J., Recommendations for implementing the strategic initiative INDUSTRIE 4.0, Final report of the Industrie 4.0 Working Group, Akatech, April, 2013.
2. Members of SMLC, Implementing 21st Century Smart Manufacturing, Workshop Summary Report, 24 June, 2011.
3. Shao Y., Strategic Vision and outlook of "Made in China 2025", Industrial and regional policies.
4. Miyazawa A., Usami R., Through IoT - Japanese factories Connected Together, METI Journal, No5, 2015.
5. Ministry of Economic Development, "Industria 4.0" Italy's national plan for industry, February, 2017.
6. European Commission, Commission sets out path to digitize European Industry, Press Release, Brussels, 19.April 2016.
7. Lee, E.: Cyber Physical Systems: Design Challenges. Technical Report. Berkeley: University of California, 2008.
8. CyPhERS - Cyber-Physical European Roadmap & Strategy, Deliverable D2.1, Characteristics, capabilities, potential applications of Cyber-Physical Systems: a preliminary analysis, November 15, 2013, www.cyphers.eu.
9. ARTEMIS-IA, https://artemis-ia.eu/projects-1.html
10. IEC-61499, International Standard IEC-61499, Function Blocks, Part 1 - Part 4, International Electrotechnical Commission (IEC), Technical Committee TC65/WG6, IEC Press, Jan., 2005.
11. IEC-61512, Batch control, part 1: Model and terminology, The International Electrotechnical Commission (IEC), Final draft, 1997.
12. Frey G., Design and formal analysis of Petri net based logic control algorithms, Dissertation, University of Kaiserslautern, Shaker-Verlag, Aachen, 2002.
13. Kim K.-D., Kumar P. R., Cyber-Physical Systems: A Perspective at the Centennial, Proceedings of the IEEE 100 (Centennial-Issue): 1287-1308 (2012).
14. Bordel B., Alcarria R., Robles T., Martín D., Cyber–physical systems: Extending pervasive sensing from control theory to the Internet of Things, Pervasive and Mobile Computing, Volume 40, September, 2017, pp.156-184.
15. Santos C., Espinosa F., Santiso E., Mazo M., Xia F., Aperiodic Linear Networked Control Considering Variable Channel Delays: Application to Robots Coordination, Sensors (Basel), Vol. 15(6), pp.12454–12473, 2015.
16. Sanfelice R. G., Analysis and Design of Cyber-Physical Systems: A Hybrid Control Systems Approach, Cyber Physical Systems: From Theory to Practice, CRC Press, pp.3-31 (2015).
17. Graham S., Baliga G., Kumar P. R., Abstractions, Architecture, Mechanisms, and a Middleware for Networked Control, IEEE Transactions on automatic control, Vol. 54, No. 7, July 2009
18. Y. Sun, G. Yang, X.-S. Zhou, A survey on run-time supporting platforms for cyber physical systems, Frontiers Information Technologies & Electronic Engineering 2017 18(10):1458-1478.
19. Mangharam R., Pajic M., Distributed Control for Cyber-Physical Systems, Journal of the Indian Institute of Science, Special Issue on Cyber Physical Systems, September 2013.
20. Kent, S. (2002). Model Driven Engineering. In Proceedings of IFM2002, LNCS 2335, Springer.
21. Holobloc Inc. (2011), Function Block Development Kit (FBDK). available: http://www.holobloc.org/
22. 4DIAC (2011), Framework for Distributed Industrial Automation (4DIAC) Available: http://www.fordiac.org
23. FBench. (2011), Open Source Function Block Engineering Tool., available: http://oooneida-fbench.sourceforge.net/
24. ICSTriplex. (2011), ISaGRAF v. 5.0. available: www.isagraf.com
25. NxtControl, (2011), nxtStudio, available: www.nxtcontrol.com
26. Dai W., Pang C., Vyatkin V., Christensen J. H., Time-Stamped Event based Execution Semantics for Industrial Cyber-Physical Systems, 13th IEEE International Conference on Industrial Informatics (INDIN), Cambridge, UK, 2015.

ANALYZING EMOTIONS FROM TEXT CORPUS USING WORD SPACE

Mohsin Manshad Abbasi, Professor Anatoly P. Beltiukov

Department of Theoretical Foundations of Computer Sciences
Udmurt State University
Izhevsk, Russian Federation
e-mail: mohsinmanshad@gmail.com, belt.udsu@mail.ru

***Abstract**: Emotion identification and analysis is a topic of growing interest. It is because of abundant amount of data available online and offline in different languages. In this paper, we propose a new methodology called Word space. In this methodology, we use the frequency of words to analyze the text corpus before its classification. Word space is the concept derived from Metric Space, which is defined as a set for which the distance between all of its members, is defined. Those distances are when taken together called a metric on the set. In Word space distances between the words and their occurrences are measures. The emotion carry words with relatively high frequency and less distance between their occurrences are strong emotions. Whereas less frequent emotions that occurs far from, each other are weak emotions in the text. Using this concept, we will analyse the document, categorize and can summarize its emotions. In the conclusion section, we explain the interesting results that we observed using this methodology.*

1. Introduction

Emotion Analysis is one of the most pursued research topics in recent times. Many researchers and companies have explored the area of opinion identification and its analysis. With the increase in the number of Internet users, there is a proliferation of opinions available on the web. This phenomenon has a huge impact on various applications such as product review summarization and the public opinion monitoring systems.

There is a need and possibility for parallel and distributed computing use in the sphere of NLP (Natural Language Processing) [1]. Various techniques and methodologies are used for the identification of Emotions from text. Word space is relatively a modern concept based on the concepts of Metric Space in Mathematics. In Word Space, the distance between the first and the next occurrence of a word is measured. Each word and its occurrences are save using a vector. In our research, we will use the similar concept for words that express emotions using the exemplary text and equations explained below.

A lot of people are happy and satisfied with the service of online shopping sites and it is because of they are secure and safe. Online shopping sites are selling good products therefore the people are happy and satisfied with them. They have been able to build trust among the different customer. It is the simplest and easiest way to shopping.

Though number of online shoppers is increasing, but still there is one biggest pitfall adding to its limitation factor, that is the risk factor it entails. Starting from credit card scams to fake product deliveries, we have clearly heard of online shopping going wrong in many ways. The risk factor is also on its peak. The products are fake and the website are not secure increases the risk to buy online. Still most of recent surveys on internet shows that people are highly satisfied buying the essential things online. The reason behind the popularity of online shopping is: it is easy and fast. It's the most simplest way of shopping from sitting at home or wherever you are and therefore, customers are happy and enjoying new shopping experience these days. It seems like online shopping stores are cheap and more efficient than the local stores we visit. Online shopping sites play significant role making human life easier. It's safe if you care some important things

Blog taken from https://www.quora.com/Is-online-shopping-safe-or-not

The text is taken from an online blog where the people express their opinion about online shopping. The words expressing positive emotion are highlighted with green marker whereas the negative are with red marker. Some of those words occurred more frequently while other are less frequent. In our work, we will analyze the similar blog expressing opinion of people resulting in big data. To analyze the data we are using some equation and formulas as explained below.

In first equation, we have C_1 that represent the number of occurrence of a word expressing an emotion in text. For every word expressing an emotion in text, the number of its occurrence counted resulting in an array C carrying the occurrence count of all emotion-carrying words from the text. Similarly F_1 represents the frequency of occurrence of a word by dividing its word count C_1 with the total number of words T_m in the text and on applying it to all the members of Word count array C, we have a new array F represent the frequency of all the emotion carrying words in the text.

$$C_1 = \text{Count (Word Occurrence)} \quad (1)$$

$$C = [C_1\ C_2\ C_3\ C_4 \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots C_n] \quad (2)$$

$$F_1 = \left(\frac{C_1}{T_m} \times 100\right)\% \quad (3)$$

$$F = [F_1\ F_2\ F_3\ F_4 \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots F_n] \quad (4)$$

The primary objective of this study is to analyze emotions from text. The emotion carrying words and their occurrences extracted in the form of a vector. This vector values are then use to identify the frequency of occurrence of the words expressing emotions in the text. On the basis of these calculations, the emotions expressed in the text can be categorize as frequent, less frequent, most frequent and least frequent emotion. It is also observed that while extracting the words carrying emotions, there should not be any negation such as may not, cannot, don't before the emotion carrying word. If it occurs then the word polarity will be change. Positive emotion will become negative and negative emotion will become positive emotion.

Table 1 Describes emotion reverses on applying negations

Operator	**Original Emotion**	**Resulting Emotion**
Negation	Positive	Negative
Negation	Negative	Positive

2. Applications

The analysis of text corpus based on emotions can be used for summarizing the text [2]. We can identify important emotions that occur more frequently in the text and are more similar to the words mentioned in the title of the text. For this purpose, two properties of the words are considered. The Local property and the Global property. Local property is the frequency of occurrence of the word in the text while Global property is the maximum semantic similarity between a word and the title of Text [3]. Using this method the words carrying important information or emotions will not be missed while summarizing the text.

It can be used to identify the occurrence of the words and to avoid its repetition. It has been observed that the occurrence of same words expressing emotions frequently results in losing the interest of the reader in the text. For making the sentences of good quality and attract the attention of the readers, the identification of word repetitions is very important. The similar idea was used on twitter data written in Indonesian Language. The idea was to remove the repeatedly characters and words so that the data to be processed in accordance with the needs [4].

It can be used to identify the sentence similarity measure for paraphrase identification. It is more difficult to distinguish between precise and loose paraphrases than between loose paraphrases and non-paraphrases [5].

It can be very helpful for making the prediction of the next occurrence of an emotion in the text based on calculations of previous occurrence of a word in the text. A similar methodology was applied for text-based emotion prediction problem using supervised machine learning. The experiment was performed on children's fairy tales and the results were convincing [6].

It can be used for approximate string matching. Such as online search engines that takes a query from the user and match each word of this query with the database of text available online. However, this matching is based on word to word matching without taking in consideration the emotions in the text. In future, the emotion expressing words can get the priority in search engines for performing the text matching based on emotions.

3. History of Emotion Analysis

The development of the General Inquirer System (1966) (Stone, [7]) by Philip Stones in Harvard was probably the first milestone to identify textual emotions. The system usually counts the positive or negative emotion instances.

After this, a lot of work has been done for identification of emotions from text in different languages. An important among them was the contributions of Jaynce Wiebe, Peter Turney and Vasileios Hatzivassiloglou during early 90's. Jaynce Wiebe in 1990 (Wiebe, [8]) defines the term "Subjectivity" for Information Retrieval research. Later on in the year of 1997 (Hatzivassiloglou et. al., [9]) identified the semantic orientation of adjectives. After a few years Peter Turney (2002, [10]) came up with his revolutionary approach of Thumbs Up and Thumbs Down for positive and negative review classification. Pang (et. al., 2002, [11]) has suggested the building of sentiment lexicon manually for a domain. (Denecke, 2009, [12]) reported an interesting study on multiple domains to demonstrate the usefulness of the prior polarity scores from the SentiWordNet. Clustering had been observed as a technique based on generalizations of graph partitioning that don't require pre specified ad hoc distance functions and is capable of automatically discovering document similarities or associations [13].

In Russia studies devoted to sentiment analysis in Russian before 2011 are not very numerous. In (Ermakov, 2009, [14]) a sentiment analysis system extracting opinions about cars from a Russian blog community is presented. The patterns are language-dependent and domain dependent, which means that patterns must capture the lexical, syntactic and stylistic features of the analyzed text. It is not possible to directly translate or map the English pattern base into Russian pattern [15].

Emotion analysis from text in Russian language appears mainly in multilingual experiments. Zagibalov in (Zagibalov et al., 2010, [16]) compare corpora of reviews related to the same books in English and Russian. In (Steinberger et al., 2011, [17]), construction of general emotion vocabularies for several languages is described. Chetviorkin and Loukachevitch (2012, [18]) described, the generation of the Russian sentiments and emotional vocabulary for the generalized domain of products and services and so on.

In bilingual experiment on Russian and Romanian languages, it is observed that the word spelling can be considered as a word phonetic equivalent. This feature allowed limiting the search to letter-based representations [19]. For emotion identification and analysis in Russian language, most of the researchers are using Rule based techniques which is quite necessary as the rules are key to develop a grammar that can be used to develop stemmer and then semantic and syntax analyzer [20]. It has been observe that the methods of machine learning performs better in text categorization and classification.

4. Related work

In 1977 (Halliday and Hasan, [21]) classified lexical cohesion into two categories: reiteration category and collocation category. Reiteration category considers repetition, synonym, and hyponyms, while collocation category deals with the co-occurrence between words in text document. After this, the major effort was the work on the surprising behavior of Distance Metrics in High Dimensional Space [22]. In this research, it was explained that the fractional distance metric provides more meaningful results both from the theoretical and empirical perspective.

The main factors that affect the efficiency when searching metric spaces are intrinsic dimensionality of the space and the search radius [23].

(Edgar C. et al. 2001 [24]) in their work present a technique to perform search in Spaces using the distance. They proposed a concept for working with general cases to model similarity with a distance function that satisfies the triangle inequality, and the set of objects.

Kruengkrai and Jaruskululchi try to determine text title [25]. Their approach takes advantages of both the local and global properties of sentences. They used clusters of significant words within each sentence to calculate the local property of sentence and relations of all sentences in document to determine global property of text document.

In 2012, Maryam Kiabod et al. [26] present an approach for summarizing the text by identifying significant words from the Text. They used the extractive method to select the subset of the sentences that contains main concept of the text. The algorithms for preprocessing the text before analyzing the emotions improves the efficiency and effectiveness of the complete process [27].

5. Methodology

In this method, the list of words carrying emotions are extracted as shown in equation below. W represent list of all words expressing emotions. D1 represents the average distance between the different occurrences of emotion carrying word W_1. On applying it to all words carrying emotion we have a list D that represents the average distances of all emotion expressing words. The strength S_1 of a word is directly proportional to its word count C_1 and inversely proportional to the average distance D_1.

$$W = [W_1\ W_2\ W_3\ W_4 \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots W_n] \quad (5)$$

$$D_1 = \frac{1}{N}\sum_{n=1}^{n}[W_1 D_n] \quad (6)$$

$$D = [D_1\ D_2\ D_3\ D_4 \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots D_n] \quad (7)$$

$$S_1 \propto C_1 \quad (8)$$

$$S_1 \propto \frac{1}{D_1} \quad (9)$$

Before calculating the average distance occurrences, the maximum and minimum distance values were identified and then normalized. For this purpose the maximum value is divided by the total number of occurrences of that word and the minimum value is multiplied.

$$Norm = \max(D)\ ,\ min\ (D) \quad (10)$$

$$Norm = \frac{\max(D)}{C}\ ,\ \min(D) \times C \quad (11)$$

The strength of an emotion is the total occurrence of it Ci divided by the average distance of occurrence D and multiplied by the total number of emotions in the textT_E. On calculating the strength, average distance and occurrence, the strong, very strong and weak emotions can be identified and the text corpus can be categorized as positive or negative.

$$S(W) = \frac{\sum_{i=1}^{n} Ci}{D} \times T_E \quad (12)$$

6. Results and Discussion

The data obtained from the blogging corpus for the purpose of experiment showing the comments of people from October 1, 2013 to September 23, 2016. More than 25 people showed their opinion using around 6000 words. They used a total of 78 emotion oriented words. Among them, 40 words are representing positive emotions, 32 are negative and 6 with inversed polarity. The ten most frequent positive and negative emotions are shown in the tables below.

Table 2 *Show most frequent positive emotions in text*

Emotion	Total Occurrence	AVG Distance	Percentage
Safe	29	66	14.5
Secure	16	81	8.04
Good	12	223	6.03
Trust	9	77.3	4.52
Easiness	6	67.6	3.01

Table 3 *Show most frequent positive emotions in text*

Emotion	Total Occurrence	AVG Distance	Percentage
Hack	7	45.7	3.51
Trouble	6	95.8	3.01
Steal	6	52.6	3.01
Unsafe	2	27.5	1.005
Bad	3	359	1.50

For analysis, an online blog on the topic, "is online shopping safe or not" has been used. The five most frequent positive and negative emotions in the text corpus are discussed here. The most frequent positive emotion is safe. It occurs almost after every 66 words and has a high percentage of 14.5. It is probably because of the reason that it is the part of the topic and most of respondents try to answer it with the same word. Then the words secure and good are the emotion that show the user feels secure and the quality of products they buy are good. They also trust and feel easiness in doing shopping online.

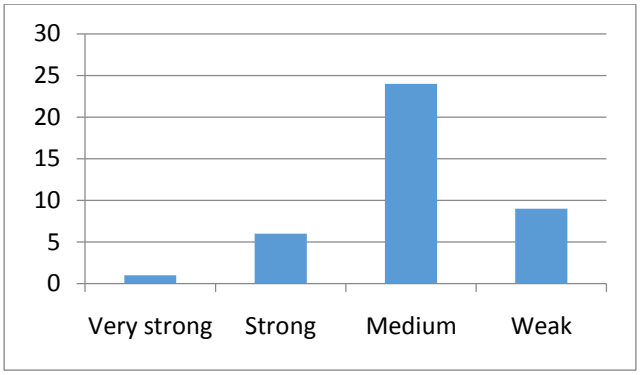
On the other hand, the most frequent negative emotion presented by users is Hack. It is because they are scared of their credit card information theft online. This emotion is shown on average after every 45 words which can be considered has an important and strong emotion. Then the second frequent negative emotion is steal. It is related with the previous emotion that describes the security threats while doing shopping online. People can steal the personal information of the buyer. This will create trouble for the buyer and the shopping can be unsafe.

It has been observed that in general the people response to online purchase is very satisfying and positive. This result is concluded by observing the positive emotion, their high frequency, average distance and the percentage of occurrence. The date of writing the blog is also very important. People now trust more online shopping than before. It is because of the modern technology. However, they are still concern about technology challenges in the form of hacking and steal.

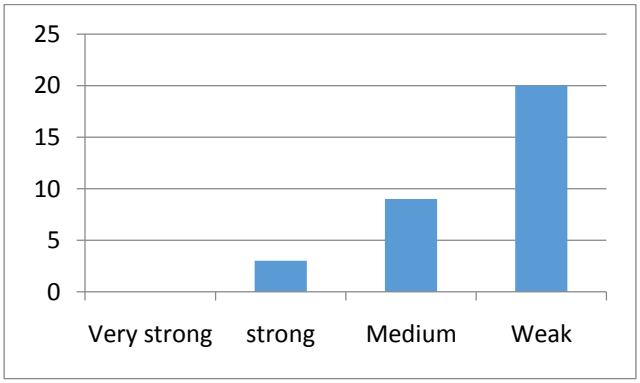
Table 4 *Comparison of positive and negative emotions from text*

Emotion Type	Total	Very Strong	Strong	Medium	Weak
Pos+	40	1	6	24	9
Neg-	32	0	3	9	20

The above table describes the text contains more positive emotion than negative emotions. The positive emotions are more strong whereas then negative emotions are mostly weak emotions.



Graph 1 *Representing Categories of Positive Emotion*



Graph 2 *Representing Categories of Negative Emotions*

The graphs are representing different strengths of positive and negative emotions. It is observable that negative emotions are mostly weak emotions with an average occurrence of less than 3 % in the text whereas the positive emotion are mostly medium strength emotions along with strong and even very strong emotions. It elaborates that text is representing positive emotions and is positive text.

7. Future Work

In future, we will apply our methodology on social media and tweets. The size of the data will be increased. The results obtained from this approach will be applied for summarizing the text based on the emotions represented in the text. We will observe other advantages of the methodology for sentiment analysis.

8. References

1. Solovyev V, Polyakov V, Ivanov V, Anisimov I, Ponomarev A. "An approach to semantic natural language processing of Russian texts". *Research in Computing Science*, 2013, pp. 65-73.
2. Ashmita S, Ruhi B. "Auto Text Summarization with Categorization and Sentiment Analysis". *International Journal of Computer Applications*, Vol. 130. November 2015, pp. 07.
3. Maryam K, Mohammad N. D, Sayed M. S. "A Novel Method of Significant Words Identification in Text Summarization". *Journal of Emerging Technologies in Web Intelligence*, Vol. 4. August 2012.
4. Fachrian A, Arif D. "Improving the Performance of Repeated Character Preprocessing in Recognizing Words in the Indonesian Sentiment Classification". *Journal of Basic and Applied Scientific Research*, 2017, pp.1-9.
5. Pronoza E, Yagunova E. "Comparison of sentence similarity measures for Russian paraphrase identification", In: *Artificial Intelligence and Natural Language and Information Extraction, Social Media and Web Search FRUCT Conference (AINL-ISMW FRUCT)*, 2015, pp.74-82.
6. Cecilia O.A, Dan R, Richard S. "Emotions from text: machine learning for text-based emotion prediction". In: *Proc. of the 5th conference on Human Language Technology and Empirical Methods in Natural Language Processing*, Canada, 06-08 October 2005, pp. 576-586.
7. Stone, P.J, Dunphy, D.C, Smith, M.S. "The General Inquirer: A Computer Approach to Content Analysis". MIT Press - Cambridge, 1966.
8. Wiebe, Janyce M. "Identifying Subjectivity characters in Narrative". In: *Proc. of the 13th International Conference on Computational Linguistics*, Helsinki Morristown NJ: Association of Computational Linguistic, 1990, pp. 401-408.
9. Vasileios H., Kathleen R.M. "Predicting the Semantic Orientation of Adjectives" *EACL '97* In: *Proc. of the eighth conference on European chapter of the Association for Computational Linguistics*, Madrid, Spain, July 07-12, 1997.
10. Peter D.T. "Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews". *In: Proc. of the 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, 2002, pp. 417-424.
11. Pang B, Lee L. "Thumbs up? Sentiment Classification using Machine Learning Techniques" *In: Proc. of the Conference on Empirical Methods in Natural Language*, Philadelphia, 2002, pp.79-86.
12. Denecke K. "Are SentiWordNet scores suited for multi-domain sentiment classification?". *In: Proc. Of the fourth International Conference on Digital Information Management*, University of. Michigan, USA, 1-4 November 2009.
13. D. Boley, M. Gini, R. Gross. "Partition based clustering for web document categorization". *Elsevier Journal for Decision Support Systems*, Vol. 27, Issue 3, December 1999, pp. 329-341.
14. Ermakov A. "Knowledge extraction from text and its processing: Current state and prospects". In*: Proc.: of the Computational Linguistics and Intellectual Technologies*, 2009. pp. 50–55.
15. Pivovarova L, Du M, Yangarber R. "Adapting the PULS event extraction framework to analyze Russian text". In: *Proc. of the 4th Biennial International Workshop on Balto-Slavic Natural Language Processing*, Sofia, Bulgaria, 8-9 August 2013, pp. 100–109.
16. Zagibalov, T , Belyatskaya et al. "Comparable English-Russian Book Review Corpora for Sentiment Analysis" , Edition 2010.
17. Steinberger J, Lenkova P, Kabadjov M., et al. "Multilingual Entity-Centered Sentiment Analysis Evaluated by Parallel Corpora", In: *Proc. of Recent Advances in Natural Language Processing*, Bulgaria, 12-14 September 2011, pp. 770–775.
18. Chetviorkin I, Braslavskiy P, Loukachevich N. " Sentiment Analysis Track at ROMIP 2011" Computational Linguistics and Intellectual Technologies: *In: Proc. of the International Conference* (Dialog 2012), Bekasovo, pp. 1–14.
19. Sokolova M, Bobicev V. "Classification of emotion words in Russian and Romanian languages". *International Conference RANLP*, Borovets, Bulgaria 2009, pp. 416–420.
20. Abbasi M.A, Beltiukov A.P. "Analysis of sentiment and emotion from text written in Russian language*" 5th All Russian Conference on Information technology for intelligent decision making support (ITIDS)*, Ufa, Russian Federation, May 16-19, 2017.
21. Halliday M, Hasan R. "Cohesion in English". London: Longman, 1975.
22. Aggarwal C.C, Hinneburg A, Keim D.A. "On the surprising behavior of distance metrics in high dimensional space". *International Conference on Database Theory*, 2001, pp. 420-434.
23. Cha E, Navarro G, Baezayates R, Jose Y, Marroquin L. "Searching in metric spaces" *ACM Computing Surveys*, Vol. 33, Issue 3, September 2001, pp. 273–321.
24. Edgar C, Gonzalo N, Richardo B.Y, Jose L.M. "Searching in Metric Spaces". *ACM Computing Surveys*, Vol. 33, No. 3, September 2001, pp. 273–321.
25. Jaruskululchi C, Kruengkrai. "Generic text summarization using local and global properties of sentences". *IEEE/WIC international conference on web intelligence*, October 2003, pp.13-16.
26. Kiabod M, Naderi M, Sharafi S.M. "A novel method of significant words identification in text summarization". *Journal of Emerging Technologies in Web Intelligence*, Vol. 4, Issue 3, August 2012.
27. Abbasi M.A, Beltiukov A.P. "Механизм предварительной обработки текста перед анализом настроений*". 6th All Russian Conference on Information technology for intelligent decision making support (ITIDS)*, Ufa, Russian Federation, May 28-31, 2018.

METHOD OF STRUCTURING THE SELF-ORGANIZED INTELLECTUAL SYSTEM ON THE BASIS OF REQUIREMENTS OF THE ISO/IEC 15288 STANDARD IN THE FORM OF THE CARTESIAN CLOSED CATEGORY. (ON THE EXAMPLE OF DESIGN OF INFORMATION AND ANALYTICAL SYSTEM)

G.G. Kulikov, V. V. Antonov, A.R . Fakhrullina, L.E. Rodionova

Faculty of Informatics of Robotics, Ufa State Aviation Technical University, Ufa, Russia
e-mail: gennadyg_98@yahoo.com, antonov.V@bashkortostan, almirafax@mail.ru, lurik@mail.ru

Abstract *In article approach to formation of structure of the self-organized system is considered. The method of transition to the quantitative multiple description of information processes and application of requirements of the ISO/IEC 15288 standard is shown. The rule of synthesis of the self-organized structure of models of process is offered. The possibility of parametrical control of a condition of processes is shown.*

1. Introduction

Today, a particular interest is given to analytical information systems with elements of intellectual management, capable to generate not only strictly regulated reports, but also inquiries upon the demand of the persons, making decisions (P.M.D).

Таким образом, становится актуальным проектирование аналитических информационных систем, которые могут быть использованы большими компаниями, решающие задачи управления персоналом, кадровым резервом, талантами, показатели, которых, являются критериями эффективности.

Thus, there is relevant a design of analytical information systems which can be used by the big companies, the solving human resource management problems, a personnel pool, talents, indicators which, are criteria of efficiency.

Compliance of traceability of information objects, traceability of real objects, parameters for the knowledge base and design of information and analytical system. At the abstract level of the studied educational production systems we will apply to definition of real objects and compliance to their information objects and traceability concepts and rules of the ISO/IEC 15288 standard "System engineering". We will note that the concept of this standard completely answers concepts to rules Cartesian the closed system.

The main objective of design of information and analytical system at decomposition of abstract level naturally consists in ensuring isolation at the lower levels. It will correspond to laws according to logic of a statement of Aristotle and formal logic of the first order. In this case, logical contradictions at the solution of problems of identifiability and traceability of information objects will be excluded. It is system model of formal system.

$$FS =< M_n, P(M_n), O >$$, where

FS – formal system;

M_n – set of carriers;

P – predicates on M_n;

O – set of operations (algebra).

If O – the empty set, formal system degenerates system model (SM). If $P(M_n)$ is an empty set, the algebra turns out. Predicates are the statements expressed by sizes the truth or a lie. If predicates aren't certain, then they can't be expressed.

We will understand the organized system including a set of components as system model: in structural, information, semantic, linguistic, mathematical and other representations [1].

If the predicate set is defined by logic of a statement or formal logic of the first order if it Cartesian is closed, then this model will be consistent on structure, but it is consistent concerning logic of the first order since it besides formal logic goes the logician of the second order or the logician Hegel. Here identifiers of the first order during time don't change. And at logic of the second order information objects can change.

Processes of life cycle of the ISO/IEC 15288 standard of "System engineering" such as, processes of the enterprise, processes of the project and technical processes help effective formation and use of systems, promoting achievement of the goals of the enterprise. And processes of the agreement represent working relationship, by the conclusion of the agreements which are of the greatest interest to us [3].

Use of this standard will allow to provide compatibility various a component among themselves (components – objects of the real world which are a part of PS).

ISO/IEC 15288 "The system engineering" represents a stock of knowledge and the integrated system of thinking about intellectual content, allowing to create a technical basis for combination of requirements of the market [1].

For structuring a program system, we will consider stages of ZhTs of a program system, having described each stage as the separate module of the program, taking into account requirements of the standard ISO/IEC 15288:

1. Concept (plan);
2. Development;
3. Manufacture;
4. Application;
5. Application support;
6. Phase-out and write-off.

We will consider this theory on the example of management of a personnel pool.

We will consider the main business process from the point of view of system approach, having applied idef0 methodology, to further formation of model of a personnel pool [4].

Methodology – set of the methods which are eaten around by common goals and tasks.

This model submits a set of hierarchical charts, each of which describes function and is divided into subfunctions. The system can be presented in the form of a set of processes And, carrying out transformation of elements of system (entrances, exits, management, mechanisms of management) (Fig 1).

Present I_i – set $\{j_1..., j_K\}$ process entrances A_j, O_i – set $\{o_1..., o_\kappa\}$ process exits A_j, C_i – set $\{c_1..., c_N\}$ managements of process A_j, M_i – set $\{m_1..., m_P\}$ mechanisms of management of processes A_j.

Therefore, process can be presented in the set form:

$$Aj = < I_i, O_i, C_i, M_i >.$$

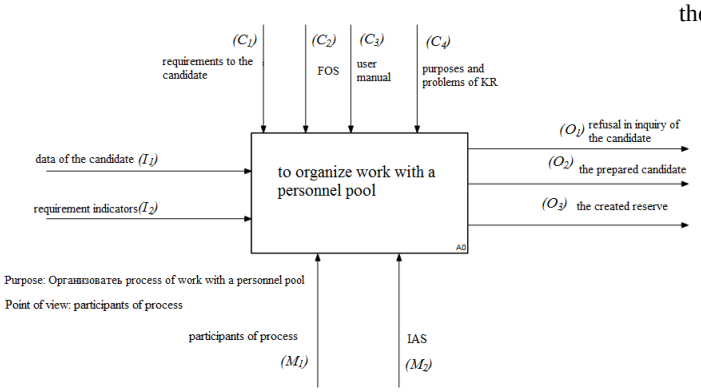


Fig 1. *– Contextual chart of process of work with a personnel pool*

With use of set-theoretic modeling, this process it is possible to present sets in the form A_0= {A_1, A_2, A_3, A_4} subprocesses [5].

We will consider separately each process of work with a personnel pool:

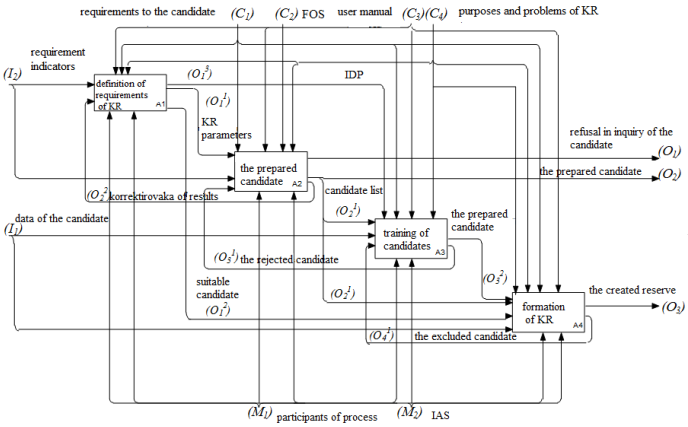


Fig 2. – Decomposition of the contextual chart of process of work with a personnel pool

A_1 – Definition of needs for a personnel pool. This process is characterized in definition of requirement of number of reservists and official structure: internal resources of the enterprise are analyzed. Upon termination of decisions are made: to attract candidates "from the outside", to partially retrain the working experts, to reduce personnel, the suitable candidate (Fig 2.).

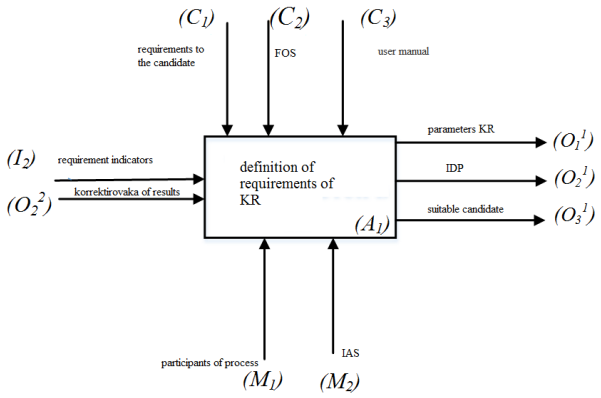


Fig 3. *– Process of definition of needs for KR*

With use of set-theoretic modeling, this process can be presented in the form (Fig 3.**)**:

$$A_1 = (I_2, O_2^2, C_1, C_2, C_3, M_1, M_2, O_1^1, O_2^1, O_3^1)$$

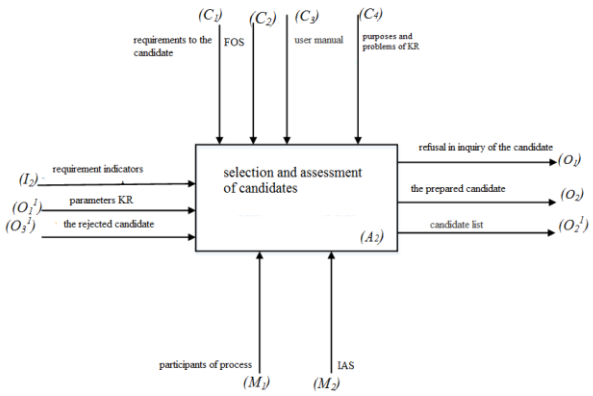


Fig 4. *– Process of selection and assessment of candidates*

A_2 – Selection and assessment of candidates. This process represents, selection of candidates for different types of a reserve and also their assessment for requirements, the purposes and problems of a personnel pool of the enterprise (Fig 4).

This process can be presented in the form:

$$A_2 = (I_2, O_1^1, O_3^1, C_1, C_2, C_3, C_4, M1, M_2, O_1, O_2, O_2^1)$$

A_3 – Training of candidates. After selection and assessment of candidates is carried out, the individual development plan (IDP) which has included training of candidates is formed. Following the results of training the analysis of efficiency of training of candidates is carried out and the decision on inclusion in a reserve, or carrying out additional training is made (Fig 5).

This process can be presented in the form:

$$A_3 = (I_1, O_2^1, O_4^1, O_1^3, C_1, C_3, C_4, M1, M_2, O_3^1, O_3^2)$$

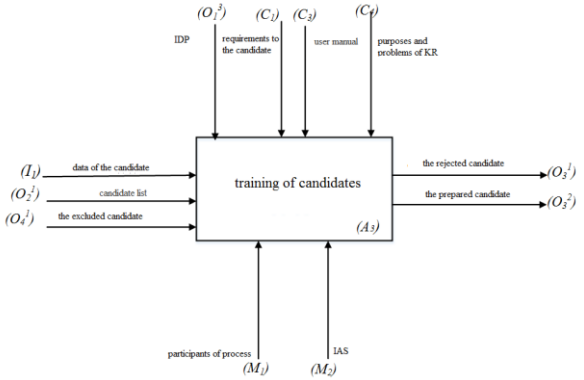


Fig 5. *– Training of candidates*

A_4 – Formation of a personnel pool. When forming a personnel pool in an information system the following operations are performed:

– coordination of the list of candidates for a personnel pool;
– inclusion of employees in a personnel pool;
– movement of the employees consisting in a reserve of other position in a reserve;
– exception of employees of a personnel pool;
– approval of structure of a personnel pool (Fig 6).

This process can be presented in the form:

$$A_4 = (I_1, O_2^1, O_1^2, O_3, O^1_4, C_1, C_2, C_3, C_4, M1, M_2)$$

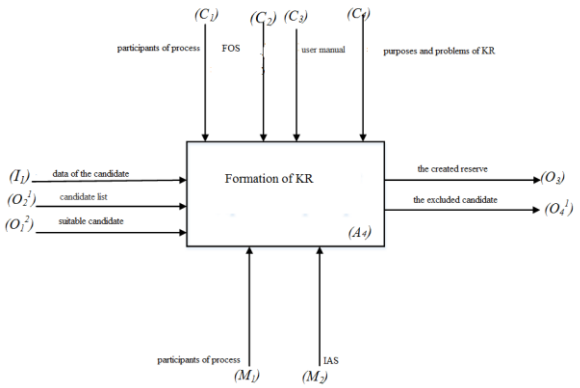


Fig 6. *– Formation of a personnel pool*

This process consists of a set $A_0 = \{A_1, \ldots, A_4\}$ подпроцессов.

With use of set-theoretic modeling, this process can be presented in the form:

$$A_1 = (I_2, O_2^2, C_1, C_2, C_3, M1, M_2, O_1^1, O_2^1, O_3^1);$$

$$A_2 = (I_2, O_1^1, O_3^1, C_1, C_2, C_3, C_4, M1, M_2, O_1, O_2, O_2^1);$$
$$A_3 = (I_1, O_2^1, O_4^1, O_1^3, C_1, C_3, C_4, M1, M_2, O_3^1, O_3^2);$$
$$A_4 = (I_1, O_2^1, O_1^2, O_3, O^1_4, C_1, C_2, C_3, C_4, M1, M_2)$$

For this purpose we will determine external parameters and modules which will be rather independent. As a result, if it is necessary to replace some stage, we replace only this module. Considering the relations which can be formalized, i.e. to exclude these relations from the text of the module, and to use them as the knowledge base i.e. as the table on which he will work.

Communication between these modules, can be defined according to the ISO/IEC 15288 standard "System engineering", namely in the form of categories of sets.

At design of information and analytical system, the theory of categories allows to describe the relations of objects in a program system, having left for later the description of properties of objects. eventually it allows to exclude the semantic description from a program code, entering elements of formalization and writing down in the form of rules in knowledge bases. Considering that IAS modules can be used as objects, it allows to construct adaptive system which will be considered as universal to change of external data.

Such mechanism allows to study subject domain on the other hand and to develop a compact and universal program system. Since the program works on the set algorithm and it can be changed, adding only other parameters to KB.

For design of the IS database the object and relational database management system with the open PostgreSQL code was used. Unlike other databases, for example, of MySQL, MariaDB and Firebird which aren't object and relational, PostgreSQL provide more powerful functionality. The choice of a DB is caused by presence of the following advantages at her, PostgreSQL:

- the reliable, integrated and scalable DBMS;
- it is started on all main platforms, including Windows;
- supports a set of types of data;
- provides the expanded capacity of data;
- have API for C/C++, Java, Perl, Python, Ruby and others;
- being DBMS of a class of the enterprise, provides such features as restoration on a point in time, preservation points, anticipatory journalizing on a breakage case;
- can process many data;
- careful attitude to integrity of data.

Thus, application of the given method allows to use PostgreSQL not only as the powerful database management system allowing to provide activity of the organization, but also as the platform of applications programming, demanding use of relational DBMS.

Conclusion

By the conducted researches it is established that this method can be used for many business processes of the organizations, and the parameters determined by this method form category, can be reduced in a multidimensional matrix (an analog of a matrix of J. Zakhman) and finally are used at the description of the specification on development or reengineering of IAS.

References

1. Kulikov G.G., Konev K.A., Suvorova V. A. Theory of systems and system analysis: The manual for students of all forms of education studying in the directions 080800 (230700) «Applied informatics», 230100 «Informatics and computer facilities». Education guidance – Ufa: UGATU, 2012 y. 159 p.

2. Kulikov G.G., Antonov V.V.., Shilina M.A., Fakhrullina A.R. Structuring content of the considered area for the further intellectual analysis. Example of formation of the structured content of OPS / Kulikov G.G., Antonov V.V., Shilina M.A., Fakhrullina A.R. // Magazine Informatsionno-upravlyayushchiye Sistemy magazine of Sankt-of the St. Petersburg state university of space instrument making № 2. 2016 –P. 95-100

3. IDEF0. [Электронный ресурс] // Access mode: https://ru.wikipedia.org

4. ISO / IEC / IEEE 15288: 2015

5. Antonov V.V. The IS set-theoretic model for the multidimensional analytical analysis meeting the requirements of storages of data [Text] / Antonov V.V., Kulikov G.G., Antonov D.V.// Bulletin UGATU: scientific magazine – 2012.- V. 16, № 6(51).-P. 189–201.

CONCEPTUAL MODEL "STAGE" AND INTELLECTUAL SUPPORT OF THE SOLUTION PROCESS OF "TACTICAL PROBLEMS" BY CREWS OF AIRCRAFTS

B.E. Fedunov, A.A. Plyatcovoi
Federal State Unitary Enterprise, «State Research Institute of Aviation Systems» Moscow, Russia
e-mail: boris_fed@gosniias.ru, e-mail: plyatcovoi.al@gmail.com

Abstract: *This paper is about the conceptual model "Stage" and its use in the development of on-board advisory expert systems of an operational goal setting (OAES-goal setting) and flight stages OAES-typical situation (TS).*

1. Introduction

In the last third of the twentieth century, DARPA of the USA announced the research work - "Pilot`s Assistant", which sets the task of developing operational recommendations to the crew on how to solve the task / problem on board of a combat aircraft. It was quickly discovered (in the USSR and in the USA) that such an "integrated system" would not be able to be created, that first it needs to work out another conceptual view of the aircraft (the conceptual model of the aircraft) and to determine the nomenclature of the systems supporting the crew decision process from this model's position, arising during the execution of the flight task. The description of the tasks and composition of such nomenclature were determined on the basis of the conceptual model "Stage" which is presented below.

2. Conceptual model "Stage" for the aircraft [1]

The model formalizes the flight process of the aircraft and classifies tasks that should be solved on its board.

The formalization of the flight process of the aircraft's flying activity includes a conceptual model of the flight process of an aircraft (a group of aircraft) (the upper part of Fig. 1).

The "Stage" model contains a set of assigned sessions of functioning (in the terms of reference (ToR) for the development of the anthropocentric object (A/O)) to the algorithmization (with the description of the general task of functioning (GTF) of each session); the presentation of each session through a semantic network of typical situations (or typical combat situations (TCSs)) and the presentation of each TS through a semantic network of problematic subsituations (PrS/Ss). Semantic networks are built on a cause-effect relationship.

As a rule, each GTF has attached expected external and intra-board threats which are described as a set of TS-threats.

The set of TS is finite and is constructed on the whole set of the functioning sessions (in the development of the aircraft).

It should be noted that each TS contains the ultimate goal of the TS and the way to achieve it, broken down by its PrS/S.

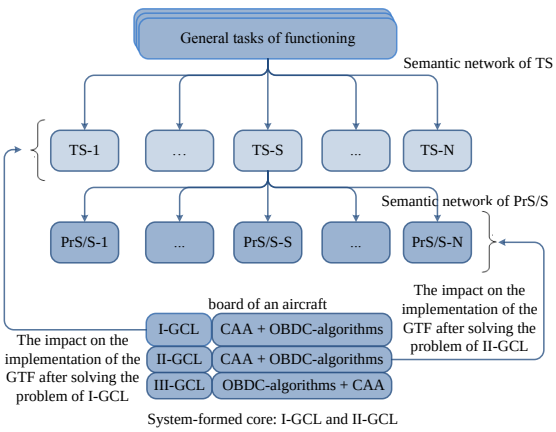


Fig.1. *The model of an aircraft "Stage" for the design of the on-board algorithmic and indication support (OA&IS) of the system-forming core of the aircraft*

Classification of aircraft tasks (the lower part of Fig. 1) includes the allocation on its board of three global control levels (GCLs), in each of them, the leading role of the crew activity algorithms (CAAs) or OBDC-algorithms programmed in on-board digital computers (OBDC) (OBDC-algorithms) are determined :

1. The first GCL is the level of operational goal-setting (I-GCL). In terms of the "Stage" model, the result of solving this problem is the assignment of the current TS. In Fig. 1, this is shown as a "feedback" from the I-GCL to the line of TS.

2. The second GCL is the level of constructing a rational method for achieving the TS (II-GCL, operatively assigned by I-GCL.) In terms of the "Stage" model, the II-GCL task is interpreted as the operational construction of a specific fragment of the semantic network of PrS/S of the assigned TS. in Fig. 1 this is shown as a "feedback" from II-GCL to the line of PrS/S of named TS;

3. The third GCL is the level of realization of the method of achieving the goal (III-GCL), designed on the II-GCL.

Tasks of the I-GCL and the II-GCL are solved in the system-forming core of the A/O. These tasks are usually called tactical.

In the "Stage" model it is established that the crew is necessarily involved in the solution of the tasks of the I-GCL and the II-GCL. In Figure 1, this is indicated by the abbreviation CAA. At the same time, the process of solving the problems of these GCLs by the crew is supported by the OBDC-algorithms. Full support of this process is possible only if on-board intelligent tactical level systems (OITLS) are deployed on board of the aircraft, which must solve the corresponding tasks and provide the crew with the recommended method for their solution. The implementation of recommendations by the onboard executive systems comes only after the crew has agreed to this. Because of this, such systems are classed as operationally advisory systems.

The "Stage" model classifies OITLS, highlighting:

- intelligent systems ((on-board intelligent systems) OIS_I-GCL), solving tasks of the I-GCL;
- intelligent systems (OIS_II-GCL), solving tasks of the II-GCL.

The model of the functioning of the aircraft group (Fig. 2). To develop an OA&IS for the operating of the aircraft group performing the same GTF, a three-level functional hierarchy of the group is adopted:

- a commander of the first level of control in the group (K-I). He's one in a group;
- the second-level commanders in the a group (K-II) subordinate to K-I. There are up to four of them in the a group;
- commanders of the third level of control in a group (K-III), subordinate to the commanders of K-II. Each K-II commander has up to four K-III commanders under control.

There are two information flows in the group: the flow of control (commands) from the top to down (KI⇒ KII⇒ KIII) and the flow of current information (the flow of notification) from the bottom to up (KI⇐KII⇐ KIII, KI⇐ KIII).

A group of aircraft before a session of functioning is preparing to perform GTF, receiving a priori information on the upcoming flight (flight assignment (FA)) to the board. During the flight, the number of the aircraft in the group may decrease, but in the remaining part the functional hierarchy of the group accepted before the session is operatively saved.

At each stage of the session, all the aircraft in the group are in the same TS.

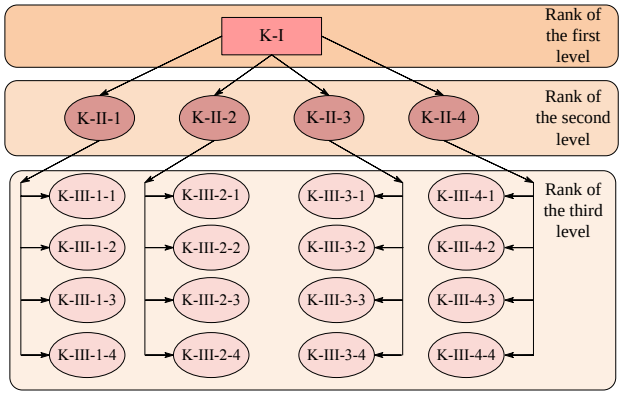


Fig.2. *The hierarchy in a group of aircraft*

Aircraft of the 4th generation in the development of OA&IS and crew`s guidance for their application were focused on the following conceptual model of the object.

In accordance with the ToR for the designing of the aircraft a number of possible episodes are identified, which can occur in flight. For each episode its OA&IS should be developed, regardless of the rest of the episodes. The concept of GTF in this model is absent. This conceptual model of an aircraft is called the conceptual model "Episode".

When using this model for the design of an OA&IS, the concepts of the GTF, flight stages and GCLs appear only in the aircraft application manual. Such episodes in the semantic network of PrS/Ss of the flight stage are shown to the crew "how to collect them".

However, even on the latest developments of the 4th generation aircraft, the designers found out the "Episode" model is inadequate, and there was a need to design an OA&IS using the new conceptual model of the aircraft, called the "Stage" model.

2.1. On-board intelligent tactical level systems

We will describe the composition and functions of the onboard intellectual systems that solve the problems of the system-forming core of the aircraft.

Computer intelligent systems to support the aircraft crew decision process of operational goal-setting (OIS_I-GCL)

The process of solving the task of operational goal setting by the crew is supported by the OIS_I-GCL: the Intelligent Information System "Situational Awareness of the Crew" (IIS-SAC) [3] and the on-board advisory expert system "Operational goal-setting" (OAES-goal setting) [4,5].

Both of these systems operate throughout the entire flight of the aircraft. They inform the crew about the emerged external and intra-board immediate threats (IIS-SAC) and recommend to the crew the current flight target (OAES-goal setting), taking into account the flight stage, the FA and the type of the threat that has arisen. We briefly describe these systems using information from [3 - 6].

The decision of the task by the crew of operational goal-setting includes three components: circumspection, situational awareness, situational confidence.

The crew's circumspection is supported by on-board algorithms (OBDC-algorithms) of integrated / complex information processing coming from on-board information systems, and OBDC-algorithms for allocating potential threats. Potential threats are presented to the crew on the information-control field (ICF) of the information frame of the information model of the external and intra-board situation, placed on the multifunctional display (MFD). An example of the presentation of the potential air threats on such model of the fighter Rafale, see [7].

Situational awareness of the crew is provided by the intellectual information system "Situational Awareness of the Crew".

The Intelligent Information System "Situational Awareness of the Crew" reveals among the observed potential threats to the implementation of the mission the immediate threats (ITs) and ranks them.

In its knowledge base (KB), which uses information from the OBDC-algorithm of the integrated processing of primary information and information from a number of on-board measuring systems, ITs are identified and their development are projected. Such a prognosis is used to determine a kind of "point / moment of non-return", when it will be impossible to effectively counteract the threat by existing resources onboard of the aircraft. Information about IT is presented to the crew on the ICF. Intelligent agents (threats, objects), mathematical models (MMs) of the collision "Threat - Object" with algorithms for calculating the "points of non-return" of these threats are included in the KB of the IIS-SAC. The threat that is most "close" to its "point of non-return" is declared a priority [3].

The ITs which are identified by the IIS-SAC are constantly presented to the crew on the ICF and through speech informants.

Situational confidence of the crew is provided by the on-board advisory expert system "Operational goal-setting" (OAES-goal setting).

The priority IT found in the IIS-SAC, which is "near" its "point of non-return" comes immediately to the OAES-goal setting. Among the other ITs, the OAES-goal setting will receive only a threat previously allocated by the crew.

The OAES-goal setting operates in three modes, each of which is activated under the following conditions [4,5].

Mode I is activated in the OAES-goal setting, when the aircraft is at the current stage of the implementation of the FA and the OAES-goal setting received an IT from the crew. In a number of cases, the OAES-goal setting can involve the crew to estimate the current situation.

In this mode, the previously positive experience of counteracting the identified threat is accumulated, formalized in its KB in the form of a knowledge matrix, is included in the OAES KB. Using the precedent conclusion in the KB, the priorities of possible solutions (precedents) of the crew are calculated [6]. The decision (selected TS) with the highest priority is recommended to the crew.

Mode II is activated in the OAES-goal setting, when the aircraft was in a situation of counteraction to the threat (TS-threat) and in it a signal "The Threat was Passed" was produced. OAES-goal setting involves the crew to estimate the current situation.

In this mode, in the KB of the OAES-goal setting is not always a reliable estimate of the current state of the aircraft and the state of its crew. Therefore, the OAES-goal setting requests this information from the crew and, in view of this, allocates those TS-flight stage, which the aircraft can still implement.

Mode III is activated in the OAES-goal setting when a threat occurs "near" its "point of non-return", received from the IIS-SAC. In this case, the IT enters the OAES-goal setting and the recommendations developed by it without the sanction of the crew activate the corresponding on-board intelligent tactical level system.

Computer intelligent systems to support the decision process of the aircraft crew for the task to design the way to achieve the operationally assigned current flight objective (OIS_II-GCL).

The on-board intelligent systems of the II-GCL include on-board advisory expert systems of typical flight situations (OAES-TS) [8].

An OAES-TS is one of the main components of the aircraft's flight intelligence (a new class of on-board algorithms) that operatively determine the rational methods of aircraft`s behavior in a typical flight situation. For each TS, a OAES-TS should be created. The conceptual model "Stage" is put in the structure of the KB of OAES-TS [2].

The "outside world" in which the OAES-TS will operate is the airborne information environment of the aircraft, formed by the output information of the on-board measuring systems, "standard" (not included in the knowledge base of the OAES-TS) OBDC-algorithms and signals from the ICF (Fig. 3).

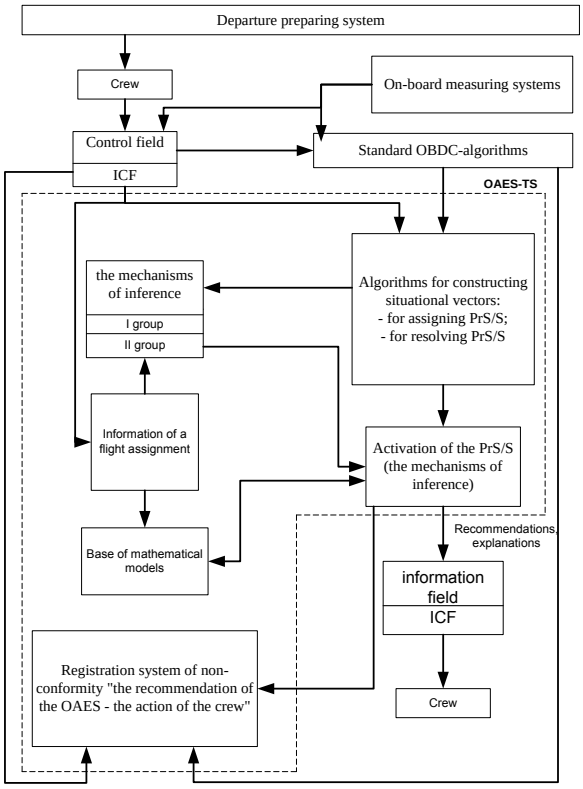


Fig. 3. *Functional diagram of the OAES-TS*

When preparing the aircraft / group of aircraft for a flight from the departure preparing system in the OAES-TS, a FA (a priori information) is loaded: a TS-FA network, a list of expected threats and a set of corresponding TS-threats.

In the process of performing the session for each significant PrS/S for performing the current TS, the OAES-TS prepares recommendations for the crew on its resolution with brief explanations. Recommendations and explanations to them appear on the ICF. The crew has the right not to accept the recommendation made by an OAES and resolve the arising PrS/S in a different way, without informing the OAES-TS about it. At the same time, the next recommendation of the OAES-TS will have to be worked out, taking into account the method implemented by the crew. Any ignoring by the crew of the constructed OAES recommendations is fixed in the on-board system of objective control (in Fig. 3 the block "Registration system of non-conformity ..."). After the end of the flight, this information is transferred to system of analyzing the results of the last flight.

The recommendations of the OAES-TS should be constantly coordinated with the activated model of the crew's behavior at the conceptual and operational levels. The crew perceives the outside world and the technical condition of the aircraft (the intra-board world) through their senses and the cabin information model of the external and intra-board situation on the ICF. Possessing professional training and having a priori information (the most general about the outside world and concrete about the upcoming session), the crew, within the framework of its activated model of behavior each time allocates the current problem (PrS/S) without informing the OAES-TS about it. The OAES-TS should give at the current time a valid and effective recommendation for this PrS/S for its resolution.

The presentation of the list of the main features of OAES-TS:

- it must solve all the problems of its TS (to be closed on TS problems);
- to have a limited dialogue with the crew (restrictions on the time limit issued to the crew by the external environment, and on the possibilities of input of information by the crew through the controls of the ICF);
- algorithms and rules in the KB should be guided by the structures of situational control algorithms;
- always be consistent with the activated conceptual model of the crew's behavior, developing recommendations for resolving the current problem at the level of a professional operator with sufficient significance for him,
- to have a "deferred" component of training.

At each stage of the flight, the crew, while solving the task of operational goal-setting (I-GCL), can activate the corresponding OAES-TS.

The KB of the OAES-TS consists of two hierarchical levels [8]. At the first level, according to the priori and current information, the corresponding PrS/S is activated on the base of production rules (Fig. 3). At the second level, the tasks of the activated PrS/S are solved and recommendations are presented to the crew.

We list the methods of solving problems used at this level (the mechanisms of inference): a multicriteria choice of an alternative, a conclusion based on a precedent, an optimization conclusion and production rules. In detail, the first two mechanisms are described in [6].

When implementing these mechanisms, mathematical models of the subject area fragments are used (Fig. 3 "Base of mathematical models").

2.2. Formalization of the general mission task "Escort of strike forces" in the interests of the commander of a group of escort fighters (an example of using the "Stage" model)

The general mission task (GMT) "Escort of strike forces (strike aircraft (SA))" is set to escort fighters (EF), which must directly accompany the SA to accomplish their mission.

The group of fighters is hierarchically ordered. The group is led by group commander K(EF)-I , which has subordinate commanders K(EF)-II subgroups.

Construction of the semantic network of typical situations in the GMT "Escort of strike forces" on the basis of the "Stage" model

When carrying out the formalization in the conceptual model "Stage" of the GMT the following TSs (flight stages) for commander K(EF)-I of the escort fighter group were identified:

- the take-off of group of the aircraft (TS "Take-off");
- the collection in air of EF and SA (TS "Collection-1"). The task of the commander of the EF group: to assemble the EF group, to lead the group to the point of meeting with the SA, to ensure the employment of EF places in the battle formations of the group;

- the flight of the EF group along the SA route (TS "Route-1"), the reflection of the hostile fighters (HF) attack by the whole EF group (TCS "Entering the group into the air battle" (TCS EGB-A)). The task of the commander of the EF group is to assign tactical reception of the HF reflection, to conduct tactical goal setting. TCS EGB-A usually occurs in the process of TS "Route-1", TS "Protection of the area" and TS "Route-2";
- collecting the EF group after completing the TCS EGB-A and meeting with SA (TS "Collection-2"). The task of the commander of the EF group: to ensure further support of the SA;
- the organization of the protection of the work area of the SA (TS "Protection of the area"). The task of the commander of the EF group is to prevent the enemy aircraft from attacking SA;
- the collection of the EF group and meeting with SA that have completed the attack of the targets (TS " Collection-3"). The task of the commander of the EF group is to ensure further support of the SA;
- the flight along the route with SA to the point of termination of their escort (TS "Route-2"). Task K(EF)-I: provide covering attack areas from which HF may appear; to repel an HF attack by a part of the EF group, to return this part of the group to the EF combat formations;
- the dissolution of the EF group and providing cover for the SA from HF attacks (TS "Dissolution of the Group"). Task K(EF)-I: establish order of priority of the landing, organize reflection of EF attacks.

The listed TS/TCS in the FA are presented by a semantic network in which the conditions for its occurrence are formulated for each TS/TCS (the result of a thorough study of the general mission task of fighters "Escort of strike forces", which is previously presented by a set of TSs/TCSs.

Construction of the semantic network of PrS/S in the TS Route-1

When carrying out the formalization in the conceptual model "Stage", the following PrS/Ss in the TS "Route-1 were selected:

- PrS/S (K(EF)-I) "correction of the formation of the EF". Task of the K(EF)-I - an operational change in the structure of the EF to cover the detected attack-dangerous direction;
- PrS/S (K(EF)-I) "Counteract the immediate threat to SA". Task of the K(EF)-I - allocation of the subgroup SA (dedicated group, commander K(EF)-II) to repel the attack of HF;
- PrS/S (K(EF)-I) "Return of the selected sub-group of fighters into the escort formation". Task of the K(EF)- having ascertained the status and capabilities of the selected group after the completion of its counteraction by the HF, it is compulsory to solve for the K(EF)-II the task of operational goal-setting before it.

The listed PrS/Ss are presented for the considered TS "Route-1" semantic network, in which for each PrS/S the conditions for its occurrence are formulated (the result of the deep study of the TS "Route-1", which is previously represented by a set of PrS/Ss).

The recommended solutions to the problems in the above PrS/Ss are prepared for the K(EF)-I by the on-board advisory expert system "OAES-Route-1".

4. Conclusion

1. A Statement of the task of determining the content and composition of tasks for operational intellectual support of the process of solving tactical decisions by the crew of an aircraft is possible only when the conceptual model "Episode" is changed to the conceptual model "Stage".
2. The composition of the onboard intellectual systems of a tactical level is defined:

- the Intelligent Information System "Situational Awareness of the Crew" (IIS SAC) and the on-board advisory expert system "Operational goal-setting" (OAES-goal-setting) that constantly work when the aircraft performs the flight task;
- a set of on-board advisory expert systems of typical situations / flight stages (OAES-TS), each of which operates at its own stage of flight.

The functions of the IIS SAC, the OAES-goal setting, the OAES-TS and the structure of their knowledge bases are described.

5. References

1. Fedunov B.E., Simkina N.D. “The conceptual model of anthropocentric objects for the onboard tactical intelligence systems”. In: *Proc. of the 19-th International Workshop on Computer Science and Information Technologies (CSIT’2017), Germany, Baden-Baden*. October 8-10, 2017. Vol.1, pp.211-214.
2. B. E. Fedunov, “The model Etap for the development of characteristics of on-board intelligent systems for anthropocentric objects. Ontologiya Proektirovaniya, No. 2(4), 36–43 (2012)
3. Kolisnichenko A.V, Fedunov B.E. “Onboard intellectual information system “Situational awareness of the helicopter crew””. Mechatronics, automation, control. 2016, No. 10, pp. 703-708
4. Zheltov S. Yu. and Fedunov B.E. “Operational Goal Setting in Anthropocentric Objects from the Viewpoint of the Conceptual Model called Etap: I. Structures of Algorithms for the Support of Crew Decision-Making”. J. Comput. Syst. Sci. Int.Vol.54, No.3, pp.384-398.(2015).
5. Zheltov S.Yu. and Fedunov B.E. “Operational Goal Setting in Anthropocentric Objects from the Viewpoint of the Conceptual Model called Etap:II. Operation Modes of the On Board Real Time Advisory Expert System and Its Dialog with the Crew Journal”. of Computer and Systems Sciences International #3, 16
6. Fedunov B.E. “Mechanisms of output in the knowledge base of on-board advisory expert systems”. –M., Izv. RAN. TiSY. 2002. № 4
7. http://aviationmoderne.chez.com/Presentation%20du%20Rafale%202.htm
8. Fedunov B.E. “Basic algorithmic shell of on-board advisory expert systems of typical situations of object functioning”. –M., Izv. RAN, TiSU. № 5, 2009. pp. 90 -101.
9. Stefanov V.A, Fedunov B.E “On - board operational - advisory expert systems (BOSES) of typical situations of functioning of anthropocentric (technical) objects”. –M., Publishing house MAI. 2006. 191 p.

IMPROVED STOCHASTIC CONTROL FLOW MODEL FOR LLVM-BASED SOFTWARE RELIABILITY ANALYSIS

V. Vidineev[1], N. Yusupova[1], K. Ding, A. Morozov[2], K. Janschek[2]
Faculty of Computer Science and Robotics, Ufa State Aviation Technical University, Ufa, Russia[1]
Technische Universität Dresden, Germany[2]
vidineev.vyachesla01@ugatu.ac.ru, yussupova@ugatu.ac.ru, kai.ding tu-dresden.de, andrey.morozov tu-dresden.de, klaus.janschek@tu-dresden.de

Abstract: *Recently we have proposed a new method for error propagation analysis of the safety-critical software using the transformation of the source code to the Dual-graph Error Propagation Model (DEPM) based on the Low-Level Virtual Machine (LLVM) compiler framework, that allows the automatic analysis of C-code or another LLVM supported front-end. The source code is compiled into the LLVM Intermediate Representation and instrumented in order to analyze control and data flow structures of the software and the control flow transition probabilities between the basic blocks. Based on this information a DEPM for further analysis is generated. The DEPM is a stochastic model that captures system properties relevant to error propagation processes such as control and data flow structures and reliability characteristics of single components, LLVM instructions in this particular case. The DEPM helps to estimate the impact of a fault in a particular instruction on the overall system reliability, e.g. to compute the mean number of erroneous values in a critical system output during given operation time. The feasibility of the method has been proven on several case studies and also reveals several limitations of the current control flow model.*
This paper address the improvement of the control flow model using a new customizable heuristic method for the analysis of control flow sequences and their mapping into discrete-time Markov chain models. The method is designed in a way to keep a required tradeoff between the model size and precision.

1. Introduction

Analysis of the software reliability is an important part of the system-level dependability evaluation for any safety-critical industrial domain. We are working on a comprehensive stochastic evaluation of the propagation of data errors through complex software structures. We consider the faults can be described stochastically, focusing on bit flips and other computing hardware or network originated faults.

The evaluation of the fault probabilities is left out of the scope of our research focus. Examples of the studies for space and automotive domains can be found in [1, 2].

The main goal is the evaluation of the probability that a data error will reach a critical system output given the fault activation probabilities and the system operation time. The mathematical model behind our analytical approach is called Dual-graph Error Propagation Model (DEPM) that will be discussed in more details in Section 2.1. We are putting effort on automatic generation of these models from the source code. An overview of our LLVM-based DEPM generation approach will be given in Section 2.2.

This paper introduces a new, improved algorithm for the generation of stochastic control flow model for the DEPM. The method is designed in a way to keep a required tradeoff between the model size and precision. Section 3 gives a reference example that will be used in Section 4 that in turn describes the proposed algorithm.

Fig. 1 shows a simple DEPM example. An *element*, e.g. *A*, *B*, or *C*, represents a fundamental executable part of a system. Each element may receive input data and provide output data. A *data storage*, e.g. *d1*, *d2*, or *d3*, represents a variable that can be read or written by an *element*. For instance, the *element C* reads *d2* and *d3*, and writes output. A *control flow edge* (black lines in Fig. 1), weighted with an attribute probability, represents a control flow transition between the *elements*. For instance, after the execution of *A*, *B* will be executed with the probability *0.7* and *C* with the probability *0.3*. A *data flow edge* (purple lines in Fig. 1) describes data transfer between the *elements*. A *data flow* connects an *element* with a *data storage* or vice versa. The *data flow* edges are considered to be the paths of the data error propagation. Fault activation and the error propagation are specified using probabilistic *conditions* of the *elements*, see the conditions of *A*, *B*, and *C* in the right part of Fig. 1.

2. State of the Art

2.1. Dual-graph Error Propagation Model

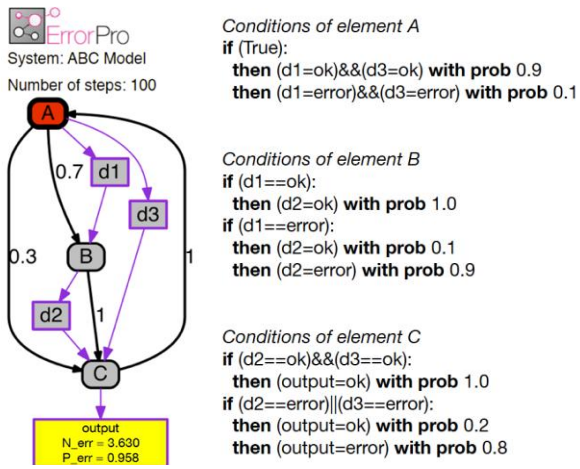


Fig 1. *A simple DEPM example and the conditions of the elements [4].*

A DEPM [3-5] is a mathematical model that captures control and data flow aspects of a system, described using the following set-based notation:

$$DEPM := < E, D, A_{cf}, A_{df}, C >$$

Where, E is a non-empty set of executable system 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, and C is a set of conditions of the elements.

During the execution of an element, faults can be activated and occurred errors propagate to its output data. For instance, in the *element A*, faults can be activated with probability *0.1*, defined in the conditions of *A* (see Fig. 1), and occurred errors propagate to its output data *d1* and *d3*. The error propagation probabilities for each element are defined also using probabilistic *conditions*. The errors can propagate from the inputs to the outputs. For instance, the *conditions* of the *element B* specify that the *element B* does not activate faults, but the errors can propagate from *d1* to *d2* with the probability *0.9*.

The DEPM allows the computation of several reliability metrics, such as the mean number of errors (N_{err}) and probability of errors

(P_{err}) in selected data storages. N_{err} stands for the average number of erroneous values in a data storage, and P_{err} is the probability of an error in a data storage during the system execution. For instance, the evaluated N_{err} in the data storage output during *100* steps (execution of one element is one step) is equal to *3.630*, and the P_{err} is *0.958*, as shown in Fig. 1. The computed reliability metrics are important measures for the system analysis, particularly for the reliability assessment and should comply with system requirements.

2.2. Generation of DEPM from source code

The DEPM generation method is based on the Low-Level Virtual Machine (LLVM) compiler framework, that allows the automatic transformation of C-code or another LLVM supported front-end into a DEPM. The source code is compiled into the LLVM Intermediate Representation and instrumented in order to analyse control and data flow structures of LLVM instructions and control flow transition probabilities. The obtained information is transformed into the formal DEPM XML for further analysis.

Fig. 2 shows the architecture of the proposed transformation method. Rounded rectangles with blue borders represent activities and grey rectangles represent data. Two LLVM passes have been developed as well as the python script that processes the outputs of the passes and generates DEPM XML file for further analysis with our tool OpenErrorPro [7]. All the steps are automated and can be run with a single shell script that calls LLVM tools (compilers, linkers etc), as well as our passes, and the DEPM generation python script. The detailed technical description of the transformation method is given in [6]. Here we would like to focus on the control flow part because it is relevant for the contribution of this paper.

The transformation pass, shown in Fig. 2, helps to identify elements of the future DEPM, control flow structure, and control flow transition probabilities. The pass takes the LLVM IR of the original C-Code generated with *clang* as input and performs two tasks:

(1) The first task is to generate the list of DEPM elements that represent single LLVM IR instructions. Using the LLVM API, the pass iterates over all LLVM IR instructions and stores the information into the *elements.txt* file. The pass gives a unique name for each instruction taking into account its location according to the LLVM IR hierarchy.

(2) The second task is the generation of the instructions execution sequence in order to define the structure of the DEPM control flow as well as control flow transition probabilities. In order to achieve this, we link an external C-code of a *"print"* function with the original LLVM IR using *llvm-link* tool and append calls to this function after each instruction of the original LLVM IR. After this instrumentation, we compile and run the program. The instructions execution sequence is stored in *sequence.txt*.

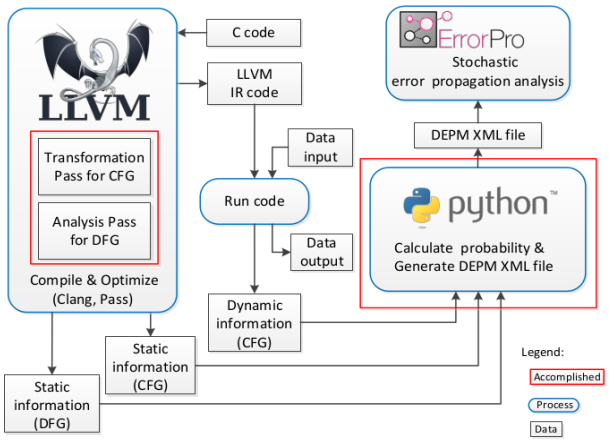


Fig 2. *Top-level architecture of the transformation method from C code to DEPM [6].*

After that the python script processes this sequence and in order to generate the DEPM control flow graph. In our original version, all control flow transitions are considered to be stochastic. We just count the number of control flow transitions between the elements and transform this information into control flow arcs and their transition probabilities.

However, this method has several drawbacks. This control flow generation method might cause a problem of discrepancy between the generated probabilistic control flow graph and the original deterministic sequence. Some parts of the original sequence can have a strong visible pattern that will be lost after the mapping of the sequence into the control flow graph that follows the Markov property and the control flow probabilities depend solely on the current element and do not take the execution history into consideration.

2.3. Serial test

The proposed solution in this paper includes the testing of the generated sequence for randomness. This can be done using the serial test method that is commonly used for testing random number generators. It is based on Chi-squared statistic to compare adjacent pairs or triples of numbers.

For instance, the serial test method is provided by the python library skidmarks [8], the detailed algorithm description is given in [9]. Listing 1 shows two examples of more and less random sequences:

```
>>> serial_test('101010101111000')
{'chi':1.4285714285714286,'p':0.6988513076924
8427}
>>>
serial_test('11000000000000011111111111')
{'chi':18.615384615384617,'p':0.0003283102182
6061683}
```

Listing 1. Examples of the skidmarks serial test on two binary sequences

To make a decision about randomness of the sequence we have to take a look at the p-values of the test. If the p-value is less than a defined constant the sequence is considered to be random.

3. Reference example

Let's consider the next abstract example in order to explain the main idea of the proposed methods for control flow analysis. Suppose that our reference piece of software operates according to the control flow graph that is shown in Fig. 3.

Assume that the discussed method for control flow analysis generated the following execution sequence:

$$ABE\ ABE\ ABE\ ACDE\ ABE\ ACE\ ABE\ ACDE\ ABE\ ABE\ ACE\ \ldots$$

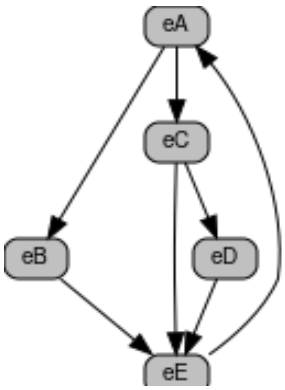


Fig 3. *The reference DEPM control flow graph.*

In this sequence there are two elements, which have two or more output control flow arcs: namely A and C. In the transitions from A to B or C we observe no regularity, while in the transitions from C to D or E we can see that it happens not by chance but according to a clear pattern:

$$CD, CE, CD, CE, CD, CE, \ldots$$

The results are confirmed by the serial test shown in Listing 2:

```
>>>
serial_test('BBBCBCBCBBCCCBCCCBCCCBCBBBCCCCBC
BBCBC')

{'p':        0.6695087450687482,        'chi':
1.5555555555555556}

>>>
serial_test('DEDEDEDEDEDEDEDEDEDEDEDEDEDEDE
DEDED')

{'p': 1.0, 'chi': 0.0}
```

Listing 2. Skidmarks serial test results for the reference example.

So, we can consider the control flow transitions after the element A stochastic and the transition after the element B deterministic.

4. Algorithm description

Generalizing these results, we can define the next steps:

Step 1: Build a control flow graph based on the original sequence of execution of elements.

Step 2: Find elements that have two or more output arcs.

Step 3: For each such element build a sequence of elements which will be executed after this element.

Step 4: Check each of these sequences for randomness using the Serial test.

Step 5: For each sequence that is considered to be random statistically estimate control flow transitions probabilities based on the number of the actual transitions in the original sequence.

Step 6: For each not random sequence add a data storage with the rules that describe the uncovered control flow pattern.

5. Results of the algorithm application

The result of this algorithm applied to the reference example is shown in Fig 4. In this DEPM, the element *A* has purely stochastic control flow transitions: we jump to the element *eB* with the probability *0.63* and to the element *eC* with the probability *0.37*.

In the other hand, the element *C* contains the following conditional, guarded by the value of the data storage *i*, control flow transitions:

1) (i=0) => (cf'=eD) & (i'=i+1)

2) (i=1)=>(cf'=eE) & (i'=0)

That define the following control flow behavior. If the value of the data storage *i* is *0* that we jump to *eD* and increase the value of *i*. If the value of *i* is *1* then we jump to *eE* and reset *i* to its initial value *0*.

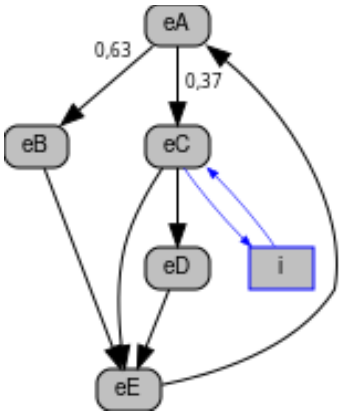


Fig 4. *The resulting DEPM control flow graph generated using the proposed algorithm.*

4. Conclusion

The data error propagation analysis is an important part of the reliability evaluation of safety critical software. In our recent works we have proposed a stochastic Dual-graph Error Propagation Model (DEPM) and an automatic method for the generation of the DEPM for the software using LLVM technology. In this paper, a new algorithm for the automatic generation of the DEPM control flow graph has been introduced. The algorithm is based on the statistical evaluation of the execution sequence. The key feature of the new algorithm that it can distinguish between stochastic and deterministic control flow transitions via the application of the serial test for randomness.

References

1. I.Verzola, A.E.Lagny, and J.Biswas, "A predictive approach to failure estimation and identification for space systems operations," SpaceOps 2014 (Pasadena, CA), 2014.
2. P. Koopman, "A case study of toyota unintended acceleration and software safety," *Presentation. Sept*, 2014.
3. A. Morozov, Dual-graph Model for Error Propagation Analysis of Mechatronic Systems. Dresden: Jörg Vogt Verlag, 2012.
4. A. Morozov and K. Janschek, "Probabilistic error propagation model for mechatronic systems," Mechatronics, vol. 24, no. 8, pp. 1189 – 1202, 2014.
5. K. Ding, T. Mutzke, A. Morozov, and K. Janschek, "Automatic transformation of uml system models for model-based error propagation analysis of mechatronic systems," IFAC-PapersOnLine, vol. 49, no. 21, pp. 439–446, 2016.
6. A. Morozov, Y. Zhou, K. Janschek, "LLVM-based Stochastic Error Propagation Analysis of Manually Developed Software Components", Proceedings of European Safety and Reliability Conference (ESREL). Trondheim, Norway, 2018.
7. OpenErrorPro on the github. https://mbsa-tud.github.io/OpenErrorPro/, 2018
8. Skidmarks library for randomness serial test https://pypi.org/project/skidmarks/, 2018
9. Morgan, Byron JT. *Elements of simulation*. Vol. 4. CRC Press, 1984

METHODOLOGY FOR CONFIGURING AN INTELLIGENT NETWORK EXPERT SYSTEM [AINES] USING THE EXAMPLE OF AN ELECTROCHEMICAL DIMENSIONAL PROCESSING MACHINE

Davletbaev A.S., Kulikov G.G., Startcev Y.V.
Ufa State Aviation Technical University, Ufa, Russia
E-Mail: DavASko01@mail.ru, gennadyg_98@yahoo.com, startcevy@mail.ru

Abstract: *The article deals with the configuration of the expert system for working with electrochemical equipment. The process of electrochemical dimensional processing (EDP), its main parameters and the influence of conditions on their change are considered. The method of setting up the system for working with the EDP machine is given and an approximate variant of the result of the system operation is considered*

1. Introduction

In modern production, especially in such complex high-tech areas as engine-building, advanced developments in the field of materials science are used. New types of metal alloys are being developed, which are used in the most severe conditions. The details are being introduced, the production of which is a huge number of technological operations. Accordingly, the cost of manufacturing it increases with each new operation. Especially heavily on its cost is affected by complex technological operations, which require huge accuracy, up to several microns. Accordingly, equipment that can perform such operations with a sufficiently high speed and performance is necessary. One such machine is the EDP machine. The use of electrochemistry makes it possible to process metals with sufficiently good accuracy, roughness and speed, which is achieved due to fine tuning of the equipment. The setup includes several dozens of parameters, each of which is very significant and directly affects the final result. This imposes a great burden of responsibility before the operator-technician (technologist) in the preparation of production, because fine tuning of the equipment, at the moment, is often on the verge of intuition, coupled with extensive experience and numerous works of the technologist. It was to facilitate the routine of the technologist that an automated intellectual network expert system [AiNES] was designed. Thanks to such a system, the commissioning of production will be greatly facilitated, and the technologist's load, in part, will fall on the lightweight technological devices that today surround us everywhere.

2 The process of electrochemical dimensional processing for the formation of GTE turbine blades

2.1. Description of the process of electrochemistry

The process of electrochemical processing is a method for processing electrically conductive materials, consisting in changing the shape, dimensions and (or) roughness of the workpiece surface due to anodic dissolution of its material in the electrolyte under the influence of an electric current [2]

For thinning, it should be added that **EDP** - dimensional processing is a directed process of dissolving the metal in order to give the workpiece the desired shape and size (Fig. 1). At **EDP** electrode tools attach a negative potential, electrode-workpiece is positive. In the space between them, liquid is supplied - an electrolyte, in the medium of which the metal of the electrode-billet is dissolved. In this case, the dissolution of the metal occurs at different rates. The smaller the interelectrode space (the inter electrode gap of the IEG), the faster the metal dissolves. Thus, during the time the rate of dissolution of metal over the entire surface tends to equalize, as a result of which the IEG is aligned along the entire gap, which gives the electrode-workpiece the same shape as the tool electrode. In this case, the dissolution of the electrode-tool itself is so small that this quantity is neglected.

Thus, the shape of the surface of the tool electrode is completely copied onto the surface of the workpiece electrode.

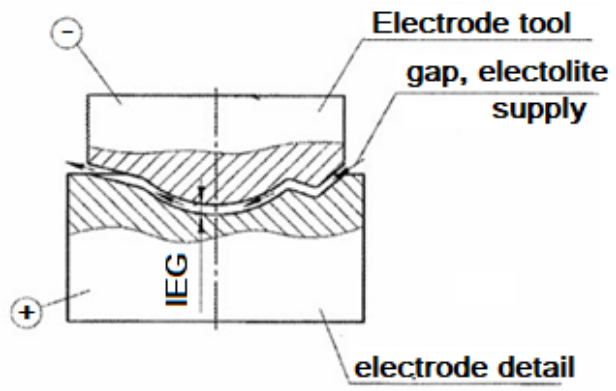


Fig. 1 *The EDP process*

This process that has become widespread in the manufacture of blades of gas turbine engines (GTE).

In our case, the process of manufacturing GTE blades on EDP machines with vibrations is considered, which is performed on ECM 38.5 machines.

2.2. Описание проблемной области ЭХРО

The problem that must be solved by the technologist is the parameters of the process itself (Figure 2). The cycle of the technological process consists of several stages.

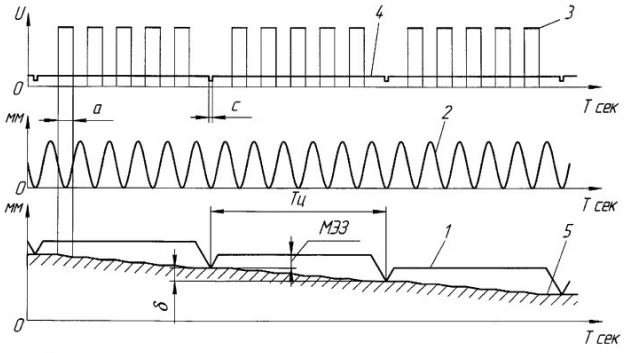


Fig. 2 *Cyclogram process EDP*

Each stage is important for the correct flow of the process itself and is carefully controlled by the machine's automation. There are the following main stages:
- search for the part coordinate;
- exit to the gap processing (Fig. 2 - δ);
- active part processing phase.

There are additional stages:
- outlet to the rinsing gap;
- exit to the starting position;
- reperforcing (special operation of bringing the coordinate axis of the drives into the initial synchronous position) of the tool electrode.

In the presented process, EDP occurs with a vibrating electrode (Fig. 2: 1 - the position of the tool electrode at time T, 2 - amplitude and frequency of vibration, conditionally), while the

zero position (the lower position of the tool electrode) is determined when the electrode is in the nearest position to the detail (Fig. 2 - 5). And the upper position zone is the electrode at the maximum distance from the part. The frequency of the vibration determines the frequency of the working current pulses (Fig. 2-a), as can be seen, the current (Fig. 2 - 3) is fed in the lower position zone, when it gives the maximum impact on the workpiece surface, in the zones "in the middle" between the lower and upper zone the voltage pulse switches its state (on / off), while in the upper position zone the most efficient washing of the metal dissolution zone takes place, during which electrolyte streams carry away the decomposition products and partially cool the working surfaces STI.

When searching for a part, they are guided by the voltage dip (Fig. 2 - 4) at the moment of contact (Fig. 2 - c), which is the evidence of the contact of the part and the electrode. The time interval between the search for the part is called the processing cycle time (Fig. 2 - Tц), so you can see the set of parameters that need to be set in the EDP machine for the correct flow of the process.

2.3. Formulation of the problem

As can be seen from the brief description of the EDP process, there are various parameters that need to be monitored. For the correct operation of the [AiNES] system, it is necessary to correctly configure these parameters and distribute them in groups:

- machine parameters;
- detail parameters;
- technological parameters of the process;
- parameters of external conditions;
- parameters of the result of processing.

It is also necessary to determine which of them are formalized and which are not, determine the limits of their values and the probable values.

In other words, it is necessary to model the parametric model of the EDP process.

Then it is necessary to determine the criteria and methods for estimating the result, which will be obtained at the output of the decision system.

It is necessary to evaluate the efficiency of using the system on the machine ECM 38.5.

3. The solution of the task

3.1. Determination and distribution by groups of parameters of EDP

3.1.1 Machine parameters

For the EDP process for the manufacture of GTE blades in practical applications, the following parameters are determined as a result of the observation, which can be changed by the technologist (in the brackets, the limits for the ECM 38.5 machine):

- min / max voltage produced by the machine (5/65);
- max current given by machine (3000A imp);
- max positioning accuracy of machine (4 mkm);
- min / max frequency of vibration of the tool electrode (0/50 Hz);
- min / max pressure of the electrolyte, issued by the pump (0.1 / 10 Atm);
- min / max cycle time that the equipment can withstand (1/900 sec);
- min / max speed of the electrode movement that the machine can provide (10/10000 Mkm / s).

These parameters do not directly affect the solution of the system, but are crucial for filtering the results of calculations. If there is a deviation in any parameter, then it is considered that for the given parameters processing cannot be performed on this equipment. It is necessary either to modernize the equipment, because to perform the operation on another machine.

3.1.2 Detail parameters

The parameters of detail refer both to the detail and to the finished part, i.e. is considered in the context of the work object:

a) material of deail (alloy 20X13 was studied);

b) total area of the treated surface (of the order of 50 cm^2 on two sides);

c) curvature of the surface (curvature of the standard GTE blade);

d) number of simultaneously processed parties (2);

e) stiffness of fixing detail (deviations at loads of 100 kg +/- 10 μm);

The parameters "a" and "c" are unformalized in the presented list, therefore we see the formalization procedure by splitting the parameters into components, if possible, or we make up a categorical dictionary for assigning an individual number and value parameters (necessary for the use and training of the neural network)

For the parameter "a" - detail material, we define:

- hardness of the material (126-197 MPa);
- specific gravity (7670 kg / m^3)
- valence of detail material (4);
- Electrochemical equivalent of the material of detail (1.94 kg / Cl);

For parameter "c" we define the curvature dictionary, the name of the curvature index (value 1).

3.1.3 Process technological parameters

This group of parameters is the required group of parameters that must be transferred to the technologist for setting up the machine. The following parameters are selected:

- processing stress (20 V);
- working gap (100 microns);
- vibration frequency of the tool electrode (25 Hz);
- number of washing cycles per period (2 times);
- washing distance (250 μm);
- pressure, under which the electrolyte is supplied (1.5 Atm);
- duration of the active phase of dissolution (5 sec.);
- delay of the beginning of the pulse with respect to the point of the upper position (90 deg);
- phase of the end of the voltage pulse (270 deg);
- speed of electrode movement during the active phase of metal dissolution (0 μm / sec).

These parameters were used during the test processing and are given for the comparative analysis of the work results (they are stored in the database as the set of parameters chosen for operation).

3.1.4 Parameters of external conditions

This group of parameters that cannot be configured programmatically, but can be indirectly changed. They also affect the EDP process. These include:

a) type of operation (complex dimensional shaping);

b) material of the tool electrode (stainless steel alloy);

c) composition of electrolyte (sodium nitrate 7%, sodium chloride 7%, water 86%)

d) temperature of the electrolyte (25 degrees);

e) electrode area (total for both electrodes is 50 cm^2)

e) positioning accuracy (5 μm);

Among the presented parameters parameter "a", "b" and "c" do not have a numerical representation. Therefore, we need to formalize them.

For the parameter "b" the formalization is done by splitting the parameter into:

- valence of the electrolyte (4);
- viscosity of the electrolyte (2.3 mPas);
- electroconductivity of the electrolyte ($3.54x10^{-2}$ S / m);
- density of electrolyte (2.81 g / cm^3).

For parameter "b" it is divided into:

- valence of the workpiece material (4);

- Electrochemical equivalent of the material of the billet (1.94 kg / Cl);
- Hardness of the material (126-197 MPa);
- Specific weight (7670 kg / m^3)

For the "a" parameter, a dictionary of operation types is defined, for our case we enter the value type 1.

3.1.5 Results parameters

The expected processing parameters have the same set as the actual processing result. It is during the comparison of these parameters that the success of processing is determined. The list of parameters is as follows:
- processing speed (less than 15 minutes per detail / 12 minutes per detail);
- processing accuracy (maximum deviation 100 μm / max deviation 70 μm);
- Rmax roughness (10 μm / 15 μm);
- percentage of defects (0% / 5%).

For parameters in parentheses are specified (set value / value obtained)

3.2. Distribution of parameters in the problem being solved

The groups of parameters considered hereafter are used to configure and train the neural network (Fig. 3). For input parameters are used:
- parameters of detail;
- parameters of external conditions;
- parameters of result (expected).

For output parameters, a group is used - process parameters of the process. They are of the greatest interest to the technologist during the setup of the equipment.

At the same time, the result of the operation of the neural network is one resulting set of parameters, which are then fed to the filter algorithm along with the hardware parameters in order to ensure that the process with the parameters found can occur within the constraints of the current machine. Also on this filter pass all the solutions found by the system. It is assumed that this particular machine should perform this operation. If the filter does not miss any of the solutions found, the most appropriate set of parameters is selected and provided to the technologist with a note that the set of parameters is not optimal and has unattainable values.

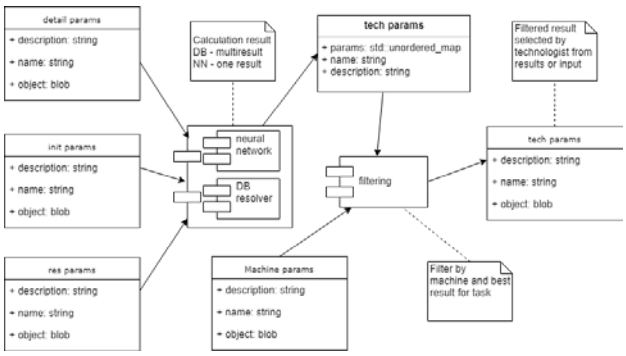


Fig. 3 *Distribution of parameters*

The resulting database for storing all the necessary data is built on the basis of the three main fields of the tables (Figure 4):

- Name- specifies the name of the parameter set (the default date and time of creation)
- Object - the saved serialized parameter object (list of parameters with type and description). The parameter is stored as a binary sequence, the entire class is saved.

Description - the description of the current object.

Also, there are tables in the database for configuring the default values (Figure 4 of the table with the prefixes "Def"). They serve as the basis for the structure of the neural network and determine the minimum necessary set of parameters for the formation of training samples of the neural network.

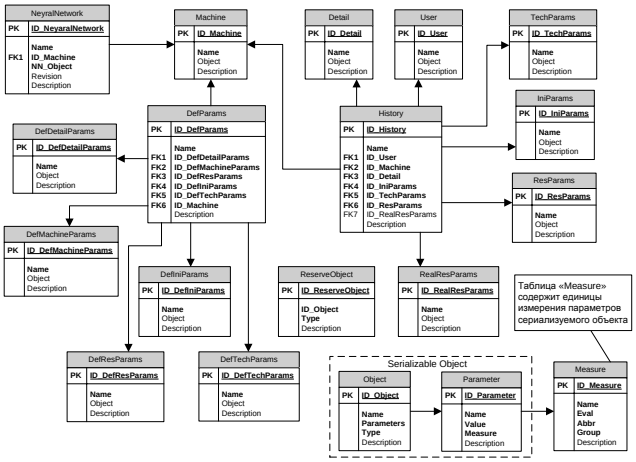


Fig. 4 *Structure of the node database [AiNES]*

All neural networks after the post-training are stored in the database in the table "Neural Network". Each under its own name is linked to the machine for which it was trained and with the revision number for further optimization and rollback when working with neural networks. The summary entry "History" is the base record of the system. It represents the history of operations and works that were conducted in the system. Data extraction is complex, and the input of new data can be combined and duplicated to speed up the input (auto-substitution mechanisms).

On the side, there is only a table of units of measurements that are stored by a separate branch and synchronized among the nodes to preserve valid consistent results between the nodes of the system.

4. Practical implementation of the system

4.1 The final structure of the artificial neural network

Based on the parameters defined above, we compose a neural network (Fig. 5).

In the design and configuration of a neural network, 26 inputs were assigned to the input, distributed in three groups (as described above). The input groups that were used for the neural network have one additional input, the so-called activation inputs. The use of such inputs (to which only one is fed) favorably affects the overall result. In the first layer (Normalize level), the parameters are normalized in such a way that the output values are given in the range from -1 to 1. This is necessary so that the weights of the neural network during training do not become too large or very small. Accordingly, at the output level (neurons from 27 to 36) perform the reverse operation of converting the values to the allowable range of the final values.

Intermediate layers: the first hidden layer - 53 neurons (26 input * 2 + 1), and the second hidden layer - 107 neurons (53 from the first hidden layer * 2 + 1) - calculated according to Hecht-Nielsen's theorems is assumed equal to 2n + 1, where n is the number of input signals and the number of neurons of the previous layer.

The above structure of the neural network is sufficient to determine the necessary technological parameters. Training samples for training were obtained, both as a result of ongoing work in the workshop, and on the basis of specially conducted tests.

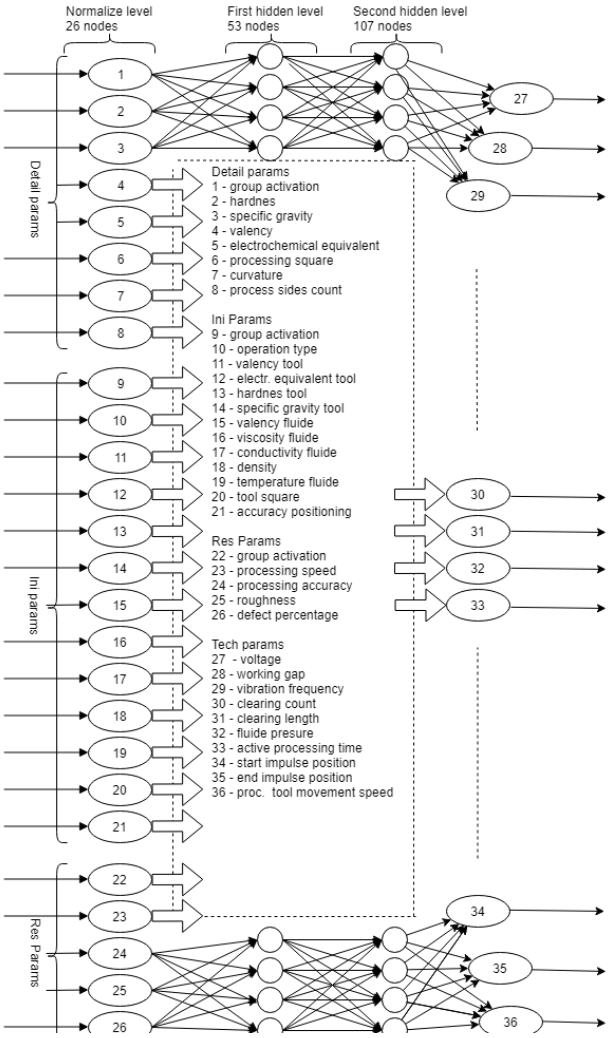


Fig. 5 *structure of the neural network for the ECM 38.5 machine*

4.2. Using the system in the workshop

Interaction with the system within the workshop (Figure 6) is carried out directly by the technologist. The technologist is authorized in the system by means of an individual login/password. After successful authorization gets access to the system. Then, within the problem of finding a solution, enters the initial parameters (selects the machine, selects the part, enters the initializing parameters and enters the parameters of the desired result).

The system saves the current parameters and forms a search task. Asks the user to confirm the beginning of the procedure for calculating the parameters. After the technologist confirms the beginning of the calculation, the system packs the task and sends it to the network to other nodes [AiNES] via the network of the system ([AiNES] network), where the free nodes of the system (another node [AiNES]) perform the common task and the decision (search decision) returns the result (decision). In turn, in parallel, the system searches for a ready solution in the database (search local DB result) and calculates the solution on the neural network (calculate on neural network), if it is considered ready for use. All results are collected in a single solution pool and filtered.

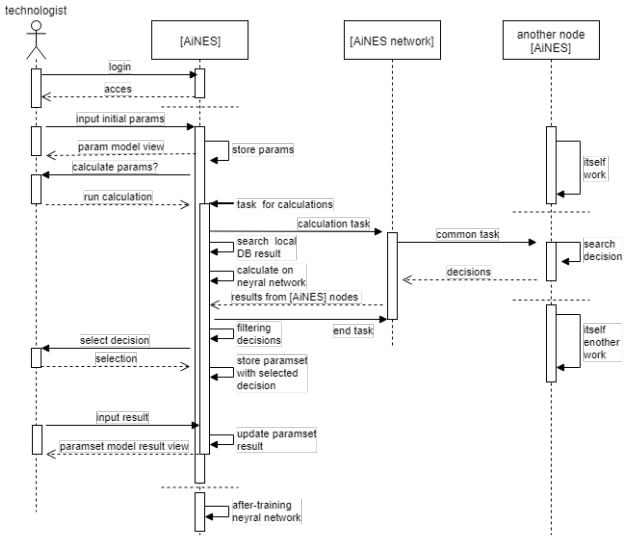


Fig. 6 *time diagram of the system operation*

As a result of the filtering, the technologist is provided with a set of several optimally optimal results. The system asks to choose one of them, or enter its own set, after which the whole set of parameters is stored in the database and waits for the input of the real part processing result, after input, to which the neural network can be trained or modified.

This completes one configuration cycle.

4.2. Preliminary results of using the system

As a result of the conducted experiments, the control measurements of the operations performed by the technologist were made (Fig. 7). The experiments were carried out on the basis of test fixtures, parts and tools. The main purpose of the experiments was the accumulation of data for the training of neural networks. The operating time of one tuning cycle was from one hour to 10.6 hours (considering 1d = 8 hours, working shift).

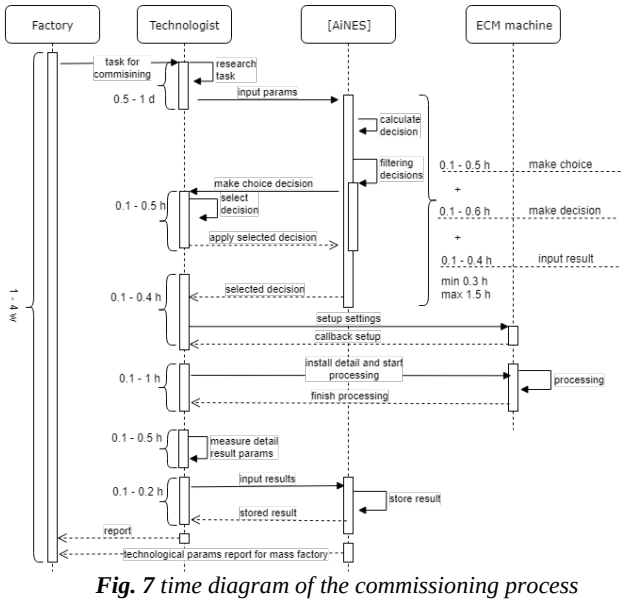


Fig. 7 *time diagram of the commissioning process*

In this case, the actual work on the device (interaction with the system) is from 0.3 hours to 1.5 hours per cycle. In relation to the work of a technologist without a system, it takes time for one cycle from 0.7 hours to 9.1 hours. It should be noted that the number of cycles using the system is reduced by an average of 4 cycles, which is from 2.8 hours to 36.4 hours, respectively. The spread is very large, but it can be seen that even with less savings (2.8 hours), this indicator is very significant. Moreover, the reduction in the number of cycles allows the use of fewer

blanks that can be spoiled. Given their cost (from 200r to several thousand rubles), the savings becomes obvious.

5. Conclusion

The article discusses the process of EDP, defines the main parameters of the process, the main difficulties in working with this technology. The example shows a method for configuring a neural network for operation in the [AiNES] system, and methods for generating parametric process models are shown. Examples of technological and other parameters that were obtained and used in the tests carried out on the machines ECM 38.5 are shown. An example of a neural network is shown and the algorithm for working with the system is shown, the time indices of the cycle of tuning the machine for processing one part are determined. A comparative analysis of the economy of the enterprise was carried out when using the system, which amounted to saving about 4 cycles of tests on the machine, which saves from 3 hours of work time of the technologist and from 800 rubles on the material of blanks.

Thanks

The author is grateful to Cand. tech. Gantsev Rustem Khalimovich for providing the equipment and consulting, ch. Technologist of JSC ISTC "Iskra" Semashko Andrey Pavlovich for consultations on the technological part of the work on electrochemical machining machines, Head of the electronic automation department of JSC ISTs "Iskra" Gilmetdinov Marat Hammatovich for help in organizing and constructing the test program for ECM 38.5 machines

List of sources used

1. 1. Senthil Kumar K.L. Modeling of metal removal rate in machining of aluminum matrix composite using artificial neural network // Journal of Composite Materials 45 (22), DOI: 10.1177 / 0021998311401083 pp. 2309-2316
2. 2. A. JONES D. A prototype expert system for manufacturing resource planning // I NT. J. COMPUTER INTEGRATED MANUFACTURING, VOL1, NO.3. pp. 165-170
3. 3. A. Macioł Ma cio P. drusik ł • St. Je¸ • Lelito J., New JL, Managing Techniques (2015) 79, DOI: 10.1007 / s00170-015-6860-5 pp. 1733-1745
4. 4. Davletbaev A.S., Startcev Y.V., Kulikov G.G. Development of the Automated Networking Expert System [AiNES] // The 17th International Workshop on Computer Science and Information Technologies CSIT'2015, (Rome, Italy Sep. 22-26 2015) volume 1, pp. 11-15
5. 5. Swingler K. "Applying Neural Networks .A practical G-uide" (translated by Y.P. Masloboev) URL: http://matlab.exponenta.ru/neuralnetwork/book4/3_2.php - (date of treatment: 01/25/2016)
6. 6. Bakhmetova NA, Tokarev S.V. MODELING TECHNOLOGICAL PROCESSES WITH NEURAL NETWORKS [Electronic resource] Modern high technology. - 2008. - No. 2. - P. 139-140 - URL: http://www.top-technologies.ru/en/article/view?id=23277 (reference date: 16.03.2016).
7. 7. D. Tarkhov, "Neural network models and algorithms. Handbook ", Radio Engineering, ISBN 978-5-88070-376-0; 2014
8. 8. MA Bosinzon "Modern CNC systems and their operation". Textbook for the beginning. prof. of education. Academia, 2009 ISBN 978-5-7695-6060-6
9. 9. Andrew S. Tanenbaum, Maarten van Steen. "Distributed systems." Principles and paradigms ". - St. Petersburg: Peter, 2003. - 877 p. - (Classics of computer science). - ISBN 5-272-00053-6.
10. 10. Wikipedia, the free encyclopedia, "Electrochemical machining" URL: https://en.wikipedia.org/wiki/Electrochemical_machining (circulation date 21.02.2018)

GEOINFORMATION ANALYSIS WITHIN THE OBJECTIVES OF WASTE MANAGEMENT SYSTEMS CONSTRUCTION

V.E Gvozdev, O. Khristodulo
Geoinformation Systems Department, Ufa State Aviation Technical University, Ufa, Russia
o-hristodulo@mail.ru

***Annotation**: This article compiles different features of waste management systems, as well as its practical implementation in Mathematic and Geoinformation analysis. Technological basis of information support of waste management and its indicators, such as adverse impact on the natural environment and human health and territorial features of the risk caused by long-term negative impact on the environment and public health, are presented. Population, economy, and environment in the Republic of Bashkortostan, Russia are strongly affected by thorough decisions made by authorities in the field of waste management, that are: finding places for storage, recycling, and subsequent distribution of waste and its derivatives, that also includes taking measures to influence people's understanding of dealing with waste and environmental problems. In this paper, the focal point is the technology of mathematical-geoinformation modelling, the basis of which, as applied to waste management problems, is a system combination of geoinformation technologies and mathematical-statistical modelling methods. Approaches to solving the problem of classifying areas of the territorial system by state indicators are presented in the article.*

1. Features of waste management systems

The waste management system falls into the class of open complex systems. Waste management involves various interested parties with different motivations and different personal ontological models of the problem situation conditioned by the generation, storing and processing of waste. This makes it necessary to improve the methods for developing consolidated decisions that are recognized by all interested parties. The provisions of the theory of convergent management and evergenetic can serve as a basis for the methodological making of consolidated decisions.

Waste management has to be carried out under uncertainty determined by a lack of knowledge about the mechanisms of waste's impact on the territorial systems; the difference in goals and motives of representatives of state structures, business and the population in resolving problem situations; limited data characterizing the processes of generation, collection, storing, and processing of waste; insufficient knowledge of the cause-effect relationships of waste, environment, and public health. The above circumstances determine the need to develop a consolidated solution for resolving the problem situation on the basis of axiological knowledge. Structured personal knowledge of subjects, as well as the results of a multidimensional analysis of data from departmental information systems for solving diverse tasks related to ensuring environmental, technogenic safety and public health, serve as the elements of the communication platform.

System-forming factors, allocating a territorially-distributed waste management system, are:

- The amount and characteristics of industrial and domestic waste generated, stored and processed at different sites of the territory;
- The territorial-temporal dynamics of the impact of waste on the environment and public health;
- Territorial-temporal aspects of collection, storing, transportation, and processing of waste;
- Territorial-temporal features of the population's attitude towards the problem of waste management.

As local control loops in the waste management system, it should be noted:

- Management of the negative impact of waste on the environment and public health;
- Management of the choice of places for permanent and temporary storing of waste, construction sites for waste processing enterprises, taking into account the ratio of the volumes of waste produced and already stored in the storage areas of various types (industrial and domestic), and the state of the transport network in various parts of the territorial system;
- Management of the requirements for the activities of enterprises in terms of waste processing;
- Management of economic consequences of waste storing and processing, including pricing for the collection, storing, transportation and processing of waste;
- Management of the values, aspirations, behavior of the population in waste management issues, while changing the processes, methods, strategies, structure of the waste management system.

The above list is not final; its composition may be adjusted as the fundamental and axiological knowledge are being obtained while practical activities accumulate.
For each local control loop, the target groups of clients and associated management objectives are defined; indicators of effectiveness of management; indicators of management quality. The mutual agreement serves for coordination of management objectives in local contours. Coordination of processes related to local control loops must be reached by balanced combination of the following elements:

- Direct coordination (chief-subordinate);
- Standardization of goals (setting targets for executors who independently choose the means in order to achieve the specified management targets);
- Standardization of activities in the contest of implementation of the process approach;
- Standardization of the population's values in order to raise its awareness of the domestic waste management.

Critical management tasks are:

- Identification of self-reinforcing growth processes that create the basis for significant changes in the waste management system;
- Identification of processes that balk positive changes that determine its limiting factors;
- Development of approaches and the implementation of actions for the matching / elimination of limiting factors.

The main goal of coordination of local loops management is to ensure the consonance of self-enabling processes in local control loops.

The complexity of solving weakly formalized problems related to the construction of local control systems is due to:

- Necessity of substantiation of selection in initial descriptions of territorial systems of necessary knowledge, enabling to allocate control systems from the environment;
- Systematization of knowledge on the components of waste management systems and their reflection on

semantic concepts, in the basis of which later ontological models will be built and will form the communication basis for solving waste management tasks;
- Structuring of knowledge on the components of the waste management system for the subsequent presentation as structured formalized models (cognitive models).

Situational management is a perspective approach to solving problems of waste management in real time. It involves the use of pre-prepared "situation-solution" scenarios, which are then selected by the management entities in accordance with current situations. Currently, the solution of the above tasks is carried out largely empirically and heuristically due to the territorial features of the impact of waste on the environment and public health in the first place, and in the second due to the territorial features of the population's attitude to the problem of collecting, transporting and processing of waste. As a result, the quality of the models that form the basis of situational management depends on many subjective factors that are:

- Uncertainness in reaching a guaranteed result in the modeling process;
- Difficultness of assessing the completeness and correctness of the situational models constructed this way.

The combination of the factors in many cases leads to the rejection of modeling as an information support for management, and the subject-operator remains completely responsible for the management tasks.

2. Technological basis of information support of waste management

The presentation in structured form of knowledge and information of the involved entities as well as multidimensional analysis of available data characterizing the negative impact of waste on the environment and public health form the basis of information support for the development of a consolidated solution for resolving problem situations.

The implementation of this activity creates the basis for the formation of alternatives to solve tasks related to the collection, organization of temporary storing, and transportation of waste.

The technological basis for the creation of the management information platform includes:

- Cognitive modeling technologies;
- Convergent control technologies;
- Mathematical-geoinformation modeling technologies.

In this paper, the focal point is the technology of mathematical-geoinformation modeling, the basis of which, as applied to waste management problems, is a system combination of geoinformation technologies and mathematical-statistical modeling methods.

The use of geoinformation modeling technologies creates the basis for presenting in reference form (in the form of thematic cartographic layers) for the following:

- Information and data characterizing the natural components of geotechnical objects;
- Characteristics of origin of various types of waste (industrial, domestic) in different parts of the territorial system;
- Places of storage and processing of wastes;
- Characteristics of the transport network.

The use of mathematical and statistical methods creates the basis for the analysis of the territorial-temporal variability of the indicators of the components of territorial systems under conditions of uncertainty of the properties of interrelated processes, conditioned by natural, social, technogenic factors, including the processes of generation, storing, transportation and processing of waste.

3. Mathematics and Geoinformation analysis of indicators of waste management system

The territorial-temporal analysis of symptoms indirectly characterizing the processes controlled by means of the local control systems identified above [15-17] serve as the basis for assessing the state of the waste management system.

The table describes a number of indicators that indirectly characterize the state of the waste management system.

In Fig. 1 - 2 as an example, the following results are given: assessing the state of the territorial system, diagrams for the distribution of one of the indicators for the municipal districts of the Republic of Bashkortostan, as well as maps where the most critical areas can be identified, regarding the current problems of waste management system.

The results were obtained in solving the problem of zoning of the territory of the Republic of Bashkortostan by private indicators, which creates the basis for analyzing the effectiveness of solving local problems of waste management.

Table 1: Waste management quality indicators

Indicator name	Indicator content	Local control loop
The amount of waste per capita / km^2	Adverse impact on the natural environment and human health	Management of the negative impact on the natural environment and human health
Number of employed and neutralized waste per capita / km^2	Parrying the negative impacts of waste on the environment and public health	• Management of the negative impact on the natural environment and human health • Management of requirements for the activities of enterprises in terms of waste processing
The amount of landfilled waste per capita / km^2	Territorial features of the risk caused by long-term negative impact on the environment and public health	• Management of the negative impact on the natural environment and human health • Control of the selection of permanent and temporary places of storage and processing of waste • Management of economic indicators

		of storing and processing of waste • Management of values, aspirations, behavior of the population in waste management issues
Volume of output per capita	Territorial features of placing production (as a source of industrial waste)	• Management of economic indicators of storing and processing of waste • Control of the location for permanent and temporary places of storage and processing of waste • Management of values, aspirations, behavior of the population in waste management issues
Population density	Territorial features of the places of residence of the population (as a source of household waste)	• Management of economic indicators of storing and processing of waste • Control of the selection of permanent and temporary places of storage and processing of waste • Management of values, aspirations, behavior of the population in waste management issues

In papers devoted to the study of complex systems, it is noted that the classification problem solution is the basis of the analysis of heterogeneous dynamic objects [1, 6, 7, 8, 11]. Regarding the analysis of the territorial systems, this is expressed in the classification of areas of the territory (in case of waste management - municipal entities) by the values of the state characteristics [5, 9, 10, 12, 13]. Classification of areas of the territory is based on the development of formal methods for the regularization of spatial data, allowing each area to be assigned to a particular class of states.

In general, the classification problem is represented in the form:

$$\varphi(x,y) \to S(x,y), \qquad (1)$$

Where $\varphi(x,y)$ is the surface topography and $S(x,y)$ is a spatial mosaic, formed by sections of the study area, with each element of the mosaic is associated with a certain class of state territorial system [5].

The basis for the transformation of $\varphi(x,y)$ into $S(x,y)$ is the use of different scales that establish a correspondence between the ranges of values of the relief indexes of characteristics and the classes of state of the territory, that is

$$\alpha_i \le \varphi(x,y) \le \beta_i \to C_i, \qquad (2)$$

Where α_i and β_i are boundaries of the i -th sub-interval of the classification scale; C_i is the i -th class of the state of the site of the territory.

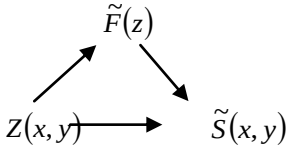
In solving practical problems related to the assessment of the state of a territorial system, it is rather typical that when only $Z(x_i, y_i) = \varphi(x_i, y_i), (i = \overline{i;N})$ are known instead of $\varphi(x,y)$ corresponding to the nodes of the non-uniform measurement grid. In this case, the model (3.1) is transformed to the form:

$$Z(x_i, y_i) \xrightarrow{P_\alpha} \tilde{S}(x_i, y_i), (i = \overline{i;N}), \qquad (3)$$

Where $\tilde{S}(x,y)$ is the estimate of $S(x,y)$ depending on N, the locations of the measurement grid nodes, and the chosen P_α transformation method of the terrain transformation into the spatial mosaic.

When solving applied problems of classification of territorial systems, one often has to deal with a situation where the classification scale is a priori unknown. In this case, $\tilde{P}_\alpha$ should strive for actually available sample data.

It is noted in [2, 3, 4, 5] that if $Z(x_i, y_i)$ are random variables, the distribution function $\tilde{F}(z)$ is an integral characteristic of the statistical relief. In this connection, the classification process should take into account features of evaluation $\tilde{F}(z)$, obtained from the values $Z(x_i, y_i), (i = \overline{i;N})$ and the model (3) is transformed to the form:



(4)

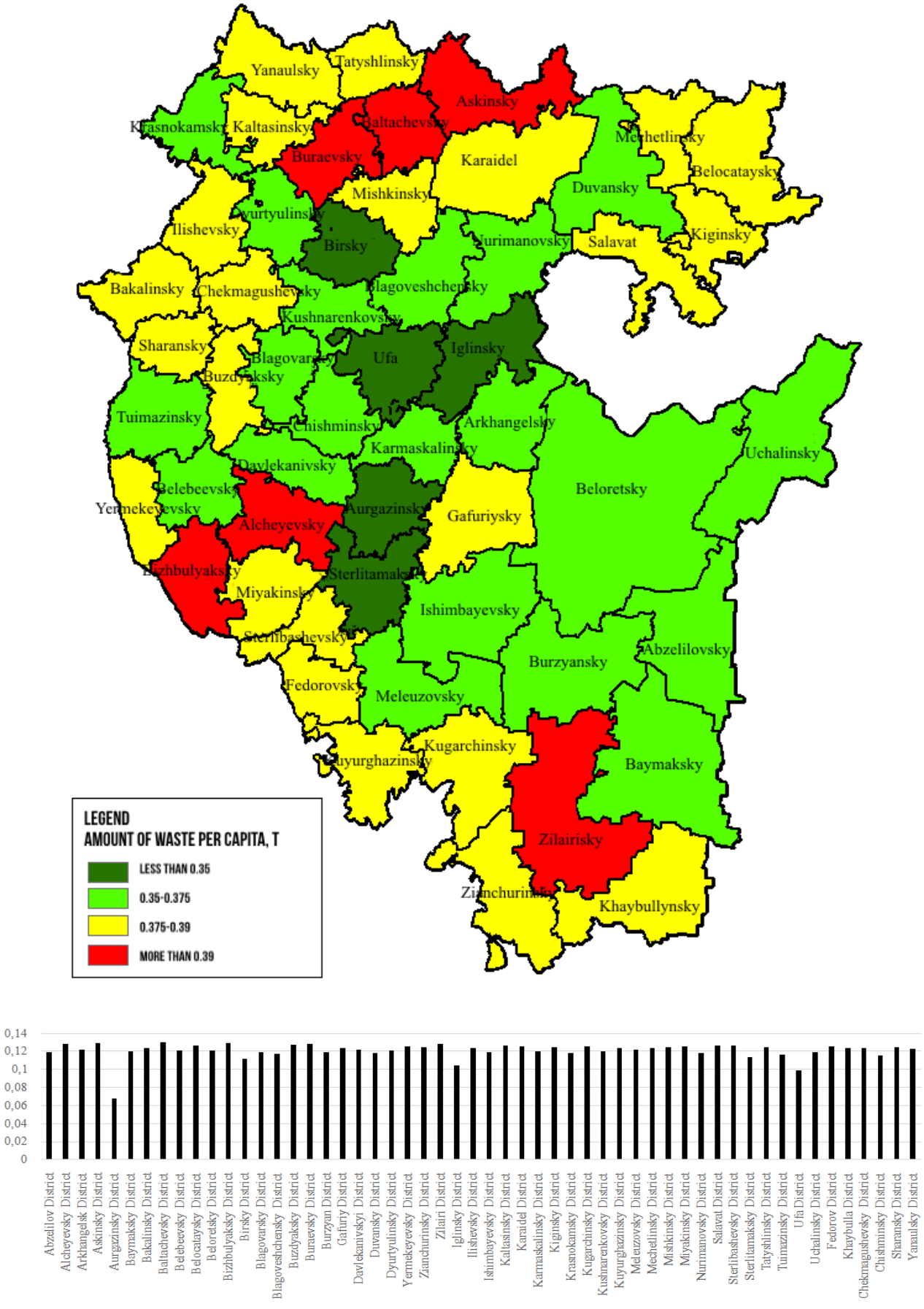


Fig 1. Amount of waste generated per capita in the Republic of Bashkortostan for 2015

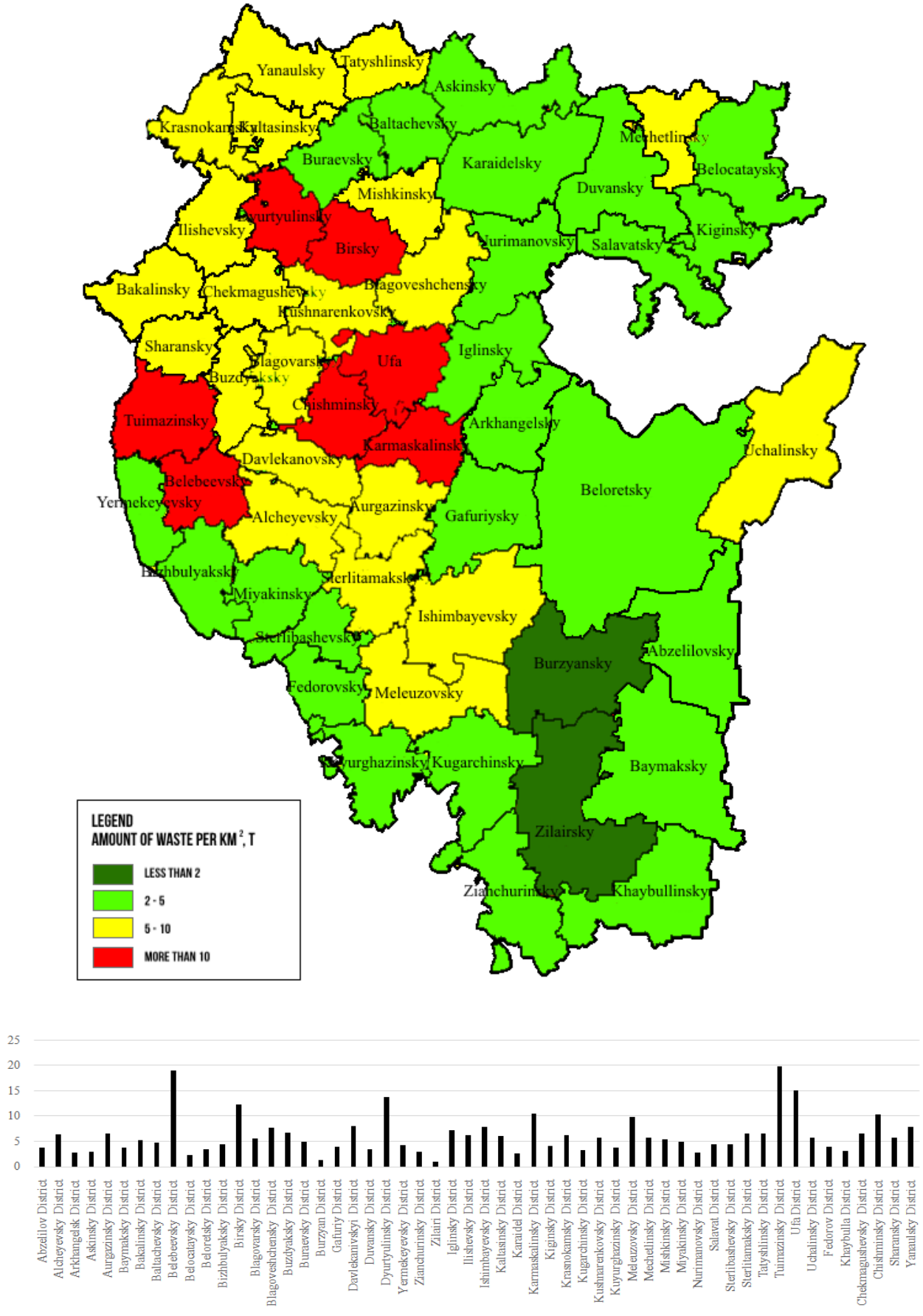


Fig 2. Amount of waste generated per 1 km^2 in the territory of the Republic of Bashkortostan for 2015

From this model it follows that the quality of the mosaic $\tilde{S}(x,y)$ is determined by the closer score $\tilde{F}(z)$ to the distribution law $F(z)$, which corresponds to sample $Z(x_i, y_i)$.

The generalized scheme for classifying areas of the territorial system from the sample data is based on the following:

- All information on the state of the territorial system is contained in the sample data, with the estimate $\tilde{F}(z)$ being an exhaustive (regarding the parameter z) state characteristic;
- The estimation of the spatial mosaic $\tilde{S}(x,y)$ is reduced to the division of the territory into areas, each of the areas being related to one of the classes of state. Sites belonging to the same class are assigned the same value of the relief characteristic, by which the integral estimate $F_{\tilde{S}}(z)$ can be constructed. Thus, the number of state classes and the boundaries of disjoint classes must be chosen in such a way that the following condition is satisfied:

$$P_j\left[\tilde{F}(z), F_{\tilde{S}}(z)\right] \to \min, \qquad (5)$$

Where P_j is j-th comparison rule for $\tilde{F}(z)$ and $F_{\tilde{S}}(z)$.

Assuming that the sample values $Z(x_i, y_i), i = \overline{i, N}$ are random variables to which the distribution function $\tilde{F}(z)$ corresponds, for formalization of the classification scale procedure it is rational to use a grouping approach to statistical data. In [14] various approaches to the construction of $\tilde{F}(z)$ estimates from samples of finite (including small) volume are described, as well as different approaches to solving the problem (5). It also describes the engineering methodology that allows to completely formalize the procedure for classifying a territorial system based on homogeneous samples $Z(x_i, y_i), i = \overline{i, N}$.

Approaches to solving the problem of classifying areas of the territorial system by state indicators should be linked to the management levels: thematic, and operational. In this connection, it is expedient to introduce the concepts of "absolute", "relative" classification scales. The purpose of constructing a "relative" classification scale is to present, in a visual cartographic form, the current (operational) situation, corresponding to the l-th time slice. In this case, the data, $Z^{(l)}(x_i, y_i), i = \overline{i, N}$ corresponding to the l-th research conditions, serve as sampling elements.

The goal of constructing an "absolute" classification scale is to create a basis for comparing the state of a territorial system in different time slices.

The absolute scale is formed on the basis of data corresponding to a set of time slices and is designed to compare the states of territories in different time slices $(l = 1,2,..., h)$. As sample elements in this case, serves the combined sample:

$$\{z\} = \bigcup_{l=1}^{h} z^{(l)}(x_i, y_i), i = \overline{1; N}. \qquad (6)$$

Limitations of this approach are the need to observe the conditions of homogeneity of sample data acquired at the nodes of the mesh measurement.

The results of the research presented in the article are partially supported by the grant 18-08-00885 - A "Methodological foundations of multi-criteria management of the process of choosing the location of industrial enterprises for processing waste based on the provisions of the power industry."

References:

1. Ayvazyan SA, Buchstaber VM, Enukov IS, Meshalkin LD Applied statistics. Classification and reduction of dimensionality. - Moscow: Finance and Statistics, 1989. - 607 p.
2. Berlyant A.M. Cartographic method of investigation. - Moscow: Moscow State University, 1988. - 252 p.
3. Berlyant A.M. Image space: map and information. -M .: Thought, 1986. - 254 p.
4. Bocharov M.K. Methods of mathematical statistics in geography. - Moscow: Thought, 1971. - 371 p.
5. Viktorov A.S. Picture of the landscape. - Moscow: Thought, 1986. - 179 p.
6. Gig J. Wang. Applied general theory of systems. - Moscow: Mir, 1981. - Book. 1. - 341 p., Prince. 2 - 730 p.
7. Gubanov VA, Zakharov V.V., Kovalenkov A.N. Introduction to system analysis. - Leningrad: Publishing House of Leningrad State University, 1988. - 232 pp.
8. Davison M. Multidimensional scaling: Methods of visual presentation of data / Trans. with English. VS S. Kamensky. - Moscow: Finance and Statistics, 1988. - 254 p.
9. Milkov F.N. Physical geography: the current state, patterns, problems. - Voronezh, 1981. - 136 p.
10. Mikhailov N.I. Physico-geographical zoning. - Moscow: Moscow State University, 1985. - 184 p.
11. Aldenderer MS, Blashfield R.K. Cluster analysis // Factor discriminant and cluster analysis, Ed. Enyukova N.S. - Moscow: Finance and Statistics, 1989. - pp. 139 - 210.
12. Rosenberg GS, Shitikov VK, Brusilovsky PM Ecological forecasting (functional predictors of time series). - Togliatti, 1994. - 182 p.
13. Fedina A.E. Physico-geographical zoning. - Moscow: MSU, 1981. - 128 pp.
14. Guzairov MB, Gvozdev VE, Ilyasov BJ. G., Kolodenkova A. Ye. Statistical study of territorial systems: monograph. - Moscow: Mechanical Engineering, 2008. - 187 p.
15. Christodoulo O.I. Development of an information system for the location of technogenic hazards with the use of fuzzy logic / O.I. Christodoulo, I.F. Salimzyanov, N.R. Gareyeva // Scientific journal "Scientific and technical statements of SPbSPU. Computer science. Telecommunications. Control". - St. Petersburg, Russia, 2015. Vol.5 (229). Pp. 47 - 58.
16. Khristodulo O., Davletbakova Z., Gvozdev V. Spatial Information Processing for Decision-making Supporting of Siting Sources of Technogenic Hazards Using Computer Technologies // 2nd International Conference on Industrial Engineering, Applications and Manufacturing (ICIEAM). - Chelyabinsk, 2017. P. 1-5.
17. Gvozdev VE, Khristodulo OI, Fakhretdinova E.B. Information support of waste management on the basis of cognitive, geoinformation and mathematical modeling // Proceedings of the International Scientific and Technical Conference Perspective information technologies, Samara, 2018 - P. 1110 - 1114.

MACHINE LEARNING APPLICATIONS FOR DESIGN OF NEW MATERIALS: A REVIEW

Sr. Lecturer Kumar S. PhD., Prof. Nezhurina I. M. PhD.
College of Information Business Systems – National University of Science and Technology, MISiS
Leninsky Ave. 4, 119049, Moscow, Russian Federation

sachinagnihotri16@gmail.com

Abstract: *The importance of materials in our life is well known to the world. Material science is one of the crucial research domain which deals with planning, development, synthesis and analysis of materials and its use in the real applications. The scientists and researcher are using highly advance technologies for the production and design of new materials in present era; however the process is time consuming. Machine Learning can be defined as a set of techniques that learn from the large amount of available data and make predictions for the new data. Recently, machine learning has gained a great attraction in material science research and has been used in several research studies. Machine learning techniques have the potential to learn from the large amount of materials data and make predictions about different properties for new materials of good quality. These predictions can help in designing new materials. In this paper, we present a comprehensive review on the machine learning applications in different material science domain. We believe that the paper would be useful for the researchers, academicians and students who are involved in the design of new materials.*

Keywords: MACHINE LEARNING, MATERIAL SCIENCE, MATERIALS DESIGN, CLASSIFICATION AND PREDICTION, CLUSTERING

1. Introduction

Material science is an important research domain which is highly involved in planning, designing, synthesis and development of materials. The design of new materials is one of the hot topics at present among material scientists [1]. Scientists and researchers are working hard to design new materials with better efficiency, quality and life expectancy. Initially, the process of designing new materials involved 10-20 years from initial scratch to the production of materials. Fig. 1 illustrates the steps used in the traditional process of designing new materials. Present world is equipped with high tech equipment, world class laboratories and computationally efficient methods for the design and discovery of new materials. Hence, the process of developing new materials achieved some advancement in terms of time spent for analysis and processing of materials.

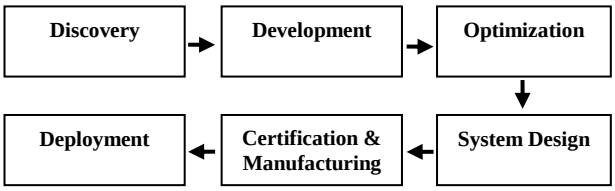


Figure 1: *Traditional steps in materials design*

The two most common approaches that have been widely used from decades in the material science research are laboratory experiments and computer simulations. In material science, laboratory experiments usually involved the study of microscopic structure of a material, analysis, synthesis and measurement of materials. All these required the use of specific instrument which is available in the laboratory only; however these processes are bound to certain temperature and pressure conditions and required a specific amount of time to complete the procedure along with expertise of the researcher involved in the study. Therefore, the importance of laboratory cannot be ignored. Computer simulation is an alternative approach in which a computer program is written to exploit the existing theoretical concepts to calculate electronics structure of the materials at macroscopic and macroscopic level. Density functional theory [2], Monte Carlo simulations [3] and Molecular dynamics [4] are the popular approaches on which computer simulations relied.

There is no doubt that computer simulations are much powerful and efficient in terms of processing time and provides full control over variables which is quite difficult to achieve with experimental analysis. However, certain challenges are still exist before computer simulation methods as well [5]. One of the issues is the dependency of microstructure of the materials. Second, these simulations are computationally expensive and require high computing clusters to run the simulations and achieve results on time which is quite expensive. Also, the previous simulations outcome does not help if a new system is introduced for study. Present demand in material science research needs a close collaboration between experimental and computational methods to provide fundamental facts about structure and properties of the materials to the scientific world. It is possible to conduct some of the experiments using computational tools due to rich technologies in present world and it has reduced the time spend on experiments from 10-20 years to 14-18 months [6]. Although, both experimental and computational methods has certain limitations that is not possible to overcome such as it is often difficult and not recommended to perform some experimental procedures over a wide temperature limit (for glass). Similarly, some the computational simulation is not possible exactly in the required form as they might be dependent on a variety of internal and external factors i.e. pressure, conformational features and structure [7].

In order to overcome these problems, several attempts have been made. MGI (Materials Genome Initiative) [8] is launched in 2011 which along with big data community provides a means to generate and store large amount of data related to material science. These initiatives made it possible for the engineers, researchers and scientists to access the properties of known materials. Several databases for materials properties are the results of these initiatives. These databases [9-11] are Inorganic Crystal Structure Database (ICSD), Super Conducting Critical Temperatures (SuperCon), Open Quantum Materials Database (OQMD) etc. Once we are equipped with huge amount of materials properties, the next task is to do some analysis to understand it and use this knowledge for decision making. The question is: How to get this information?

Machine learning [12] is a well-known powerful set of techniques that has shown its potential in different real world domains including material science. Machine learning methods enable a computer to learn from the huge amount of data and do predictions for unseen data. Recently, machine learning and big data methods have been effectively used to resolve complex issues in modeling the materials. The aim of the present study is to provide a review of the effectiveness of machine learning methods in the different domains of material science. In the next section, a brief introduction of machine learning with respect to material science is provided. In Section 3, application of machine learning in material's property prediction is provided. In Section 4, we discussed the machine learning applicability in molecular dynamics aspect of

material science. Finally, the most relevant conclusions are provided in conclusion section.

2. Machine Learning in Material Science

Fig. 2 illustrates the general idea of the knowledge discovery process in material science using machine learning. It shows that materials properties data can be generated through laboratory experiments and computer simulations. Next, data preprocessing methods should be performed to select the relevant materials properties and cleaning the data. Further, machine learning can be applied to develop a prediction model. This model can be used to predict the desired properties for unseen data of new materials properties.

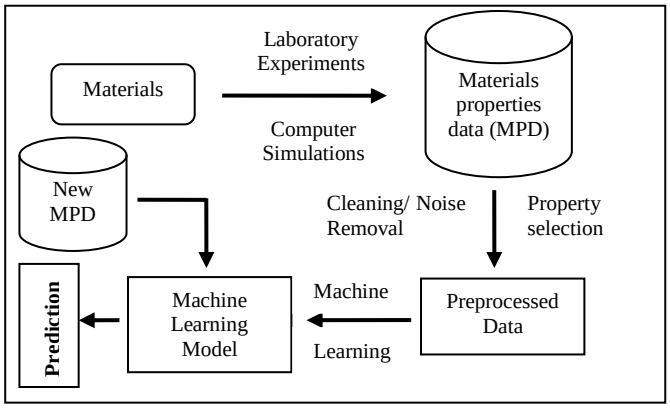


Figure 2: *Knowledge Discovery Process using Machine Learning in Material Science*

The crucial steps in knowledge discovery process for materials data analysis are data preparation and selection of suitable machine learning algorithm for model development. The properties of materials for the analysis must be efficiently done because not all properties are useful for specific analysis. Hence, the selection must be accurate. The next important task is to select the machine learning algorithm. This selection is based on the nature of the task to be performed and the features in the data set. Therefore, this has to be done carefully.

Machine learning has already shown its potential in material science both in terms of prediction accuracy and time [13]. In material science, machine learning can be used for property prediction (both macroscopic and microscopic level), discovery of new materials and force field prediction at atomic level for materials.

3. Materials Property Prediction using Machine Learning

Computer simulation and laboratory experiments are the two common approaches to study the materials properties. Melting point of a material, glass transition temperature, ionic conductivity, lattice constant and molecular atomization energy are some of the properties of materials which can be described both on macroscopic and microscopic levels. Although computational simulation and laboratory experiments are efficient enough to study the materials properties but if the results are not accurate as desirable then the time and resources spent in computation and laboratory experiments are wasted [14]. Moreover, in some cases it is very difficult and almost impossible to study the properties even with complex computer simulations and extensive efforts in laboratory experiments. Therefore, it is highly desirable to develop a system to understand the properties of materials with high efficiency and low computational cost.

Machine learning is a constructive way that enables the computer to learn from the huge amount of materials properties and create a linear or non-linear mapping (as applicable) between materials properties and related factors. Further, this model can be used for prediction of properties for new materials. The problem of materials property prediction can be divided in to macroscopic and microscopic level. At the macroscopic level, specifically we are interested in mechanical and physical properties of materials and its relationship with microstructure of materials [15]. Machine learning algorithms such as Support Vector Machine (SVM), Regression and Artificial Neural Network (ANN) with optimization algorithms is widely used for the study and analysis of macroscopic properties of materials. One of the popular versions of ANN is back propagation ANN that has been used to capture the relationship between materials properties and external factors [16]. Previously, few attempts have been made to predict the materials properties such as temperature responses, elongation, corrosion etc. using back propagation ANN [16-17]. Although back propagation ANN is an effective method to be used for property prediction in material science, slow convergence rate and local minima is a problem of concern.

Another version of ANN without such problems of local minima and low convergence rate is RBF-ANN (Redial Basis Function-ANN). A successful attempt with RBF-ANN has been made to investigate the crack propagation has been made in a study on bituminous layered pavement structure [18]. In another study [19], prediction of melting point of a material is investigated using ANN. There are other studies as well which tried to develop prediction models using ANN for excited state energy prediction [20], diffusion barriers [21] etc. Another aspect of ANN that must be taken into consideration is that ANN provides better results if the size of the data is relatively large enough. For relatively small samples, ANN results may not be reliable. Support vector machine (SVM) is another machine learning methods for data analysis and pattern recognition in the data. With a relatively small sample of data, SVM can be used for reliable results as it can efficiently handle large dimensions and over fitting problem. An approach based on Genetic algorithm (GA) and (Support Vector Regression) SVR has been proposed [22] with an objective to forecast the corrosion behaviors of zinc and steel. SVM has shown its potential in several domains and can be deployed for ionic conductivities prediction, glass transition temperature prediction [23-24].

On the other hand, microscopic properties of a material such as electronics crystal structure, lattice constant, band energy, molecular atomization energy etc. are also one of the important aspects in design of new materials. Logistic Regression (LR), SVR and ANN have been deployed for prediction of lattice constant in various studies [25-28]. In these studies, it was found that ANN has a better accuracy and performance than LR method; however the accuracy for training data was better than the accuracy on test data. Further, SVR found to achieve better accuracy and performance than ANN for lattice constant prediction. Also, SVR has the better learning and testing efficiency that ANN for the smaller datasets. In order to improve the accuracy of machine learning techniques, Ward et al. [29] used the concept of ensemble learning methods. They suggested that ensemble learning can overcome the shortcomings of machine learning algorithms and can provide improved efficiency and accuracy.

Therefore, it is evident that machine learning algorithms has played a big role in providing more accurate materials property predictions at both macroscopic and microscopic levels. However, not all machine learning methods are suitable in all cases. If the size of data sample of materials properties is not larger enough, the SVR performs better than ANN. On the other hand, if the size of the sample is high than ANN and RBF-ANN can be used to develop more effective and reliable prediction models.

4. Atomic Force Field Prediction using Machine Learning

Quantum based mechanical approaches also known as Ab-initio methods or first principle quantum calculations are very demanding at present in materials modeling. Based on first principles quantum mechanical simulations, a benchmark data set can be created which

consists of positions of the atoms (atomic configurations) and the corresponding forces working on them. This benchmark data set can be utilized to investigate the relationship between atoms positions and the potential energy using machine learning methods [30]. Many active research groups in the field have faith in machine learning that it can lead a pathway to create interatomic potentials with lower computational cost than quantum mechanical methods, with better accuracy and versatility than conventional interatomic potentials and provides better efficiency [31].

Zhenwei et al. [32] presented a machine learning approach along with first principle quantum calculations to perform molecular dynamic simulations. They mentioned that their proposed scheme is more accurate and takes lesser quantum mechanical calls as required in repeatedly encountered new chemical process. In recent studies, it is realized that there is a vectorial force experienced by every atom in the system, which can be learned from the given positions of the atom and can be predicted further for new positions [33, 34]. This approach is more appealing as the conventional methods tend to predict the potential energy of the system which is a global component of the systems, whereas atomic force on an atom is a local component which can be determined by the surrounding environment of the system.

Furthermore, Gaussian models, neural networks have already shown their potentials in the development of machine learning based models for molecular dynamic simulations [32-36]. Huan et al. [33] proposed a universal strategy to generate machine learning based atomistic force field in order to perform more accurate and powerful molecular dynamic simulations. They proposed to first create a dataset with low noise which can be obtained from density functional theory methods, second to create fingerprints by numerically representing the atomic configurations and corresponding forces. Next, select the optimal and non-redundant datasets from reference data for training purpose. Further, different prediction algorithms can be used to learn from the data and compute atomic forces directly from atomic configurations.

Assessment of Intra and inter-molecular energies can be done with quantum mechanical simulations which depend on the accuracy and preciseness of atomic force field. Friederich et al. [37] mentioned that non-local association between dihedral potentials which has been ignored in many state of the art dihedral force fields, has an important role to play in the description of total molecular energy. They used ANN to study the degrees of freedom of atoms within molecules of complex organic materials. Their results were quite promising in stating that multilayer ANN can provide better predictive results than the conventional approaches.

Therefore, it is worth saying that machine learning is being proved an important approach which has potential to develop fore field models which can perform better than conventional computational simulations.

4. Machine Learning: Is there any complexities?

There is no doubt that machine learning potentials in material science came up with several advantages, however certain complexities associated with machine learning potentials must be taken into consideration. One of the crucial factors is the preparation of the data that is the basic need for a machine learning task. So, most of the efforts must be done with preparation phase as if the data is not well prepared the machine learning will no longer be able to provide the reliable results. Furthermore, in material science expert knowledge is very important in preparation of the data i.e. selection of appropriate properties of materials that can be utilized in property prediction task. In atomic force field creation, fingerprinting is very important task in which atomic configurations and forces must be numerically presented for better analysis. Therefore, expert knowledge is required for effective data preparation.

Once the preparation of data is done effectively then the next important step is to select the efficient machine learning algorithms to develop prediction model. As shown in earlier sections, ANN is a reliable approach for materials property prediction but SVR performs better when the size of the data is relatively small [22, 23]. Furthermore, if the size of the data is relatively large that it is a better option to with ANN and its other version such as RBF-ANN. Multi-layered ANN also known as deep neural networks has been used recently in some studies related to material science. Mills et al. [38] used convolutional deep neural network to predict the ground state energy of electrons in two dimensional electrostatic potentials and the accuracy obtained are within chemical boundaries. In a study, Ryczko et al. [39] proposed an approach to compute the total potential energy for the atomic systems. They used deep convolutional neural networks and named this method as CNNAS (convolutional neural network for atomic systems). They mentioned that their method is able to predict energies for different types of 2D hexagonal lattices with a great accuracy. Therefore, if you have a large database of materials properties or DFT calculations, you may choose the deep neural networks as it is able to achieve a higher level of accuracy.

5. Conclusion

This study attempts to provide an understanding about the machine learning applicability in the materials science to the students, researchers and academicians engaged in the design of novel materials. The study took into consideration the traditional approaches such as experimental analysis in the laboratory and the computer simulation on powerful computers for the study of materials. Next, the issues associated with these conventional approaches such as their limitations and the computational cost etc. is discussed that required a new method to be developed to overcome these limitations. Machine learning is already being used by material scientists and engineers to study and investigate the materials. Different machine learning methods such as Genetic algorithms, Gaussian model, SVM, SVR, ANN and deep ANN has already been deployed to develop prediction models for different materials. Although, machine learning is contributing well in design and discovery of new materials; yet more extensive research is needed to fully understand its applicability in design of new materials with better quality.

Acknowledgement

The authors gratefully acknowledge the financial support of the Ministry of Education and Science of the Russian Federation in the framework of Increase Competitiveness Program of **NUST «MISiS» (№ К4-2017-052).**

References

1. About J., Materials genome initiative for global competitiveness, National Science and Technology Council, USA, 2011.
2. Hohenberg P., W. Kohn, Inhomogeneous electron gas, Phys Rev, 136, 1964, 864-871
3. Binder K., A. Baumgartner, The Monte Carlo method in condensed matter physics, 71, 1995.
4. Alder B. J., T. E. Wainwright, Studies in molecular dynamics. I. General method, J Chem Phys, 31 (2), 1959, 459-466.
5. Rajan K., Materials informatics, Mater TodCay, 8, 2005, pp. 38-45.
6. Hautier G., A. Jain, S. P. Ong, From the computer to the laboratory: materials discovery and design using first-principles calculations, J mater. Sci, 47, 2012, 7317-7340.

7. Yu X. L., B. Yi, X. Y. Wang, Prediction of the glass transition temperature for polymers with artificial neural network, J Theor Comput Chem, 7, 2008, 953-963.
8. Materials Genome Initiative: https://www.mgi.gov accessed on 30.07.2018.
9. Belsky A., M. Hellenbrandt, V. L. Karen, P. Luksch, New developments in the inorganic crystal structure database (ICSD): accessibility in support of materials research and design, Acta Crystallogr, Sect B Struct Sci, 58, 2002, 364-369
10. Super Conducting Critical Temperatures (SuperCon): https://supercon.nims.go.jp accessed on 30.07.18
11. Kirkli S., J. E. Saal, B. Meredig, A. Thompson, J. W. Doak, M. Aykol, The open quantum materials database (OQMD): assessing the accuracy of DFT formation energies, npj Comput Mater, 1, 2015, 15010.
12. Mueller T., A.G. Kusne, R. Ramprasad, Machine learning in materials science: recent progress and emerging applications, Rev Comput Chem, 29, 2016, 186.
13. Yue L., Z. Tianlu, J. Wangwei, S. Siqi, Materials discovery and design using machine learning, J Materiomics, 3, 2017, 159-177.
14. Ning X., M. Walters, G. Karypisxy, Improved machine learning models for predicting selective compounds, J Chem Inf Model, 52, 2012, 38-50.
15. Bertinetto C., C. Duce, A. Micheli, R. Solaro, A. Starita, M. R. Tine, Evaluation of hierarchical structured representations for QSPR studies of small molecules and polymers by recursive neural networks, J Mol Graph Model, 27, 2009, 797-802.
16. Guo Z., S. Malinov, W. Sha, Modelling beta transus temperature of titanium alloys using artificial neural network, Comput Mater. Sci, 32, 2005, 1-12.
17. Topçu B., M. Sarıdemir, Prediction of properties of waste AAC aggregate concrete using artificial neural network, Comput Mater Sci, 41, 2007, 117-125.
18. Altun F., Özgür Kişi, K. Aydin, Predicting the compressive strength of steel fiber added lightweight concrete using neural network, Comput Mater. Sci, 42, 2008, 259-265.
19. Gajewski J., T. Sadowski, Sensitivity analysis of crack propagation in pavement bituminous layered structures using a hybrid system integrating artificial neural networks and finite element method, Comput Mater Sci, 82, 2014, 114-117.
20. Häse F. , S. Valleau, E. Pyzer-Knapp, A. Aspuru-Guzik, Machine learning exciton dynamics, Chem Sci, 7 (8), 2016, 5139-5147.
21. Wu H., A. Lorenson, B. Anderson, L. Witteman, H. T. Wu, B. Meredig et al., Robust FCC solute diffusion predictions from ab-initio machine learning methods, Comput Mater Sci, 134, 2017, 160-165.
22. Fang S. F., M.P. Wang, W.H. Qi, F. Zheng, Hybrid genetic algorithms and support vector regression in forecasting atmospheric corrosion of metallic materials, Comput Mater Sci, 44, 2008, 647-655.
23. Liu X., W.C. Lu, C.R. Peng, Q. Su, J. Guo, Two semi-empirical approaches for the prediction of oxide ionic conductivities in ABO_3 perovskites, Comput Mater Sci, 46, 2009, 860-868.
24. Ahmad A., A. Amjad, M. Denny, C. A. S. Bergström, Experimental and computational prediction of glass transition temperature of drugs, J Chem Inf Model, 54, 2014, 3396-3403.
25. Javed S. G., A. Khan, A. Majid, A. M. Mirza, J. Bashir, Lattice constant prediction of orthorhombic ABO_3perovskites using support vector machines, Comput Mater. Sci, 39, 2007, 627-634.
26. Li C. H., Thing YH, Zeng YZ, Wang CM, Wu P. Prediction of lattice constant in perovskites of $GdFeO_3$ structure. J Phys Chem Solids, 64, 2003, 2147-2156.
27. Majid A., A. Khan, T. Choi, Predicting lattice constant of complex cubic perovskites using computational intelligence, Comput Mater Sci, 50, 2011, 1879-1888
28. Majid A., A. Khan, G. Javed, A.M. Mirza, Lattice constant prediction of cubic and monoclinic perovskites using neural networks and support vector regression, Comput Mater Sci, 50, 2010, 363-372.
29. Ward L., A. Agrawal, A. Choudhary, C.Wolverton, A general purpose machine learning framework for predicting properties of inorganic materials, npj Comput Mater, 2, 2016, 16028.
30. Hastie T. R. Tibshirani, J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed.; Springer: New York, 2009.
31. Behler, J. Atom-centered symmetry functions for constructing high-dimensional neural network potentials. J. Chem. Phys. 2011, 134, 074106.
32. Li Z., J. R. Kermode, A. De Vita, Molecular Dynamics with On the-Fly Machine Learning of Quantum-Mechanical Forces. Phys. Rev. Lett. 2015, 114, 096405.
33. Huan T. D., R. Batra, J. Chapman, S. Krishnan, L. Chen & R. Ramprasad, A Universal strategy for the creation of machine learning-based atomistic force fields, npj Computational Materials 3, 37, 2017.
34. Botu V. R. Ramprasad, Learning scheme to predict atomic forces and accelerate materials simulations. Phys. Rev. B: Condens. Matter Mater. Phys. 92, 2015, 094306.
35. Behler, J., M. Parrinello, Generalized neural-network representation of high-dimensional potential-energy surfaces. Phys. Rev. Lett. 98, 2007, 146401.
36. Bartók A. P., G. Csányi, Gaussian approximation potentials: a brief tutorial introduction. Int. J. Quant. Chem. 115, 2015, 1051–1057.
37. Friederich P., M. Konrad, T. Strunk, W. Wenzel, Machine learning of correlated dihederal potentials for atomistic molecular force fields, Scientific Reports 8, 2018, 2559.
38. Mills K., M. Spanner, I. Tamblyn, Deep learning and the Schrodinger equation, Phys. Rev. A, 96, 042113
39. Ryczko K., K. Mills, I. Luchak, C. Homenick, I. Tamblyn, Convolutional neural networks for atomistic systems, Comput Mater Sci, 149, 2018, 134-142.

DYNAMIC COGNITIVE MAPS FOR THE SUBSTANTIATION OF STRATEGIC DECISIONS ON MANAGEMENT OF ENERGY SECTOR DEVELOPMENT

A.G. Massel
Laboratory of Information Systems in energetics, Melentiev Energy Systems Institute of SB RAS, Irkutsk, Russia, e-mail: amassel@gmail.com

Abstract[1]: *This article describes the tools of intelligent decision-making support in the research of energy security problems using the cognitive approach. The application of dynamic cognitive modeling and the notion of dynamic cognitive modeling are described. The author proposes the future directions of this approach developmen.*

Introduction

The task of situational analysis is to identify the parameters and significant factors, or circumstances that determine the situation, the relationship between the factors and the character of their mutual influence. A situation is understood as a set of circumstances that determine the internal state of an object or system, and the circumstances that determine the state of the environment in relation to a given object or system. The first are described by parameters characterizing the state of the system, the second - the environmental conditions or significant factors affecting on the development of the system [1,2]. Situational analysis includes: analysis of problem situations (for example, emergency situations in power engineering); identification of ways to resolve problem situations (alternatives); the definition of criteria for the evaluating alternatives (for example, economic); analysis of alternatives; choice and implementation of the best alternative [3,4]. Given that the presence of uncertainties complicates the adequate state assessment of the object and environment, the author suggested to use the semantic technologies of situational analysis, which include ontological, cognitive and event modeling, and probabilistic simulations using Bayesian networks of trust.

Cognitive modeling is the construction of cognitive models, or, in other words, cognitive maps (oriented graphs) in which vertices correspond to factors (concepts), and arcs describe the relationships between factors (positive or negative), depending from the nature of the cause-effect relationship. The mathematical apparatus for constructing of cognitive models is the graphs theory.

The foundations of cognitive modeling were developed in due time by Wang Hao (1956), Axelrod (1976), D.A. Pospelov (1981), this work was continued by E.A. Trakhtengerts [5], and is actively developing in the Institute of Control Problems of the Russian Academy of Sciences (Abramova N.A., Maksimov V.I., etc.) for describing of weakly structured situations and for effect analysis in weakly structured situations management. Ideas of development and application of fuzzy cognitive maps are actively developing in Greece (Groumpos P.P., etc.) [6-8]

It should be noted the relevance of the new direction related to 3D-visualization, called "Situational Awareness". The concept of Situational Awareness was formed at the turn of the 1990s and is associated primarily with the pioneering work of Mica R. Endsley [9]. It is proposed to joint semantic modeling tools and situational awareness within the OODA cycle (proposed by D. Boyd) [10] for the purposes of situational management.

MESI SB RAS has traditionally conducted research on energy security (ES) problems. Currently, the research uses primarily a quantitative approach to assessing the ES level, supported by the traditional software systems. The use of these software systems, as a rule, requires a lot of time for the preparation of information, the formation and updating of information models, the determination of emergency situations (EmS) and the choice of computational experiments strategy.

To reduce expert burden, the team in which the author works offers a two-level technology of research, supported by an intelligent IT environment in which qualitative analysis is performed at the first level using ontological, cognitive and event modeling [11,12]. Based on the results of qualitative analysis, decision-makers select variants requiring a detailed justification, these variants are calculated at the second level.

This approach is currently being applied in MESI SB RAS to substantiate strategic decisions on energy development management. One such work was carried out at the request of the Russian Energy Agency to solve the problem of logical modeling of the improvement / deterioration of the situations with ensuring of the country's energy security at the federal level. In the course of the work, some shortcomings of cognitive modeling in the classical sense were noted, which can be eliminated using the author's proposed dynamic cognitive maps.

1. Intelligent IT-environment

The concept of an intelligent IT-environment supporting a two-level technology for research into the directions of the development of the fuel and energy complex, taking into account the requirements of the ES [13], is proposed. The intelligent IT environment is defined as

$V_{IT} = \{O, E, M_C, M_S\} \cup T_V$, where $\{O\}$ - the set of ontologies, $\{E\}$ - the set of emergency precedents descriptions, $\{M_C\}$ - the set of cognitive models, $\{M_S\}$ - the set of event models, T_V - the tools for supporting the IT-environment, including the description of knowledge presented

in the form of ontologies, descriptions of emergency precedents, cognitive and event models and manipulating tools of them.

Thus, the intelligent IT environment includes a knowledge space that integrates: ontological models of knowledge in the field of ES research, the knowledge base of emergencies precedents in energy sector and knowledge bases containing cognitive models of ES strategic threats and event models of the development and consequences of emergencies in energy, and tools for describing knowledge and manipulating them (Figure 1). The joint use of ontological, cognitive and event modeling gives the expert a more complete representation of the extreme situation in the energy sector.

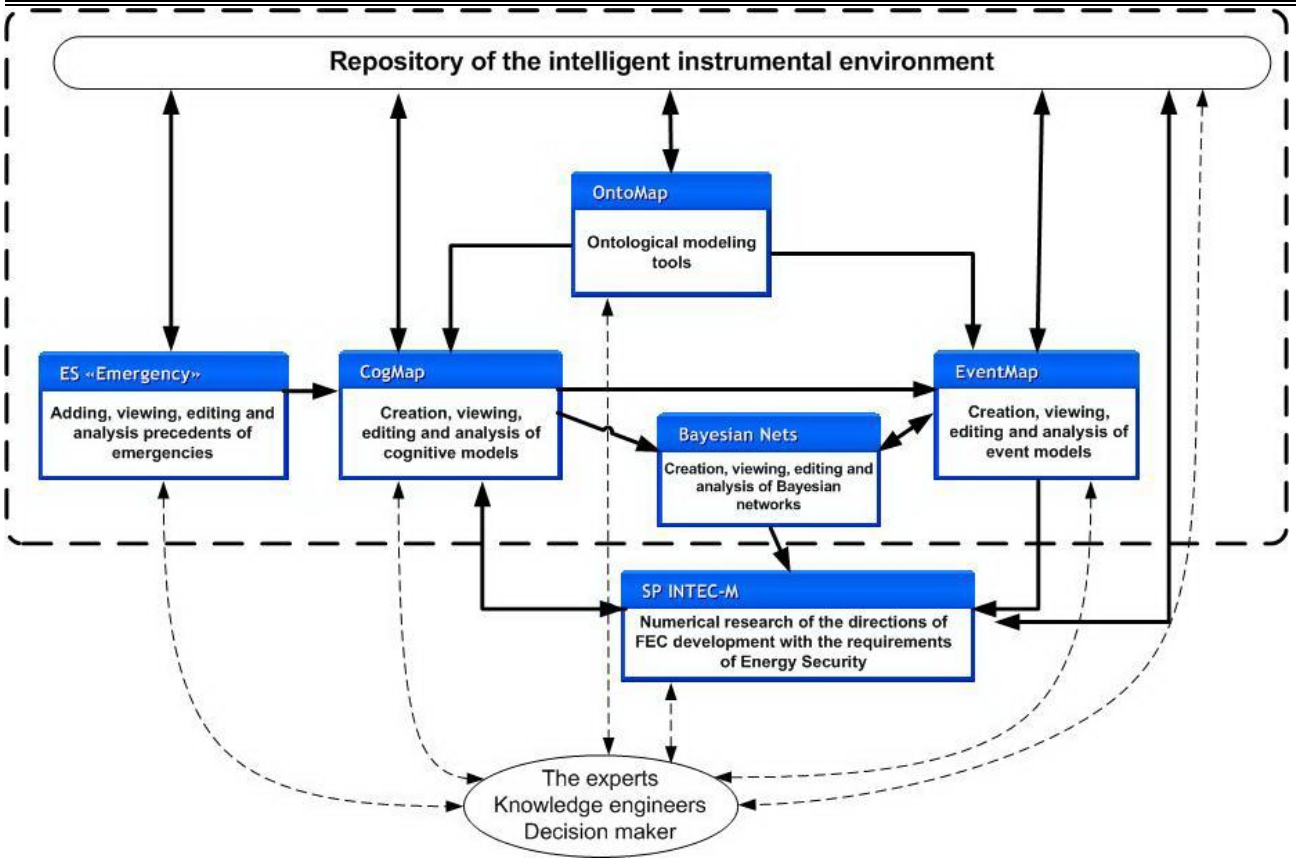


Figure 1 - Interconnection of the tools in the intelligent IT-environment

2. Cognitive approach to research on ES problems.

Cognitive modeling is the construction of cognitive models, or, in other words, cognitive maps (oriented graphs) in which vertices correspond to factors (concepts), and arcs describe the relationships between factors (positive or negative), depending from the nature of the cause-effect relationship. The mathematical apparatus for constructing of cognitive models is the graphs theory. Most often, fuzzy cognitive maps are used, and in the simplest case, the connection weight is defined as +1 or -1.

Cognitive modeling is considered, in particular, in [13, 14] and is used primarily for the analysis of socio-economic situations. The author proposed to use the cognitive approach in research of ES problems [15]. This approach makes it possible to obtain scenarios for the sustainable and crisis development of the energy sector in the region, to identify factors that affect on the scenarios of the energy system development, and to develop plans of counteract to the ES threats. Scenario programming was chosen as a methodology for modeling the region's energy system, from a mathematical point of view it is a logical and probabilistic extension of the cognitive map system [15].

We illustrate cognitive modeling on the example of the natural ES threat, namely the threat of "Coldness" (Fig. 2).

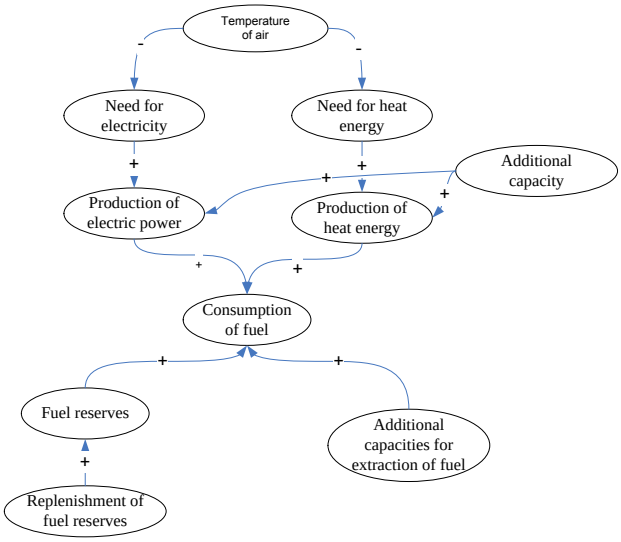


Figure 2 - Cognitive map of the natural threat "Coldness"

On the cognitive map of the natural threat "Coldness", the main factors influencing the decision-making of the decision maker are revealed, factors – threats and factors – events are identified. This map shows the cause-effect relationship between the factors. The weight coefficients of the connections are assigned by experts, by default the weight coefficients are +1 or -1.

Cognitive maps, in turn, can not always give a complete, and most importantly unambiguously interpreted, view of the situation to the expert. To eliminate this shortcoming, the author proposes introducing the concept of dynamic cognitive maps.

3. Dynamic cognitive maps

The cognitive map is inherently static and at the same time reflects only the nature of the factors interconnection, and in the modeling of extreme situations in the energy sector, a dynamic representation is desirable. The author proposes to use dynamic cognitive maps for this goal.

In contrast to event modeling, aimed at analyzing the development and consequences of emergencies, dynamic cognitive maps retain the ability to identify the cause-effect relationships of concepts aimed at analyzing ES threats, described in conventional cognitive maps (CM). The difference from the fuzzy cognitive maps used in ES research earlier is that the expert will be able to see and evaluate the influence of connections relative to the time scale, and most importantly, in modeling, he can confine himself to considering individual factors of influence.

The dynamic cognitive map (DCM) differs from the conventional cognitive map in that it is a set of cognitive maps presented at different times, ie:

$\{D\} = \{D_{t_0}, D_{t_1} ... D_{t_n}\}$, where $\{D\}$ - a set of dynamic cognitive maps; $\{D_{t_0}\}, \{D_{t_1}\}, \{D_{t_n}\}$ - sets of cognitive maps at time instants t0, t1, t_n.

As a result, it becomes necessary to include in the DCC the weights of the relationships that would reflect the influence of factors (and / or objects) on each other as a function of time.

Then we can represent the bond weight as some function W (t). In most cases, the connection weights in DCM can be represented by simple functions.

Analyzing the existing knowledge base of cognitive models, in accordance with the methodology of cognitive modeling of ES threats [16], we distinguish several basic classes of factors for which we can determine the simplest cases of interaction: 1) the class of ES threats; 2) a class of events (preventive, operational and liquidation); 3) the class of energy objects.

Consider the simplest case, typical for a class of ES threats: the occurrence of an event. It can be represented in the form of a single impulse. This dependence is characterized by the fact that the period of occurrence and action of the event is much less or equal to the chosen time step.

The next type of connections and their weights is more complex, since the degree of its impact can continue for several steps, and for the simplest connections, the weight functions W(t) will have a linear form.

One of the problems in constructing DCM is the mapping of the mutual influence of various connections on the same concepts (factors). To determine the degree of interference, it is necessary to introduce for the factors a set of parameters (characteristics of factors) that can be influenced by links with other factors (for example, see the concept "Fuel Consumption" in Figure 2). Due to the fact that different factors can be included at different time periods, it becomes necessary to display implicit relationships.

An example of such implicit connections is the presence in the cognitive map of a preventive action and a threat that at the moment *t* develops into an EmS. The effective implementation of preventive measures can prevent the outbreak of the ES threat in EmS.

4. Fuzzy DCM (FDSM)

Based on the proposed approach, within the framework of two-level technology it is supposed to use cognitive maps for qualitative analysis of ES problems. In this case, for using of DCM the expert have to determine the weight of the connection and the speed of its influence on other factors in the CM. Therefore it is proposed to introduce a linguistic scale similar to used in fuzzy CM, with the difference that the introduced gradation will display the speed of weight changes and its effect on the selected factor.

If in a fuzzy cognitive map it is possible to determine the weight in the range "weakly-strong", then in FDCM communication can be determined in the "long-fast" range.

In other words, if we take t_n as the time of the final step, then the event that lasts the time interval $T \geq t_n / 2$, is a long event.

The next stage in the development of methods for constructing FDCM is their integration with various mathematical models. MESI SB RAS uses the software package INTEK-M [17], which is designed to support the computational experiment of research into the directions of the development of the fuel and energy complex, taking into account the requirements of energy security. These studies are considered as the basis for justifying of strategic decisions on energy development.

This software package INTEK-M uses the traditional TRV model (Figure 3), where:

T is a technology, R is a district and V is a vector that denotes different modes of operation for different objects.

After the construction, in accordance with the proposed methodology, of cognitive models, the user is given the opportunity to specify characteristics of factors (basic concepts) of the cognitive model. Further, these characteristics will be transformed into corrections of the mathematical model, which are transmitted to the INTEK-M in the form of a special file.

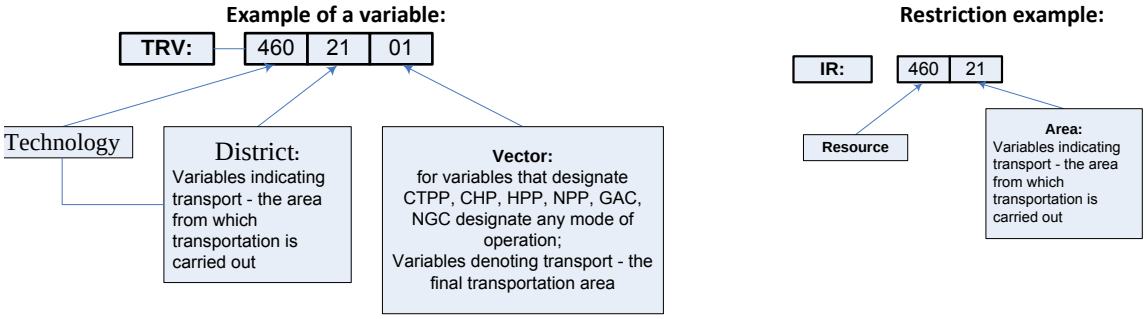


Figure 3 - Description of the TRV-model.

Until recently, traditional cognitive maps were used in energy research. The author proposes to extend their application by implementing two additional steps: 1) to reduce the expert's load when analyzing cognitive models by visualizing the development of cause-effect relations in dynamics; 2) to integrate semantic and mathematical models.

Then the general scheme of the tools interaction within the intelligent IT- environment will look like this (Figure 4).

To facilitate the construction of cognitive models of this type, it is necessary to add the directories of technologies, regions and vectors. The technology of connecting directories in the form of ontologies for constructing cognitive models has already been debugged and described in [18]. At this stage, work is underway to create an ontological knowledge space, the ontologies of which describe the main elements of the used TRV-model.

In addition to the use of directories, it is necessary to specify as DCM characteristics, and such information as the time step, various coefficients and limitations of mathematical models.

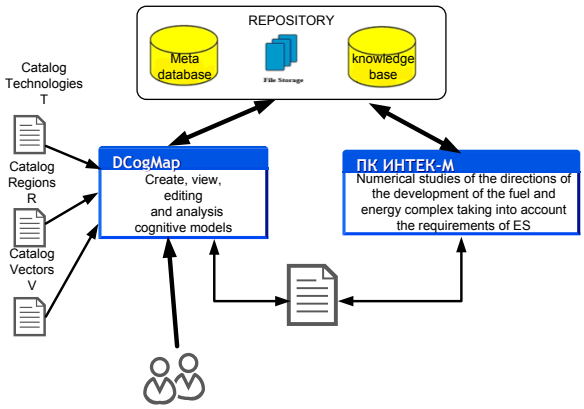


Figure 4 - General scheme of tools interaction within the intelligent IT environment

In addition to the use of directories, it is necessary to specify as DCM characteristics, and such information as the time step, various coefficients and limitations of mathematical models.

An inverse problem can also be solved: automating the construction of cognitive models using the results of a computational experiment.

Conclusion.

The article deals with the use of cognitive maps in energy research, suggests a transition to the use of FDCM, and also the possibility of integrating semantic and mathematical models is shown.

The proposed approach will provide a new quality of research by reducing errors (reducing the influence of the human factor) in the construction of cognitive models. Integration of mathematical and semantic models will help solve problems with incompleteness and possible unreliability of the initial data for a computational experiment, allowing in some cases to replace exact data in mathematical models with expert estimates.

The use of visual analytics and the ability to visualize calculations using cognitive maps will help improve the quality of the scientific justification for the recommended strategic solutions.

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. Pospelov D.A. Principles of situational management. – News of USSR AS, Technical Cybernetics. - 1971.- №2 .- Pp. 10-17 (in Russian)
2. Klykov Yu. I. Situational management of large systems. - M.: Energy.- 1974.- 134 p. (in Russian)
3. Pospelov D.A. Situational management. Theory and practice. - Moscow: Science, 1986. - 284 p. (in Russian)
4. Vasiliev VI, Ilyasov BG Intelligent control systems. Theory and practice - Moscow: Radio Engineering, 2009. - 392 p. (in Russian)
5. Trakhtengerts E.A. Computer support for decision-making. - Moscow: SINTEG, 1998. –376 p. (in Russian)
6. Mpelogianni, V., Marnetta, P., Groumpos, P.P.Fuzzy Cognitive Maps in the Service of Energy EfficiencyIFAC-PapersOnLine 48 (24), p. 1-6
7. Anninou, A.P. Groumpos, P.P. Polychronopoulos, P.Modeling health diseases using Competitive Fuzzy Cognitive Maps. IFIP Advances in Information and Communication Technology. Volume 412, 2013, Pages 88-95
8. Groumpos, V.P. Biniari, K. Groumpos, P.P.A new mathematical modelling approach for viticulture and winemaking using fuzzy cognitive maps. ELEKTRO2016 - 11th International Conference, Proceedings. Pages 57-61
9. Mica R.Endsley, Daniel J. Garland, Situation awareness: analysis and measurement, Lawrence Erlbaum Associates, 2000, ISBN 0805821341, 9780805821345
10. The Boyd Cycle (OODA cycle) and second-order cybernetics. - Open systems .- №7. - 2013. - Pp. 23-26.
11. Masel L.V. Application of ontological, cognitive and event modeling for the analysis of the development and consequences of emergencies in the energy sector / Problems of safety and emergency situations. - № 2.- 2010. - Pp. 34-43.
12. Massel L.V., Arshinsky V.L., Massel A.G. Intelligent information technologies for decision support in research and energy security providing / Proceedings of the International Conference "Intelligent decision-making systems and problems of computing experiment", Evpatoria, 2010. - Pp. 192-196.
13. Massel A.G. Intelligent IT- environment for research of the energy security problem // Information technologies in science, education, telecommunications and business: Proceedings of the International Conference - Ukraine, Gurzuf, 2010. - Pp. 306-309.
14. Massel A.G. Methodological approach to the organization of intelligent support for research on energy security problem/ Information technology. - №9. - 2010. - Pp. 32-36.
15. Massel A.G. Cognitive modeling of threats to energy security / Mining Information and Analytical Bulletin (scientific and technical journal), separate issue №17. - M.: Publishing house "Mining Book", 2010. - Pp. 194 - 199.
16. Massel A.G. Cognitive modeling of energy security treats / Proceedings of the 11th International Conference “Computer Science and Information Technologies”, vol. 1, 2009.- Greece, Crete. - P. 78-80.
17. Arshinsky V.L., Massel A.G., Fartyshev D.A. Multiagent software complex for research of the energy security problem // Proceedings of the XIV Baikal All-Russian Conference "Information and Mathematical Technologies in Science and Management". Vol. 3. - Irkutsk: MESI SB RAS, 2009. - Pp. 283-289.
18. Massel A.G., Kopaygorodsky A.N., Pascal K.K. Application of the cognitive modeling system in the research of energy security problems // Vestnik IrSTU. - 2012. - № 10 (69). - Pp. 29-33

SOLVING THE POINT-PLANE PROBLEM FOR THE CLASS OF AFFINE TRANSFORMATIONS AND DEVELOPMENT OF A FAST ITERATIVE ALGORITHM FOR REGISTERING OF 3D POINT CLOUDS

A. Vokhmintsev[1], A. Melnikov[2], T. Botova[3]
«Information Technology and system», Research Laboratory, Chelyabinsk State University, Chelyabinsk, Russian Federation[1]
Ugra State University, Khanty-Mansiysk, Russian Federation[2]
«Intelligent information technology and systems», Research Laboratory, Chelyabinsk State University, Chelyabinsk, Russian Federation[3]
vav@csu.u, melnikovav@uriit.ru, botova@csu.ru

Abstract: *Solution to the point-plane problem for the class of affine transformations will be found, and a fast accuracy iterative algorithm for registering 3D point clouds will be designed.*

1. Introduction

The scientific problem at solving which the project is aimed consists in the development of methods of generating a three-dimensional combined dense map are of the accessible environment around mobile a robot. To construct the qualitative three-dimensional models it is necessary to share data received from different sensors (images, clouds of three-dimensional points, depth maps) at each moment of time. Depth sensors and laser scanners are using to obtain information about the shape of a three-dimensional object [1-3]. Kinect camera is widely used to obtain the depth map of a scene. Existing algorithms for solving this problem can be divided into two groups. The algorithms of the first group are based on the classical algorithm ICP [4] and obtain a sparse cloud of three-dimensional points by calculating the structure of the motion sensor through a set of images. The disadvantage of this approach is the need for user input during the initial mapping phase. Algorithms of the second group are based on comparison of features in two-dimensional space of descriptors (SIFT, SURF, ORB) [5, 6]. The disadvantage of the second group of algorithms is the high computational complexity and the assumption of a sufficient number of common features on images of neighboring frames.

A common registration algorithm solves the variational problem of finding the optimum geometric (orthogonal or affine) transformation, which best matches two clouds of points with a given correspondence between points [4]. Different functionals lead to various methods for registration of point clouds. The most commonly used methods for searching correspondence between a pair of clouds are point-to-point and point-to-plane. For the class of orthogonal transformations, the solution of the point-to-point problem was explicitly given by of Horn [7,8]. An exact solution of the point-to-point problem for the case of an arbitrary affine transformation is given in following paper6. The variational problem of the point-to-point in the class of orthogonal transformations is usually solved using either the Levenberg-Marquardt iteration algorithm or the linearization method for small angles [9]. For the variational problem of the point-plane for the class of affine transformations, we proposed the exact solution [10]. An approximate solution of the point-plane problem for the class of orthogonal transformations was obtained using Khoshelham algorithm [11]. Note that the point-plane method for the class of orthogonal transformations is more robust to noise of sensors, but for this method the explicit solution of the problem has not been found yet. This complicates the use of the method in real-time registration applications.

One of the main problems that has not been acceptably solved, in the field of SLAM is that of dynamic investigated environment. The preliminary results of the project have shown that the use of an approximate 3D map of the surrounding environment [12] significantly improves the quality of recognition and localization of objects in dynamic, bulky scenes, especially with partial or complete occlusion of a object by other scene objects [13, 14]. Basically, the quality of the depth map obtained from the Kinect camera is good. However, there are some problems with the output information. So in the output data, invalid areas are formed due to the following reasons: the structured light of the radiation does not come to the camera after reflection, the resolution of the measurement of the depth decreases by a quadratic law with increasing depth, and a fast movement of the camera also leads to loss of data. Within the framework of this project, we are going to analyze existing calibration and correction algorithms and to develop a new algorithm for constructing an accurate 3D map of the surrounding environment for dynamic, contextually complicated and large-scale scenes [15-17]. In the context of this project new accurate methods for reconstruction of three-dimensional maps of the environment will be developed in terms of complex navigation tasks of mobile robot in unknown dynamic environment with obstacles [18, 19]. We will present the foundation for required estimates of accuracy of the reconstruction of the environment and computational complexity based on adaptive approach. The proposed method will show the accurate reconstruction of scenes are not less than a specified threshold.

Within the framework of the proposed project, new accurate method for reconstruction of three-dimensional map of the unknown dynamic environment will be provided. The paper is organized as follows. In Section 2, the 3d scene morel is presented. In Section 2 the proposed registration algorithm of 3D point clouds is presented and discussed. Computer simulation results are provided in Section 3. Section 4 summarizes our conclusions.

2. Fast iterative algorithm for registering of 3D point clouds

Let $P = \{p_1, \dots, p_n\}$ be a source point cloud, and $Q = \{q_1, \dots, q_n\}$ be a destination point cloud in $\mathbb{R}3$. Suppose that the relationship between points in P and Q is given in such a manner that for each point p_i exists a corresponding point qi . The ICP algorithm is commonly considered as a geometrical transformation for rigid objects mapping P to Q:

$$Rp_i + t, \qquad (1)$$

where R is a rotation matrix, t is a translation vector, $i = 1, \dots, n$. The S-ICP algorithm uses a slightly different geometrical transformation given as

$$SRp_i + t, \qquad (2)$$

where S is a scaling matrix.

The group of affine transformations in the dimension of three has 12 generators. It means that the affine transformation in the dimension of three is a function of 12 variables. Let us consider the ICP variational problem for an arbitrary affine transformation in the point-to-plane case. Denote by $S(Q)$ a surface constructed from the cloud Q, by $T(q^i)$ denote a tangent plane of $S(Q)$ at point q^i . Let $J(A, T)$ be the following function:

$$J(A) = \sum_{i=1}^{n} \left(< Ap^i - q^i n^i > \right)^2, \qquad (3)$$

where $<\cdot,\cdot>$ denotes the inner product, A is a matrix of an affine transformation in the homogenous coordinates:

$$A=\begin{pmatrix} a_{11} & a_{12} & a_{13} & t_1 \\ a_{21} & a_{22} & a_{23} & t_2 \\ a_{31} & a_{32} & a_{33} & t_3 \\ 0 & 0 & 0 & 1 \end{pmatrix}. \quad (4)$$

p^i is a point from the cloud P, n^i is the unitary normal for $T(q^i)$:

$$p^i=\begin{pmatrix} p_1^i \\ p_2^i \\ p_3^i \\ 1 \end{pmatrix}, \quad n^i=p^i=\begin{pmatrix} n_1^i \\ n_2^i \\ n_3^i \\ 0 \end{pmatrix}. \quad (5)$$

Then new project solutions to 3D total variation regularization will be obtained. The function $J(A,\mathrm{T})$ can be rewritten:

$$J(A)=\sum_{i=1}^{n}\left(<Ap_i-q_i,n_i>\right)^2= \\ \sum_{i=1}^{n}\left(<Ap_i,n_i>-<q_i,n_i>\right)^2= \quad , (6) \\ =\sum_{i=1}^{n}<Ap_i,n_i>^2-2<Ap_i,n_i><q_i,n_i>+<q_i,n_i>^2$$

where p_i is a point from the point cloud P, n_i is the unitary normal for $T(q_i)$.

The term $< q_i, n_i >$ is a constant with respect to A, thus the variational problem (6) takes the form:

$$\sum_{i=1}^{n}<Ap_i,n_i>^2-2<Ap_i,n_i><q_i,n_i>+ \\ +<q_i,n_i>^2\to\min_{A} \quad .(7)$$

Let us rewrite the inner product $< A\,p_i, n_i >$ as follows:

$$<Ap_i,n_i>=\left(a_{11}p_1^in_1^i+a_{12}p_2^in_1^i+a_{13}p_3^in_1^i+t_1n_1^i\right)+ \\ +\left(a_{21}p_1^in_2^i+a_{22}p_2^in_2^i+a_{23}p_3^in_2^i+t_2n_2^i\right)+ \\ +\left(a_{31}p_1^in_3^i+a_{32}p_2^in_3^i+a_{33}p_3^in_3^i+t_3n_3^i\right)+(0) \quad .(8)$$

$J(A)$ is a function of 12 variables. Let us consider components of the gradient $\nabla J(A)$.

Partial derivatives with respect to t

Let us consider the partial derivatives with respect to a_{ij}, i,j = 1,…,3:

$$\frac{\partial J}{\partial t_j}=\sum_{i=1}^{n}2<Ap^i,n^i>n_j^i- \\ -2n_j^i<q^i,n^i>=0,\ j=1,\cdots,3. \quad (9)$$

$$\sum_{i=1}^{n}<Ap^i,n^i>n_j^i-n_j^i<q^i,n^i>= \\ =\sum_{i=1}^{n}<Ap^i,n^i>n_j^i-\sum_{i=1}^{n}n_j^i<q^i,n^i>=0 \quad .(10)$$

The sum $\sum_{i=1}^{n}n_j^i<q^i,n^i>$ is constant with respect to A,

denote it by cj: $\sum_{i=1}^{n}n_j^i<q^i,n^i>=c_j$.

The expression (10) takes the form:

$$\sum_{i=1}^{n}<Ap^i,n^i>n_j^i=c_j,\ \mathrm{j}=1,\ldots,3. \quad (11)$$

Denote by the $(PN)^i$ the following matrix:

$$\left(PN\right)^i=p^in^{it}=\begin{pmatrix} p_1^in_1^i & p_1^in_2^i & p_1^in_3^i & 0 \\ p_2^in_1^i & p_2^in_2^i & p_2^in_3^i & 0 \\ p_3^in_1^i & p_3^in_2^i & p_3^in_3^i & 0 \\ n_1^i & n_2^i & n_3^i & 0 \end{pmatrix}. \quad (12)$$

Let us consider the following expression:

$$<Ap^i,n^i>=tr\left(A\square\left(PN\right)^i\right). \quad (13)$$

It follows from

$$<Ap^i,n^i> \\ =<\begin{pmatrix} a_{11}p_1^i & a_{12}p_2^i & a_{13}p_3^i & t_1 \\ a_{21}p_1^i & a_{22}p_2^i & a_{23}p_3^i & t_2 \\ a_{31}p_1^i & a_{32}p_2^i & a_{33}p_3^i & t_3 \\ 1 & 1 & 1 & 1 \end{pmatrix},\begin{pmatrix} n_1^i \\ n_2^i \\ n_3^i \\ 0 \end{pmatrix}>= \\ =\left(a_{11}p_1^in_1^i+a_{12}p_2^in_1^i+a_{13}p_3^in_1^i+t_1n_1^i\right)+ \quad (14) \\ +\left(a_{21}p_1^in_2^i+a_{22}p_2^in_2^i+a_{23}p_3^in_2^i+t_2n_2^i\right)+ \\ +\left(a_{31}p_1^in_3^i+a_{32}p_2^in_3^i+a_{33}p_3^in_3^i++t_3n_3^i\right)+(0)$$

and

$$<\begin{pmatrix} a_{11} & a_{12} & a_{13} & t_1 \\ a_{21} & a_{22} & a_{23} & t_2 \\ a_{31} & a_{32} & a_{33} & t_3 \\ 0 & 0 & 0 & 1 \end{pmatrix}\begin{pmatrix} p_1^in_1^i & p_1^in_2^i & p_1^in_3^i & 0 \\ p_2^in_1^i & p_2^in_2^i & p_2^in_3^i & 0 \\ p_3^in_1^i & p_2^in_2^i & p_3^in_3^i & 0 \\ n_1^i & n_2^i & n_3^i & 0 \end{pmatrix}>=$$

$$=\begin{pmatrix} a_{11}p_1^in_1^i+a_{12}p_2^in_1^i+a_{13}p_3^in_1^i+t_1n_1^i & & & \\ & & & \\ a_{21}p_1^in_1^i+a_{22}p_2^in_1^i+a_{23}p_3^in_1^i+t_1n_1^i & & & \\ & & & \\ a_{31}p_1^in_3^i+a_{32}p_2^in_3^i+a_{33}p_3^in_3^i+t_3n_3^i & & & 0 \end{pmatrix} \quad .(15)$$

Let us rewrite Eq. (13) taking into account Eq. (15) as

$$\sum_{i=1}^{n} n_j^i tr\left(A\square(PN)^i\right)c_j, \mathrm{j}=1,\ldots,3. \quad (16)$$

Denote by $(n_jPN)^i$ the following matrices:

$$\left(n_jPN\right)^i=\begin{pmatrix} p_1^i n_1^i n_j^i & p_1^i n_2^i n_j^i & p_1^i n_3^i n_j^i & 0\\ p_2^i n_1^i n_j^i & p_2^i n_2^i n_j^i & p_2^i n_3^i n_j^i & 0\\ p_3^i n_1^i n_j^i & p_3^i n_2^i n_j^i & p_3^i n_3^i n_j^i & 0\\ n_1^i n_j^i & n_2^i n_j^i & n_3^i n_j^i & 0 \end{pmatrix}. \quad (17)$$

The following condition holds:

$$n_j^i tr\left(A\square(PN)^i\right)=tr(A\square(n_j\left(PN\right)^i). \quad (18)$$

Taking into account Eq. (18), we rewrite Eq. (16) as

$$\sum_{i=1}^{n} tr(A\square(n_j\left(PN\right)^i)=c_j. \quad (19)$$

Using the property of the matrix trace, we arrive to

$$\begin{aligned}&\sum_{i=1}^{n} tr(A\square(n_j\left(PN\right)^i)\\ &=tr\left(\sum_{i=1}^{n} A\square(n_jPN)^i\right)\\ &tr\left(A\left(\sum_{i=1}^{n}(n_jPN)^i\right)\right)=c_j\end{aligned}. \quad (20)$$

Denote by Mj, $j = 1, \ldots, 3$, the following matrices:

$$\sum_{i=1}^{n}(n_jPN)^i=M^j, \quad (21)$$

$$M^j=\begin{pmatrix} m_{11}^j & m_{12}^j & m_{13}^j & 0\\ m_{21}^j & m_{22}^j & m_{23}^j & 0\\ m_{31}^j & m_{32}^j & m_{33}^j & 0\\ m_{41}^j & m_{42}^j & m_{43}^j & 0 \end{pmatrix}. \quad (22)$$

The equation (22) can be rewritten as

$$tr\left(A\square M^j\right)=c_j,\ j=1,\cdots,3, \quad (23)$$

$$\begin{aligned}&A\square M^j=\\ &=\begin{pmatrix} a_{11} & a_{12} & a_{13} & t_1\\ a_{21} & a_{22} & a_{23} & t_2\\ a_{31} & a_{32} & a_{33} & t_3\\ 0 & 0 & 0 & 1 \end{pmatrix}\begin{pmatrix} m_{11}^j & m_{12}^j & m_{13}^j & 0\\ m_{21}^j & m_{22}^j & m_{23}^j & 0\\ m_{31}^j & m_{32}^j & m_{33}^j & 0\\ m_{41}^j & m_{42}^j & m_{43}^j & 0 \end{pmatrix}=\\ &=\begin{pmatrix} a_{11}m_{11}^j+a_{12}m_{21}^j+a_{13}m_{31}^j+t_1m_{41}^j\\ a_{21}m_{12}^j+a_{22}m_{22}^j+a_{23}m_{32}^j+t_2m_{42}^j\\ a_{31}m_{13}^j+a_{32}m_{23}^j+a_{33}m_{33}^j+t_3m_{43}^j\\ 0 \end{pmatrix}\end{aligned}. \quad (24)$$

We get the following equations for $j = 1, \ldots, 3$:

$$\begin{aligned}&\left(a_{11}m_{11}^j+a_{12}m_{21}^j+a_{13}m_{31}^j+t_1m_{41}^j\right)+\\ &+\left(a_{21}m_{12}^j+a_{22}m_{22}^j+a_{23}m_{32}^j+t_2m_{42}^j\right)+\\ &+\left(a_{31}m_{13}^j+a_{32}m_{23}^j+a_{33}m_{33}^j+t_3m_{43}^j\right)=c_j,\\ &j=1,\cdots,3, i=1,\cdots,n\end{aligned}. \quad (25)$$

Partial derivatives with respect to a

Let us consider the partial derivatives with respect to a_{ij}, $i, j = 1, \ldots, 3$.

$$\begin{aligned}&\frac{\partial J}{\partial a_{ij}}=\sum_{k=1}^{n} 2<Ap^k,n^k>p_j^k,n_i^k-\\ &-2p_j^k,n_i^k<q^k,n^k>=0,\ j=1,\cdots,3,\\ &k=1,\cdots,\mathrm{n}\end{aligned}, \quad (26)$$

$$\begin{aligned}&\sum_{i=1}^{n}<Ap^k,n^k>p_j^k,n_i^k-p_j^k,n_i^k<q^k,n^k>=\\ &=\sum_{i=1}^{n}<Ap^k,n^k>p_j^k,n_i^k-\\ &-\sum_{i=1}^{n}p_j^k,n_i^k<q^k,n^k>=0\end{aligned}. \quad (27)$$

The partial sum $\sum_{k=1}^{n} p_j^k n_i^k<q^k,n^k>$ is a constant with respect to A:

$$\sum_{k=1}^{n} p_j^k n_i^k<q^k,n^k>=c_{ij}, i, j=1,\cdots,3. \quad (28)$$

The expression (27) takes the form:

$$\sum_{k=1}^{n}<Ap^k,n^k>=p_j^k n_i^k=c_{ij}. \quad (29)$$

Taking into account Eqs. (3) and (4) can be rewritten as

$$\sum_{k=1}^{n} p_j^k n_i^k tr\left(A\square(PN)^k\right)=c_{ij}, i, j=1,\cdots,3. \quad (30)$$

Denote by $(p_j^k n_i^k PN^k)$ the following matrices:

$$(p_j^k n_i^k PN^k) =$$

$$\begin{pmatrix} p_1^k n_1^k\, p_j^k n_i^k & p_1^k n_2^k\, p_j^k n_i^k & p_1^k n_3^k\, p_j^k n_i^k & 0 \\ p_2^k n_1^k\, p_j^k n_i^k & p_2^k n_2^k\, p_j^k n_i^k & p_2^k n_3^k\, p_j^k n_i^k & 0 \\ p_3^k n_1^k\, p_j^k n_i^k & p_3^k n_2^k\, p_j^k n_i^k & p_3^k n_3^k\, p_j^k n_i^k & 0 \\ n_1^k\, p_j^k n_i^k & n_2^k\, p_j^k n_i^k & n_3^k\, p_j^k n_i^k & 0 \end{pmatrix}. \quad (31)$$

The expression (1) can be rewritten as

$$\sum_{k=1}^{n} tr\left(A \square p_j^k n_i^k PN^k\right) = c_{ij} \ . \quad (32)$$

Denote by Mij, $i, j = 1, \ldots, 3$, the following matrices:

$$\sum_{k=1}^{n}\left(p_j^k n_i^k PN^k\right) = M_{ij} \ , \quad (33)$$

$$AM^{ij} = \begin{pmatrix} a_{11} & a_{12} & a_{13} & t_1 \\ a_{21} & a_{22} & a_{23} & t_2 \\ a_{31} & a_{32} & a_{33} & t_3 \\ 0 & 0 & 0 & 1 \end{pmatrix} \times$$

$$\times \begin{pmatrix} m_{11}^{ij} & m_{12}^{ij} & m_{13}^{ij} & 0 \\ m_{21}^{ij} & m_{22}^{ij} & m_{23}^{ij} & 0 \\ m_{31}^{ij} & m_{32}^{ij} & m_{33}^{ij} & 0 \\ m_{41}^{ij} & m_{42}^{ij} & m_{43}^{ii} & 0 \end{pmatrix} =$$

$$= \begin{pmatrix} a_{11}m_{11}^{ij} + a_{12}m_{21}^{ij} + a_{13}m_{31}^{ij} + t_1 m_{41}^{ij} \\ \\ a_{21}m_{12}^{ij} + a_{22}m_{22}^{ij} + a_{23}m_{32}^{ij} + t_2 m_{42}^{ij} \\ \\ a_{31}m_{13}^{ij} + a_{32}m_{23}^{ij} + a_{33}m_{33}^{ij} + t_3 m_{43}^{ij} \\ \\ 0 \end{pmatrix}. \quad (35)$$

Finally, we get 9 equations for $i, j = 1, \ldots, 3$:

$$\begin{aligned} &\left(a_{11}m_{11}^{ij} + a_{12}m_{21}^{ij} + a_{13}m_{31}^{ij} + t_1 m_{41}^{ij}\right) + \\ &+\left(a_{21}m_{12}^{ij} + a_{22}m_{22}^{ij} + a_{23}m_{32}^{ij} + t_2 m_{42}^{ij}\right) + \\ &+\left(a_{31}m_{13}^{ij} + a_{32}m_{23}^{ij} + a_{33}m_{33}^{ij} + t_3 m_{43}^{ij}\right) = c_{ij}, \\ &i, j = 1, \cdots, 3, \end{aligned} \quad . \quad (36)$$

We get the set of 12 linear equations (25), (36) with 12 variables a_{ij}, t_k, $i, j, k = 1, \ldots, 3$. We can write this system as the matrix equality:

$$Z \square \Omega = C \ , \quad (37)$$

where matrix Z has size 9×12 and consist of the elements m_{ij}, C is the column consist of 12 numbers c_i and c_{ij} , and the column Ω :

$$\Omega = \left(a_{11}, a_{12}, a_{13}, \mathrm{t}_1, a_{21}, a_{22}, a_{23}, \mathrm{t}_2, a_{31}, a_{32}, a_{33}, \mathrm{t}_3\right)^t . \quad (38)$$

3. Computer simulation

In this section, computer simulation results are presented and discussed. We used Kinect as RGB-D camera. The following Indoor Segmentation and Support Inference from RGB Images ECCV 2012 (Indoor Segmentation and Support database) was used for the tests (http://cs.nyu.edu / ~silberman/datasets / nyu_depth_v2.html#raw_parts), which contains semantic space labeling. We have selected four datasets of large areas of the reference database with insufficient RANSAC inliers. To obtain this result we disovered denendncy the accuracy of the three-dimensional reconstruction with respect to the number and spatial distribution of singular points in the frame, in a sequences of experiments we found out a minimal number of features providing a given accuracy of the three-dimensional reconstruction for the type of each method (descriptor) in Table 1. We have made a argumentation for estimates of the accuracy and computational complexity of the proposed method. It was found that the accuracy of 3D reconstruction of scene depends with rescpect to the number of feaures in the frame in a nonlinear way. This function has one distinct extreme for all types of descriptors. This conclusion is very important to assess the convergence of the visual SLAM algorithms in real-time.

Table 1. *Quality of reconstructing a three-dimensional combined dense map with respect to descriptor*

Method	Dataset-1	Dataset-2	Dataset-3	Dataset-4
Precision, %				
SIFT	93	93	94	95
SURF	57	61	63	59
ORB	78	82	75	75
HOGs	93	96	97	96
The processing time, s				
SIFT	17.2	26.21	12.82	18.9
SURF	0.23	0.7	0.4	0.44
ORB	3.34	3.42	5.15	3.33
HOGs	0.79	1.11	0.68	0.81

4. Conclusion

In this project the following main results were obtained:

- Solving the point-plane problem for the class of affine transformations;
- Development of a new fast iterative method for constructing a 3D accurate map of the surrounding environment using sequences of sparse point clouds obtained from depth sensors for dynamic, contextually complicated and large-scale scenes. It is necessary to design an algorithm for registration cloud clouds with arbitrary spatial resolution and scale relative to each other.
- Development of algorithms that take into account a 3D map of the surrounding environment, for reliable recognition and localization of objects in dynamic, contextually complicated scenes, especially with partial or complete occlusion of persons by other scene objects.

Acknowledgments

The work was supported by the RFBR, project no 16-08-00342.

References

1. Hertzberg C., Wagner R. and Birbach O. “Experiences in building a visual slam system from open source components”. In: Proc. IEEE International Conference on Robotics and Automation, 2011, pp. 2644-2651.
2. Henry P., Krainin M. and Herbst E. “RGB-D mapping: Using depth cameras for dense 3D modeling of indoor environments”. In. Proc. 12th International Symposium on Experimental Robotics, 2014, pp. 477-491.
3. Vokhmintsev A. and Yakovlev K. “A Real-time Algorithm for Mobile Robot Mapping Based on Rotation invariant Descriptors and ICP”. In: Proc. of 5th Analysis of Images, Social Networks and Texts. Springer. Communications in Computer and Information Science, Vol. 661, 2017, pp. 357-369.
4. Besl P. and McKay N. “A method for registration of 3-D shapes”. IEEE Transactions of Pattern Analysis and Machine Intelligence, 1992, 14 (2): 239–256.
5. Vokhmintsev A., Timchenko M. and Yakovlev K. “Simultaneous localization and mapping in unknown environment using dynamic matching of images and registration of point clouds”. In: Proc. of IEEE Proc. 2nd International Conference on Industrial Engineering, Applications and Manufacturing (ICIEAM), 2017.
6. Chen Y. “Kalman filter for robot vision: a survey”. Journal IEEE Transactions on Industrial Electronics, 2012, 59: 4409–4420.
7. Horn B. “Closed-Form Solution of Absolute Orientation Using Unit Quaternions”. Journal of the Optical Society of America A, 1987, 4(4): 629–642.
8. Horn B., Hilden H. and Negahdaripour S. “Closed-form Solution of Absolute Orientation Using Orthonormal Matrices”. Journal of the Optical Society of America A, 1988, 5(7): 1127-1135.
9. Du S., Zheng N., Ying S. and Liu J. “Affine iterative closest point algorithm for point set registration”. Pattern Recognition Letters, 2010, 31: 791–799.
10. Low K.L. “Linear least-squares optimization for point-to-plane ICP surface registration”. Technical Report TR04-004, Department of Computer Science, University of North Carolina at Chapel Hill, 2004.
11. Khoshelham K. “Closed-form solutions for estimating a rigid motion from plane correspondences extracted from point clouds”. ISPRS Journal of Photogrammetry and Remote Sensing, 2016, 114: 78–91.
12. Vokhmintsev A. and Yakovlev K. “A Real-time Algorithm for Mobile Robot Mapping Based on Rotation invariant Descriptors and ICP”. In: Proc. of 5th Analysis of Images, Social Networks and Texts. Springer. Communications in Computer and Information Science, Vol. 661, 2017, pp. 357-369.
13. Tam G., Cheng Z.-Q., Lai Y.-K., Langbein F., Liu Y., Marshall D., Martin R., Sun X.-F. and Rosin P. “Registration of 3D point clouds and meshes: A survey from rigid to nonrigid”. IEEE Trans. Vis. Comput. Graph., 2013, 19(7): 1199–1217.
14. Picos K., Diaz-Ramirez V.-H., Kober V., Montemayor A.-S. and Pantrigo J.-J. “Accurate three-dimensional pose recognition from monocular images using template matched filtering”. Optical Engeenering, 2016, 55(6): 0631-02.
15. Vidal-Calleja T. A., Berger C., Solà, J. and Lacroix S. “Large scale multiple robot visual mapping with heterogeneous landmarks in semi-structured terrain”. Robotics and Autonomous Systems, 2011, 59(9): 654-674.
16. Se S., Lowe D. and Little J. “Mobile robot localization and mapping with uncertainty using scaleinvariant visual landmarks”. The international Journal of robotics Research, 2002, 21(8): 735–758.
17. Cheng S., Marras I. and Zafeiriou S. “Active nonrigid ICP algorithm”. In: Proc. of 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition, 2015, pp. 1- 8.
18. .Makovetskii A., Voronin S., Kober V. and Tihonkih D. “An efficient point-to-plane registration algorithm for affine transformations”. In: Proc. of SPIE 10396, Applications of Digital Image Processing XL, 2017, pp. 103962J 7.
19. Vokhmintsev A., Timchenko M., Melnikov A., Kozko A., Makovetskii A. “Robot path planning algorithm based on symbolic tags in dynamic environment” In: Proc. of SPIE, Applications of Digital Image Processing XL, 2017, pp. 103962E-8.

USING DATA MINING TECHNIQUES TO CREATE AN AUTOMATED MODEL THAT MAKES COMPARISONS BETWEEN MARKET DEMANDS AND UNIVERSITY CURRICULA

M.Sc. Ylber Januzaj PhD (C).[1], Prof. Assoc. Artan Luma PhD.[1], Prof. Assoc. Azir Aliu PhD.[1], Prof. Assoc. Besnik Selimi PhD.[1], Prof. Assoc. Bujar Raufi PhD.[1], Prof. Assoc. Halil Snopce PhD.[1], Prof. Ass. Vehbi Ramaj PhD.[2]
Faculty of Contemporary Sciences and Technologies – South East European University, Macedonia [1]
Faculty of Economics – University of Peja "Haxhi Zeka", Kosovo [2]

{yj16535, a.luma, azir.aliu, b.selimi, b.raufi, h.snopce}@seeu.edu.mk [1]

vehbi.ramaj@unhz.eu [2]

Abstract: *Studying the right field has great importance in human life and perspective. Rather than affecting the greatest employer's ability, some studies see higher education as one of the leading factors that directly affects the style of life that we do. Therefore, today's demands have increased significantly for skilled people, and prepared in complex areas, and a consolidation between market demands and university curricula is needed. This paper examines Data mining techniques which are used in order to create an automated model which makes comparisons between market demands and university curricula. We also present how proposed model is able to give recommendations, based on the comparison between market demands and university curricula.*

Keywords: TECHNOLOGY, DATA MINING, JOB MARKET, UNIVERSITY CURRICULA, WEB MINING

1. Introduction

It is precisely technology that has affected the lives of people today to be more problematic and more complex, making the vast majority of information available today in the electronic format. And precisely this large volume of information, and the exceeded number of technological devices, has made demand for the labor market in the field of technology to be increased every day and more.

Latest achievements in Data Mining are huge, contributing directly to the development of different fields. Increment of data from time to time has made it necessary to apply Data Mining as a field in order to produce results as accurate and timely as possible. In [1,3] it is mentioned that we are now living in information time.

Lastly, we can easily see that we are dealing with digitalization of almost all areas, ranging from social to scientific ones, and that is precisely why we are dealing with such a large volume of data.

If we stop at social networks, nowadays we can see how from seconds to seconds we are dealing with increasing data in different formats: textual, photo, audio, video, etc.

On the other hand, if we stop at the scientific area, we know that the various exact sciences have reached in a high point of research by applying the technology and digitization their data [6,7,9]. Therefore, this huge amount of data that is crossing the world day by day, it is necessary to apply professional techniques that enable us to process accurate results in a short time.

2. Related Work

The data we are talking about, which grow dynamically at any moment may not be understandable to us when they are in the original format, but applying Data Mining techniques they are processed and converted into a format that is understandable to us. Therefore, as we can see, the main purpose of Data Mining is to adapt to the human language by becoming more and more inseparable part of us.

When we say inseparable part of ours, it should be noted that thousands of applications today function as an inseparable part of the human world by giving different people recommendations on different activities [2]. These recommendations given to different people through various applications today are made possible through Data Mining techniques. And it is precisely the collection and processing of various information that has influenced these applications to give us recommendations and conclusions through the Data Mining techniques.

A concrete example is the case of medical analyzes, where digitized devices are able to produce results on our health condition based on earlier samples of different persons that are stored in the system. The greater the capacity of these data stored in the system, the greater the precision of the result determined by the digital system. This is an exact example of how Data Mining converts data into a format that is understandable to the human world as well.

In [4,5] it is mentioned that evolution and technology development has occurred precisely because of the tremendous development that Data Mining has suffered. This dependence between technology development and Data Mining lies in the fact that the data available today on the internet is of a great variety.

The data that Data Mining can handle are: textual, numeric, structured data and unstructured data, graphical data, and data that are distributed through web systems.

In the following we explain how Data Mining has found the application on each of them.

2.1. Textual data

When we talk about textual data, the Data Mining application process in the text is known as Text mining or Text Data mining. Text mining includes the process of text processing to the presentation in a professional and high quality. The process of text processing through Data mining techniques is the process of removing a word in the text, adding a new word, or completely structuring a document based on a model that the system possesses as a training model [8,11,14]. And once the text has been processed, then it becomes ready for presentation in a format that is understandable to the human eye.

So the main purpose of the Data mining application in text is to convert textual data into data that are readily available for analysis. A simpler method for Text Data mining is converting documents from hard copy to electronic format in order to enable the application of Data mining techniques. Some of the areas that directly depend on Text mining are state intelligence, search engines, publishing houses, social networks, and so on.

2.2. Numeric Data

Creating different models that are able to provide predictive results based on preliminary results is precisely the application of Data mining to numerical data. Nowadays, many predictive analyzes are needed that help us to make decisions based on the result that the system exits. A method by which we can make predictions based on preliminary results is through Bayesian. Through this method we can make data sorting by dividing the data

into two or more groups that the system classifies. Even in this case, the more data that the system possesses the more accurate the classification that the system determines. Some of the domains where Numerical Data mining has found application are: math's, medicine, physics, chemistry, and so on.

2.3. Structured and unstructured Data

The data that can be found in a database can be structured and unstructured [1,16,17]. Structured data are the data that are organized in the best possible way. While unstructured data are data that do not have an organization and a structure. When we are in structured data, it's easier to apply Data mining techniques as we have tables that are related to one another, so when a given data is needed then we know exactly which table and we which line should we ask for it. While the main problem and concern lies in unstructured data as we have no information on where to find it. However, through Data mining techniques we can create intelligent systems using unstructured data and creating in this form of structured and understandable human data forms.

2.4. Graphical Data

A kind of data which is difficult to process are the graphical data due to the complexity they possess and because of the difficulty of conversion in data for analysis. Currently in almost all areas, the results are converted into graphs in order to be more understandable during different presentations. With Data mining techniques we can also process these data in order to obtain desired results. Data mining application in graphical data is known as Graph mining, and unlike other Data mining techniques this technique is less accurate [10,12].

2.5. Web Data

Nowadays companies of all areas promote themselves through the web platform [9]. Apart from the information that should contain a website, designers today are also focused on the dynamics that the website incorporates, including information as part of design and multimedia. It is precisely this fact that has troubled the work of extracting and processing information from web sites. Our automated model depends exactly on the websites, as the main information will be extracted from them. And the best solution is to apply Data mining techniques to the web. Two techniques that enable us to extract this content are Web Scraping and Web Crawling, which we will later discuss in more detail.

2.6. Data mining evolution

In the following we will present the evolution of Data mining and its history of how it has reached this peak point where it is today.

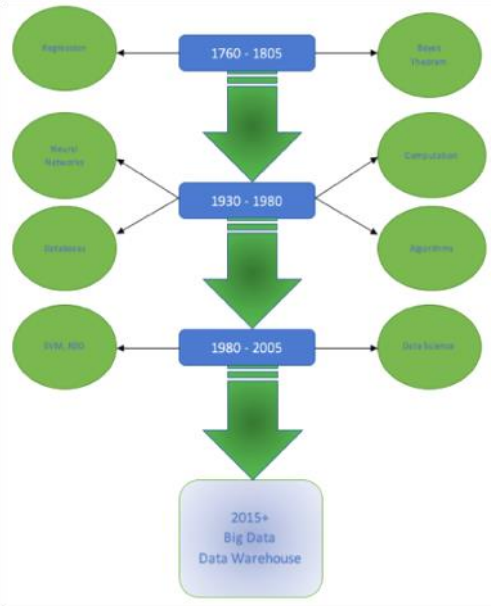


Fig. 1. Data mining evolution over centuries

In Fig. 1 we can see how Data mining has been developed since the 17th century, where from 1760 to 1805 we have the discovery of two theorems: Bayesian and Regression. Currently, through these two theorems, science has reached high discovery by yielding accurate results and predictions regarding certain cases [19,22]. We can also note that in the 19th century there have been great discoveries and great advancements in the Data mining field. Specifically, from 1930 to 1980 is the period when the first databases were created, also the Neural Networks was discovered. It is precisely this time when important algorithms that day-to-day use are being discovered, and since then, they have been sophisticated and advanced in different languages. Also this period is well known for the digitization of these discoveries since all the theorems that were discovered and all the algorithms that were discovered began to be applied through the computer, directly affecting them at the time of their implementation and in the accuracy of the processed results.

Not by chance that these algorithms were applied to the computer, since at this time the volume of data that was stored on the Internet began to increase as the Internet has begun to evolve. Also, in this period, from 1980 to 2005, other discoveries of various algorithms that are known today and are very applicable were made, such as SVM, Deep Learning, KDD, etc., Therefore, this period is also known as the period of the Data Science, at that time scientific methods began to be implemented to produce data and to present the data in a human-accessible format, whether structured or unstructured.

In Fig. 1 we can also note that the last period since 2015 is known as the Big Data period, where the data format that is stored is petabyte above. So this is exactly the time when different businesses and corporations began to use technology in order to create intelligent systems that represent their business in the best possible way [21]. Also in this period, the Data Mining application is started on giant and data sets known as Big Data. Also, the Data Mining application at Big Data has made it possible to make data processing in a precise and fast way. Finally the areas where Data mining has found application on Big Data are different, such as: Medicine, Education, Social Science, Marketing, Finance, etc.

3. Techniques used to build automated model that makes comparisons between market demands and university curricula

Technology development has directly influenced other areas as well. It has greatly influenced the application of different techniques in order to make consolidation between the labor market and the study programs offered by universities has greatly influenced. Next we present the techniques used to create an automated system that will make a comparison between the labor market and curriculum requirements offered by universities.

3.1. Clustering

The division of data into certain groups based on the similarity of objects is known as Clustering. Like other Data mining techniques, clustering is one of the techniques that has managed to develop alongside other techniques. Developing and advancing clustering techniques has also led to the development of many other areas such as medicine, education, finance, marketing, machine learning, etc.

The way clustering works is by creating as many groups of objects which are the same with each other. The greater the likeness of being within a group, and the bigger the difference between the groups, the greater the clustering accuracy.

In addition to clustering there are many other techniques that make data collection in different groups, but in some literature we can also find that clustering can also be known as a form of classification. We say that it is known as a form of classification, as it divides into data groups by classifying them into different groups. But distinction from the classification technique that classifies the

data into different groups, and when it comes to a data that is classified as a new group, clustering these new data tries to group them with the actual data being compared to the most similar figure. In the following we will show graphically how clustering works.

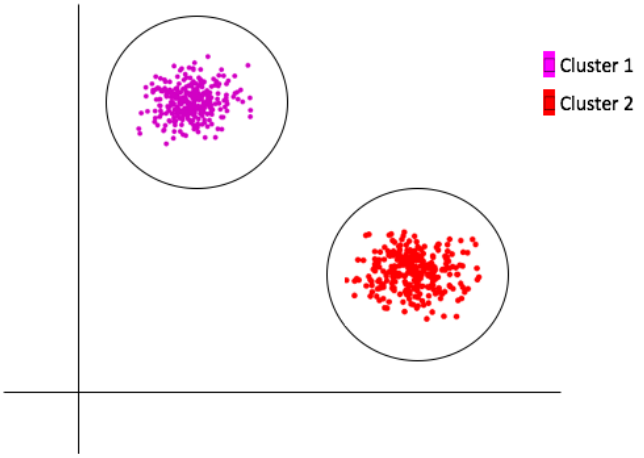


Fig. 2. Clustering different groups

In Fig. 2 we can see how the clustering technique is divided into two sets of data that are located in our dataset. These data are divided into two groups that differ with each other, but with data that are close to each other. In our case the data are divided into two groups only because of the difference between them, but in other cases we have more groups. Also in Fig. 2 we can see that we have a high level clustering because the data are completely separated, but there may also be times when the data cannot be completely separated but only join to the group which is more closest. When we are in clustering, we can point out that clustering types are: well separated, center - based, contiguity - based, density - based and conceptual clusters.

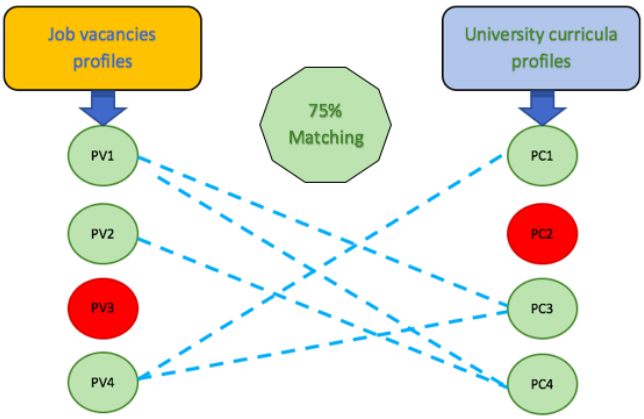


Fig. 3. Matching between market demand and university curricula

In Fig. 3 we can see how the matching algorithm works, which in this case is illustrated how it will be applied to our model. As we can see on the left hand we have job vacancies profiles that are named from PV1 to PV4, and under the right hand we have university curricula profiles that are named from PC1 to PC4. After applying the algorithm we can note that the PV1 has PC3 and PC4 adaptability, as well as PV2 has PC4 compatibility, and PV4 has PC1 and PC4 compatibility. As we can see profiles vacancies 3 and profile curricula 2 have no relation to any of the profiles. And what results from this result is that profile 3 of the job remains uncovered, and in that case our system will be able to give recommendations that this profile should be covered so that a new curriculum can be added.

Or in the other case we have profile curricula 2 which also does not have any links to any of the profiles of work, then we can conclude that this curriculum is not needed in the labor market, and our system will be able to make recommendations on the changes that need to be made in this curriculum so that it responds to the demands of the labor market.

Also based on the results that derived our algorithm after the application, it can be concluded that out of 4 profiles, 3 of them respond to labor market demands or 75% is the level of adjustment between the demands of the labor market and the curricula offered from universities in the field of technology.

3.3. Web crawling

The program that allows us to automate browsing through websites that are active is known as web crawling [14]. The way the web crawler works is by checking webpages that are active in certain phrases that we define by themselves. Some websites today use web crawling as a perfect way to keep their web pages updated.

The other way web crawlers work is by visiting the web pages automatically and downloading them to a local disk that we set as a destination to maintain the content. The content of the webpage we can download, starting from static to dynamic, but depending on the dynamics of the webpage depends also on the script we need to build in order to have as much information as possible. Below we will graphically show how the web crawler works.

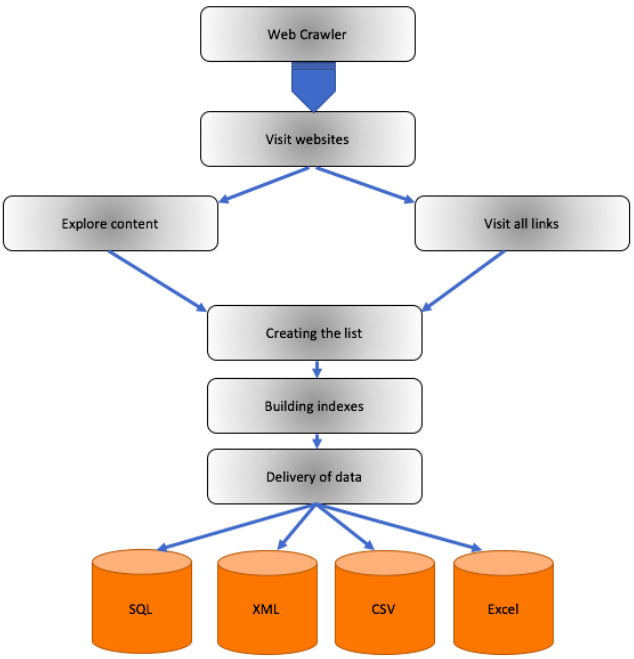


Fig. 4. Extraction of vacancies websites

In Fig. 4 we can see that the first step in which the web crawler application passes is the site visit that we define at the beginning. After a visit to the web, the crawler makes an exploration of the content based on the phrases that we have assigned at the beginning, and at the same time the crawler visits the other links that are defined within the webpage.

Once the textual content has been extracted, the crawler creates lists of those web pages, in order to create indexes that will be downloaded later.

Once the indexes are created then the last step is to save the data, or as we have presented it as delivery of the data, where the data we extracted from the content of the webpage can be stored in any format that we have need. The most used formats that can be converted to the data extracted from the web site are: SQL, XML, CSV, Excel, etc [20, 21].

Based on this, the main reason for web crawling will be applied in our research is to extract information that is published by some websites on vacancies. These data will be extracted in certain keywords, where their clustering will also be based on them.

Initially in the region there are web pages that contain data on university programs, these data also contain descriptions of the subject and description of the study program.

Based on this information, will be the clustering of study programs that are offered by universities. These later clustering will be used to match the market requirements.

On the other hand, Crawling will also be applied on web sites that provide information on competitions offered by different companies [13,17]. Also, the web crawling application will be made to specific keywords, knowing that each contestant has additional information on the specific position requests.

So, every position that is required to keep information in addition to the required position, the applicant must have knowledge in several different areas such as: programming, databases, networking, etc. All of these descriptions of later positions will be used to gather different positions.

4. Conclusion

The relevance of adapting the study programs to the labor market demands has a great importance. In our paper, we presented the tools that helps us to create a model that will compare the demands of the labor market and the curricula offered by universities.

First we presented the latest achievements in the Data mining field and the stages that Data Mining has passed to reach where it is today. Then we have seen what are the techniques we have used to create the model that will make comparisons between market demands and the curricula offered by universities.

Also in our paper we presented parts of the automated model such as: the application of the algorithm that will find the level of the actual adjustment between the labor market demands and the curricula offered by the universities. Then we presented the steps that are passed in order to extract the webpages that publish competitions in the field of technology.

In the end we conclude that our country and the region has a great need for a better adjustment between study programs and labor market demands. We also conclude that the creation of such an automated model will help meet the demands of the labor market, and our universities will be able to offer students also for the European market.

References

[1] M. Agaoglu, "Predicting Instructor Performance Using Data Mining Techniques in Higher Education". In IEEE Access, May 2016.

[2] T. Xie, Q. Zheng, W. Zhang, H, Qu, "Modeling and Predicting the Active Video – Viewing Time in a Large – Scale E – Learning System". In IEEE Access, June 2017.

[3] A. M. Njeru, M. S. Omar, S. Yi, "IoTs for Capturing and Mastering Massive Data Online Learning Courses". In IEEE Computer Society, ICIS, Wuhan, China, May 2017.

[4] R. Heartfield, G. Loukas, D. Gan, "You are probably not the weakest link: Towards Practical Prediction of Susceptibility to Semantic Social Engineering Attacks". In IEEE Access, October 2016.

[5] E. J. Fortuny, D. Martens, "Active Learning – Based Pedagogical Rule Extraction". In IEEE Transaction on Neural Network and Learning Systems, Vol. 26, No. 11, November 2015.

[6] A. Mukhopadhyay, S. Bandyopadhyay, "A Survey of Multiobjective Evolutionary Algorithms for Data Mining: Part I". In IEEE Transaction on Evolutionary Computation, Vol. 18, No. 1, February 2014.

[7] Zh. Song, A. Kusiak, "Optimization of Temporal Processess: A Model Predictive Control Approach". In IEEE Transaction on Evolutionary Computation, Vol. 13, No. 1, February 2009.

[8] S. Malgaonkar, S. Soral, Sh. Sumeet, T. Parekhji, "Study on Big Data Analytics Research Domain". In International Conference on Reliability, Infocom Technologies and Optimization ICRITO,Noida, India, September 2016.

[9]K. P. Anicic, B. Divjak, K. Arbanas, "Preparint ICT Graduates for Real – World Challenges: Results of a Meta – Analysis". In IEEE Transactions on Education, Vol 60, No. 3, August 2017.

[10] A. Haskova, D. V. Merode, "Professional Training in Embedded Systems and its Promotion". In IEEE Transacions on Education, 2016.

[11] S.C. Smith, W. K. Al-Assadi, J. Di, "Integrating Asynchronous Digital Design into the Computer Engineering Curriculum". In IEEE Transactions on Education, Vol. 53, No. 3, August 2010.

[12] M. D. Koretsky, D. Amatore, C. Barnes, Sh. Kimura, "Enhancement of Student Learning in Experimental Design Using a Virtual Laboratory". In IEEE Transactions on Education, Vol. 51, No. 1, February 2008.

[13] B. G. Member, V. S. Sheng, K. Y. Tay, W. Romano, Sh. Li, "Incremental Support Vector Learning for Ordinal Regression". In IEEE Transactions on Neural Networks and Learning Systems, Vol. 26, No. 7, July 2015.

[14] J. Li, T. Zhang, W. Luo, J. Yang, X. T. Yuan, J. Zhang, "Sparseness Analysis in the Pretraining of Deep Neural Networks". In IEEE Transactions on Neural Networks and Learning Systems, Vol. 28, No. 6, June 2017.

[15] Y. Qian, F. Li, J. Liang, B. Liu, Ch. Dang, "Space Structure and Clustering of Categorical Data". In IEEE Transactions on Neural Networks and Learning Systems, Vol. 27, No. 10, October 2016.

[16] Y. Xiao, B. Liu, Zh. Hao, "A Maximum Margin Approach for Semisupervised Ordinal Regression Clustering". In IEEE Transactions on Neural Networks and Learning Systems, Vol. 27, No. 5, May 2016.

[17] B. Gu, V. S. Sheng, K. Y. Tay, W. Romano, Sh. Li, "Incremental Support Vector Learning for Ordinal Regression". In IEEE Transactions on Neural Networks and Learning Systems, Vol. 26, No. 7, July 2015.

[18] P. Navrat, L. Molnar, "Curricula Transformation in the Countries in Transition: An Experience from Slovakia". In IEEE Transactions on Education, Vol. 41, No. 2, May 1998.

[19] S. Nalintippayawong, K. Atchariyachanvanich, "IT Management Status in Public Higher Education Institutions in Thailand". In IEEE ICIS 2016, June 26-29, 2016, Okayama, Japan.

[20] J. I. Godino – Llorente, R. Fraile, J. C. Gonzales de Sante, V. Osma – Ruiz, N. Saenz – Lechon, ""Design for All in the Context of the Information Society": Integration of a Specialist Course in a Generalist M.Sc. Program in Electrical and Electronics Engineering". In IEEE Transactions on Education, Vol. 55, No. 1, February 2012.

[21] M. Dolores Cano, "Students' Involvement in Continuous Assessment Methodologies: A Case Study for a Distributed Information Systems Course". In IEEE Transactions on Education, Vol. 54, No. 3, August 2011.

ANALYSIS OF DATA TO DETERMINE CUSTOMER LOYALTY IN THE EDUCATIONAL SPHERE USING KOHONEN MAPS

I. R. Gallyamova, N. I. Yusupova, O. N. Smetanina, L. Y. Uzbekova
Department of Computer Science and Robotics, Ufa State Aviation Technical University, Ufa, Russia
e-mail: g_irida@mail.ru, yussupova@ugatu.ac.ru, smoljushka@mail.ru, uzbekovalili@mail.ru

Abstract: *This paper analyses the loyalty of consumers of the university services (i.e., university students). If students are not only interested in receiving a degree, but also proud of the university, recognizing their social responsibility for it, being ready to lead an active life within the university, then the university gets an advantage among its competitors. This study investigates the loyalty of students with regards to the USATU. The main method of investigation was a questionnaire for students. The results of the research can be used by universities when they offer their services on the market.*

1. Introduction

The growing intensity of competition in the market of educational services, along with increasingmarket power of consumers in education make it necessary to develop a new model of identifying loyalty to the university from its direct consumers,the students.

In the field of educational services, just as in any sphere of the economy, it beneficial to study customer loyalty. However, unlike consumer loyalty in the economy, the region in question has a number of features that must be taken into account when speaking about the evaluation of the loyalty of university students. Consumers of educational services are constantly involved in the educational process and, in the final analysis, influence the activity of the university, processes in it, its image (reputation), etc. Formation of students' loyalty to the university is a multi-stage process, determined, first, by the terms of training; from course to course, loyalty can significantly differ. Constant updating of the student body does not allow talking about such a concept as "repeated services" (unless, of course, it is a question of additional educational services or additional (second higher) education) [2].

In the article, questions of the intelligent analysis of the collected personal data are considered. The results obtained should make it possible to identify the most significant factors influencing the loyalty of students, the nature of the answers to the questionnaire of students who have a similar level of satisfaction with the university, and therefore, to make decisions about increasing loyalty with the use of this knowledge.

The issues related to the term «loyalty», the formulation of the task of data processing and the formation of rules of logical inference for supporting decision-making, the data processing scenario are reflected in detail.

2. Related work

It is important to note that the study of customer loyalty, coming from foreign countries, is intended for trade and public consumer services. Studies in this area were pioneered by foreign scientists [5]. There are many viewpoints regarding the topic of loyalty, the most prominent of which are shown in Table 1.

Table 1. *Comparison of existing definitions of the term «loyalty»*

Definition of loyalty	Autor
«Regular (repeated) acquisition of a product of this brand, based on a long acquaintance with it and a favorable attitude towards it»	John R. Rossiter, Larry Percy [4]
«When a product is attractive to consumers in a degree sufficient to make a repeat purchase»	William D. Wells [5]
« A measure of consumer loyalty to the brand». In the opinion of David A. Aaker, loyalty shows the degree of probability for switching the consumer to another brand, especially when it is undergoing changes in price or some other indicators. With increasing loyalty, the tendency of consumers to perceive the actions of competitors decreases	David A. Aaker [6]
«Preferring one brand at every purchase of a product». Gordon R. Foxall notes that «the consumer follows the model of re-purchase, because it is this brand that satisfies his needs quite well, or because he has a personal attachment to the brand». In addition, «this brand provides all the desired benefits, corresponds to the lifestyle of the consumer or its character (image) is the same as the nature of the consumer»	Gordon R. Foxall [7]
«Devotion to your source of value». A loyal buyer does not change the source of values, that is, from time to time acquires a product or service, and recommends it to his environment	Frederick F. Reichheld [8]

The term loyalty means a positive attitude towards a product, service, or company, which manifests itself in a positive response (recommendations, reviews, etc.).

Researchers approach the problem from many different points of view. The problems of the management tools for the competitiveness of goods and services of the enterprise are devoted to the work of many domestic and foreign scientists and specialists. Among the researchers are, just to name a few: Ansoff [9], C.B. Barsukova, M.I. Gelvanovsky, A.L. Denisova, A.A. Dynkin, P.S. Zavyalov, T. Kono, V.M. Kudrov, Zh.Zh. Lamben, A. Marshall, G. Mintzberg, N.K. Moiseeva, V.V. Ovichinnikov, M. Porter, A. Strickland, N.V. Telnykh [10], J. Thomson, V. Feldman, P.A. Fathutdinov, I.P. Cherednichenko, and others.

The theoretical basis for control of relations with customers on the basis of loyalty programs, specifically the building of loyalty programs have been developed and analyzed in the works of these foreign researchers: D. Aaker, S. Brown, S. Butcher [12], N. Boodcock, R. Goldsmith, T. Goki, K. Gronroos, S. Coils, F. Reyheld, V. Sesser, G.Wollard, G.Foxhall, G. Hofmayer, G.L. Bagiyev, V.P. Lukin, Z.V. Kulikova, O.V. Firsanova,O.Y. Tsuvenskaya, I.P. Shirochenskaya.

A number of researchers [1] investigated the questions regarding loyalty among alumni, giving greater attention to increasing the quality of education and the reputation of the university, in order to increase funding and employment of graduates, because of a significant increase of the competition.

Considering the issues of the tools used to analyze data related to loyalty, either directly or indirectly, it should be noted:

– statistical tools for a hybrid model of interaction between alumni and the university, which allows us to assess the situation at a given

time and concentrate on factors that affect the loyalty of both students and alumni to the university – modeling structural equations [1],

– a toolkit in the form of a semantic analysis of bank customers' feedback to identify the key of the text, and as a result, calculation of the NPS index for the bank, comparing the dynamics of the index with the values of the indices of banks that are industry leaders in servicing small businesses [13]

– tools in the form of probabilistic-static models and methods to assess the quality of services based on the opinions of consumers [14], etc.

Despite a lot of attention being devoted by specialists to research, theoretical and practical methods of forming loyalty among customers, many of its aspects have not been sufficiently investigated until recently: the factors which form loyalty towards educational institutions have not been systematized, the criteria determining the degree of influence of the loyalty program on the attitude of students towards the university have not been investigated.

3. Data processing tasks for determining customer loyalty

The actual task is to analyze data to determine the loyalty of consumers of educational institutions with the subsequent development of the mechanism for its management.

To solve this problem, it is required to use data that is aimed at identifying students' satisfaction and develop means of processing that data to identify factors that have the greatest impact on students' loyalty, and possible clusters that combine questionnaires (learners) with similar answers to questions. The authors attempt to reveal what is common between those students who are most satisfied with the university.

There are two ways to obtain information for student loyalty research:

– develop a special questionnaire to identify factors that affect the loyalty of students and conduct a survey for a fairly large number of students;

– use the data from the questionnaire, conducted earlier to solve various analytical problems.

A survey of students at USATU is conducted regularly to address issues related to improving the quality of education [14]. The collected statistical material is used as data for solving various problems.

One of the questionnaires used by USATU experts from the Testing Center Rozanova L.F., Turutina A.D. and Markevich I.A. [14], is aimed at the study of the probabilistic-statistical model of consumers' opinions for the quality assurance system. It helps to identify the requirements of consumers, assess the degree of compliance of these requirements with the performance indicators of the organization.

The difference between the research of the specialists of the USATU testing center from the article conducted by the authors is that in the first case the model of multiple choice to the task of assessing the quality of the educational process from the point of view of trainees.

This article discusses students' loyalty to the university, related in addition to the 7 factors.

In the second case, these questionnaires will be used to determine the loyalty of consumers in the field of education, as well as to determine a list of specific recommendations aimed at increasing students' loyalty to the university.

Analysis of data to determine customer loyalty in the educational sphere in the article implies the solution of the following tasks considering the available statistical materials:

– definition of the technology for processing available statistical data – the results of the questionnaire;

– selection of adequate theoretical apparatus for data processing at each stage of the selected data processing technology;

– conduct data processing based on the chosen technology, taking into account theoretical tools with interpreting the results of students' loyalty.

4. Data processing technologies

Collection of data for analyzing the loyalty of students at the University is made on the basis of the students questionnaire results made earlier for solving other problems.

The questionnaire includes several questions that allow one to identify the impact of certain factors on the satisfaction of the university, among which are:

– quality of the educational process;
– further cooperation with the university;
– maintenance of educational process;
– extracurricular work with students;
– level of motivation;
– accessibility of information on the educational process;
– social work of the university.

The algorithm for processing the questionnaire data assumes the following sequence of steps. It should be noted that the first three steps constitute the preliminary preparation of data for their processing.

Step 1. Formation of an initial data set in the form of questionnaires presented in natural language, in the form of an Excel table and checking for the completeness and correctness of filling the fields.

Step 2. Forming the results of the questionnaire as a data set in the text format .csv, which is digitized with the regular expression [^ 0-9; \ n] as a function of replacing Notepad ++ with an empty cell to clear the text content.

Step 3. Introduction of an additional criterion for the level of loyalty for data processing. According to the selected factors, various sets of questions are compared with the answers and a scale is introduced for assessing the students' satisfaction from 1 to 100. Next, a rating assessment of the options for answering questions on students' satisfaction – from positive (100) to negative (1).

Graduation of the loyalty criterion from 1 to 4 is introduced depending on the satisfaction (1 – dissatisfied, 2 – partially satisfied, 3 – satisfied, 4 – absolutely satisfied). To each questionnaire is assigned the evaluation of expert opinion on student satisfaction, depending on the quantitative ratio of factors.

At this stage, the preliminary processing of the results of the questionnaire is completed and the data is ready for further analysis. To perform the data mining, the Deductor platform was chosen.

Step 4. Data validation is performed using the «Data Quality Assessment» handler to determine the degree of their suitability for solving the problem of student loyalty analysis. The fitness indicator is a value of 1. If the data is less than 1, then it is necessary to identify the cause of unfitness and outline ways to improve the quality of the data.

Step 5. Using methods of factor analysis on the preliminary prepared data of the questioning of students with the purpose of reducing the dimension.

Step 6. Correlation analysis to assess the dependence of the output data fields on input factors and eliminate insignificant factors. The principle of correlation analysis consists in the search for such

values, which are least correlated (interrelated) with the output. Such factors can be excluded from the resulting data set with practically no loss of useful information. The criterion for the decision to exclude is the threshold of significance. If the correlation (degree of interdependence) between the input and output factors is less than the significance threshold, then the corresponding factor is discarded as insignificant.

Step 7. Cluster analysis of the obtained normalized data using Kohonen maps. The card is based on the neural network of Kohonen, whose architecture is represented by two layers of neurons: input and output.

The mechanism of clustering using Kohonen self-organizing maps in the analytical platform Deductor is represented by the following: in accordance with the values of the input parameters (values of loyalty factors), the sampling element correlates with the corresponding cluster.

Self-organizing feature cards allow representing the results of clustering in the form of two-dimensional maps, where the distances between objects correspond to the distances between their vectors in a multidimensional space, and the characteristic values themselves are displayed in different colors and hues. So, the areas marked in red indicate a high value of the indicator, and the dark blue ones indicate the lowest value of the indicator.

Step 8. Analysis of the obtained clustering results using Kohonen maps for obtaining meaningful conclusions for the considered area of consumer loyalty in the educational sphere and the formation of logical inference rules.

5. Mathematical statement of a problem and theoretical frameworks of processing of data

Let $X=\{X_1, X_2, \ldots, X_m\}$ – a lot of the completed and processed questionnaires, where m is the number of such a questionnaire. Furthermore, each element from a lot of X_i is described vector of the features and additional criteria values:

$$X_i = \{x_{i1}, x_{i2}, \ldots, x_{in}, x_{in+1}\}, \quad (1)$$

where $x_{i1}, x_{i2}, \ldots, x_{in}$ – the details of the preferences of each student (n – number of the attributes); x_{in+1} – student evaluation i (i.e. results from the questionnaire survey).

Required:

– Get a vector of reduced dimension of features:

$$X_i' = \{x_{i1}, x_{i2}, \ldots, x_{il}\}, \quad (2)$$

where l – the number of elements of a vector of reduced dimension. For further processing using a set of X'.

– Identify clusters with $\cup Y_j = X'$, $j = \overline{1,c}$, where Y_j – cluster j, $0 \subset Y_j \subset X'$, $Y_i \cap Y_j = \varnothing$, $i \neq j$.

Thus, it is necessary to find a function f': $X_i' \rightarrow Y_j$, which any element X_i' assigns the cluster number j.

– It is necessary to analyze the characteristics of each cluster, interpret the results obtained in accordance with the semantics of the field.

As an analysis tool, an analytical platform Deductor is used.

The algorithm for factor analysis of this platform is based on the principal component analysis (PCA).

For factor analysis, the initial matrix is used, which is transformed and the eigenvalues λ are calculated. If we consider this procedure from the point of view of geometric meaning, then we can say that in the original space the geometric axes are subjected to rotation. Thus, a new factor is introduced, created by each axis. Then for each factor it is necessary to calculate the value in the new feature space. Due to the reduction in size, only new factors with high values will be left. PCA is based on the following formulas, where λ is the eigenvalue R; R – covariance matrix; V – the eigenvector R.

$$RV = \lambda V. \quad (3)$$

Then

$$RV - \lambda V = 0, \quad (4)$$

$$V(R - \lambda E) = 0. \quad (5)$$

And there is a solution when

$$|R - \lambda E| = 0, \quad (6)$$

where E is matrix unit.

Next, we need to find the determinant for the matrix of dimension, determine the vector V, insert eigenvalue (formula 5) and solve the corresponding systems of equations. The sum of the eigenvalues is equal to the number of variables, and the determinant of the correlation matrix is the product. The eigenvalue is the variance of the axis, the largest is the first and, finally, descent to the smallest value is the amount of information along the axis. The proportion of the variance for this component is as follows: it is necessary to divide the eigenvalue by the number m. The load factors for these components are calculated by dividing the eigenvectors by the square root of the corresponding eigenvalues.

For a complete analysis of data and factors, one should consider the Kohonen maps. The Kohonen map is a powerful self-learning mechanism for data clustering that displays results in the form of compact and two-dimensional maps for interpreting data.

The training sample for cluster analysis is selected

$$X'' \subset X'. \quad (7)$$

The yield is a set of clusters that include similar characteristics

$$Y = \{Y_1, Y_2, \ldots, Y_j\}. \quad (8)$$

Specified distance function between objects $\rho(X_i, X_j)$. It is required to break the training sample into clusters based on the similarity of the characteristics of the object for one cluster and the difference in the characteristics of objects of different clusters. In this case, each object $X_i \in X''$ is assigned a cluster number Y_j. Then each object belongs to one of the classes.

The solution of the problem using Kohonen's self-learning network is based on the fact that the neural network, guided by the structure of input vectors, connects the object to a certain class.

When developing the Kohonen network, the number of neurons at the input (in terms of the number of features of the object) and the output layer (the number of clusters), the learning rate η (by default 0.3) and the stopping criterion (the default error does not exceed 0.05) shall be defined.

When sending object attributes to the input, neurons in the output layer compete for the right to be a «winner». The incoming weights of the «winning» cell are closest to the input image (in the sense of the Euclidean distance). It can be noted that each output cell determines the «proximity» to the input image, and for the winning cell j', the relation $\forall j$ gives

$$|\omega_j - x| \leq |\omega - x|. \quad (9)$$

The winner has the right to regulate its weight in accordance with the following rule of weight change [15]: $\Delta \omega_{ij} = \eta\wedge(j, j')(x_j - \omega_{ij})$, where $\eta\wedge(j, j')$ – the neighborhood function, which is equal to unity for $j = j'$ and decreasing with distance $|r_{j'} - r_j|$ between neurons j u j' in the output layer.

The Kohonen network learning algorithm is determined by the following steps: weight initiation, setting of neighborhood function parameters and learning speed; computation $\forall j$

$$D(j) = \sum_i (\omega_{ij} - x_i)^2, \qquad (10)$$

where ω_{ij} shows synaptic weights between neurons i and j if the stopping rule is not executed.

The next step is to find such an index j', that $D(j') = min$, calculation for all cells j within a certain neighborhood from j' and $\forall\ i$ detecting

$$\omega_{ij}(n) = \omega_{ij}(o) + \eta \wedge (j, j') |x_i - \omega_{ij}(o)|, \qquad (11)$$

where $\omega_{ij}(n)$ and $\omega_{ij}(o)$ – new and old values of training sample weights, then the change in the speed of training. The next step is to reduce the radius of the neighborhood function; check the stop condition.

Thus, this section describes methods for analyzing students' assessments (reviews).

6. Results of experimental data processing

The questionnaire used by researchers contains 44 questions related to different aspects of the university. The student can choose one of the available answers from 3 to 10. Examples of questions are given in Figure 1.

Whether you are satisfied that you learn at this university and on the given direction of preparation (specialty)?
1: Perfectly satisfied
2: Partially satisfied
3: No
4: Not sure

Your opinion about your specialty has changed during the education?
1: Yes, for the better
2: Yes, changed for the worse, disappointed
3: It did not change what I expected, it is
4: Not sure

Do you imagine what you will do after graduation?
1: I'm going to work
2: I will continue my studies in the magistracy
3: I'm going to get additional education
4: I'm going to postgraduate school
5: I will open my business
6: I will go abroad
7: I will not work at all
8: I'm going to join the army
9: Not sure
10: Other

Figure 1. *Examples of questions*

It is important to note that the questionnaire form was not created specifically for this study but it's results were used.

Step 1. This study used 1405 forms filled out by students after checking their responses based on completeness. The results were put into an Excel table. A section of it can be seen in Figure 2.

	A	B	C	D	E	F	G
1				1	2	3	4
2	Department	Course of study	Specialty	Why did you graduate?	Whether you are satisfied that you learn at this university and on the given direction of preparation (specialty)?	Your opinion about your specialty has changed during the education?	Evaluate, please, quality of formation on education program on given direction of preparation (specialities) as a whole.
3	Q01_ Department	Q02_ Course _of_ study	Q03_Specialty	Q04_1	Q05_2	Q06_3	Q07_4
4	6 : FGS	2:05	36 : Mathematics a	3 : Parents (relati	2 : Partially satisfie	3 : It did not chan	3 : Good
5	6 : FGS	2:05	36 : Mathematics a	2 : Diploma is ne	1 : Perfectly satisfie	2 : Yes, changed fc	3 : Good
6	3 : FAEEIT	2:05	68 : Industrial elect	4 : Acquisition of	1 : Perfectly satisfie	1 : Yes, for the bet	4 : Great
7	3 : FAEEIT	2:05	68 : Industrial elect	5 : Higher educat	1 : Perfectly satisfie	1 : Yes, for the bet	2 : Satisfactorily
8	3 : FAEEIT	2:05	68 : Industrial elect	2 : Diploma is ne	1 : Perfectly satisfie	1 : Yes, for the bet	3 : Good

Figure 2. *Example of personal data for analysis*

Step 2. Results of the questionnaire are stored in a .csv file for further processing (Figure 3).

Students	Quality of the educational process												The normalized sum
	Evaluate, please, quality of formation on education program on given direction of preparation (specialities) as a whole.	Evaluate knowledges and skills on speciality produced by You.	How you consider, does formation correspond which gives USATU labour market requirements?	To learn in the USATU	Whether the structure of education program corresponds on given direction of preparation (specialities) your expectations? (All disciplines are present, studying of which, on Your opinion, is necessary for management of future professional work; there is no disciplines duplicating; there is no infringement of disciplines teaching logic and so forth) *	Are you satisfied with the content of the taught subjects?	Please note how often you had to face a breakdown due to the late attendance of teachers.	Please note how often you had to face a breakdown due to lack of discipline among students.	Do you have classes in an interactive form? (business and role-playing games, lecture-dialogue, lecture-presentation, group and scientific discussion, discussion, seminar in dialogue mode (seminar-dialogue), case studies, lecture press conference, lecture with pre-planned errors, problem lecture) *	What problems did you encounter when doing research work?		
Student1	2	3	3	6	4	3	2	3	3	3	29		
Student2	2	3	2	4	2	1	3	2	2	2	49		
Student3	1	1	1	2	1	1	1	2	3	5	60		
Student4	3	3	2	2	2	3	3	1	1	5	44		
Student5	2	4	2	1	1	1	2	1	3	5	51		
Student6	2	1	1	3	2	1	3	1	1	4	58		
Student7	2	2	2	3	2	2	2	1	1	5	51		

Figure 3. *Fragment of processed personal data*

Step 3. For the succeeding data processing, there is a further criterion for loyalty which is the overall score. After relating questions to the given factors, a score is given out of 100 total marks. Every questionnaire is given an expert rating based on an expert opinion of how well the questions relate to the factors.

Figure 4 shows a section of the processed data.

	A	B	C	D	E	F	G	H	I
1	Students	Quality of the educational process	Further cooperation with the university	Ensuring the educational process	Extracurricular work with students	Level of motivation	Accessibility of information on the educational process	Social work of the university	Loyal student
2	Student1	29	24	24	32	65	20	23	3
3	Student2	49	14	26	26	65	30	31	3
4	Student3	60	38	53	58	84	70	62	2
5	Student4	44	14	29	26	77	30	69	3
6	Student5	51	48	21	45	74	70	69	2
7	Student6	58	18	62	32	81	70	62	2
8	Student7	51	71	53	39	77	70	62	2
9	Student8	51	52	44	39	77	60	38	2
10	Student9	56	52	38	39	74	70	46	2
11	Student10	40	29	35	32	77	60	46	3
12	Student11	51	52	56	61	74	70	69	2
13	Student12	44	52	44	45	77	70	62	2
14	Student13	60	52	44	42	74	20	23	3
15	Student14	42	48	29	19	65	40	31	3

Figure 4. *Processed data for analysis*

As a result of the introduction of the loyalty criterion, it was found that in this sample, there were none absolutely satisfied students. Students were split into groups 1 (satisfied), 2 (partly satisfied), 3 (unsatisfied).

In total, 38% of students fall into the class «Satisfied Students», 45% - in the «Partially Satisfied Students» class, and 17% - in the class «Dissatisfied Students», out of 1405 students interviewed.

Thus, during further analysis, 8 factors are taken into account – 7 allocated and the 8th – the loyalty characteristic, that is, a vector of higher dimension is used for analysis.

At this stage, the preliminary processing of the results of the questionnaire is completed and the data is ready for further analysis.

Step 4. When assessing the quality of survey results, the capabilities of the Deductor platform are used. Since the usability indicator is close to one, all the answers to questionnaires are suitable for carrying out factor analysis of data.

Step 5. Based on the results of the factor analysis, a set of data is established, which is included in the resulting set of factors (Figure 5).

Factor analysis × Table × Statistics ×

Relevance threshold 0,50

Variables	Final factors (Varimax method)						
	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6	Factor 7
Quality of the educational p ...							0,9029
Further cooperation with th ...		0,9900					
Ensuring the educational pr ...	0,9422						
Extracurricular work with st ...						0,9357	
Level of motivation				0,9504			
Accessibility of information ...					0,9352		
Social work of the university			0,9815				

Figure 5. *Results of factor analysis*

It is established that all factors have a value greater than 0.9. Thus, we can say that each factor is significant, and the set does not need to reduce the dimension in determining the loyalty of students.

The next stage is the study of a set of questions related to certain factors. Issues that did not play any role for this task were identified. Namely, when the set of questions was correlated with the identified factors, it turned out that the answers to two questions out of 44 did not in any way affect the results of the study of students' loyalty. Thus, the questions «Did you have classes in physical culture? How many semesters did they last?» and «In what form are physical education classes conducted?» are excluded from the subsequent processing of the data.

As a result, the questionnaire questions are grouped into blocks of relevant factors (Table 2).

Table 2. *The structure of the questionnaire «The identification of students' satisfaction with the university»*

Name of question block	Question number of the questionnaire
Quality of the educational process	4; 5; 9; 12; 15; 17; 25; 26; 27; 35
Further cooperation with the university	6; 7; 10; 11
Ensuring the educational process	16; 18; 19; 20; 21; 22; 32; 33; 44
Extracurricular work with students	23; 24; 28; 29; 34; 39; 40; 41; 42
Level of motivation	1; 2; 3; 8
Accessibility of information on the educational process	13; 14; 36
Social work of the university	37; 38; 43

Step 6. The results of the correlation analysis of the data from the survey answers show the degree of influence on the students' loyalty (Figure 6).

	Input fields	Correlation with output fields
№	Field	Loyal student
1	Quality of the educational process	-0,548
2	Further cooperation with the univer ...	-0,338
3	Ensuring the educational process	-0,533
4	Extracurricular work with students	-0,517
5	Level of motivation	-0,455
6	Accessibility of information on the ...	-0,627
7	Social work of the university	-0,444

Figure 6. *Results of correlation with output fields*

Factors can be ordered by the degree of influence on the factor «student loyalty» in descending order:

1. accessibility of information on the educational process;
2. quality of the educational process;
3. ensuring the educational process;
4. extracurricular work with students;
5. level of motivation;
6. social work of the university;
7. further cooperation with the university.

The factor that has the greatest impact on students' loyalty is «Accessibility of information on the educational process». According to the results of the survey, it turned out that students do not reflect on further employment and consider the factor «Further cooperation with the university» as the least significant at the university. Perhaps this is due to the fact that the questionnaires were not subjected to graduate students.

A student at the level of training in the 1– 3 courses is not yet aware of himself as a future representative of the real sector of the economy and does not yet think on this topic, as psychological attitudes are aimed at absorbing new information while studying at the university.

Step 7. Cluster analysis of data is carried out using the Kohonen map. The best result of cluster analysis is a breakdown into 7 clusters. Figure 7 shows a map of Kohonen for revealing the loyalty of consumers in the educational sphere.

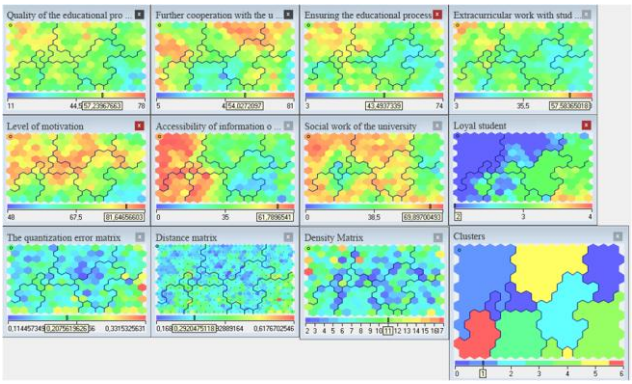


Figure 7. *Results of cluster analysis*

Cluster № 0 includes partially satisfied students, whose responses are characterized by below-average level of extracurricular work with students, average level of social work of the university, accessibility of information about the educational process, quality of the educational process and above the average level of motivation of students, ensuring the educational process and further cooperation with the university.

Cluster № 1 includes satisfied students, characterized by a high level of all the factors of loyalty.

Cluster № 2 includes partially satisfied students, characterized by a lower average level of access to information about the educational process, extracurricular work with students, providing the educational process, an average level of further cooperation with the university, quality of the educational process and above the average level of motivation of students and social work of the university.

Cluster № 3 includes partially satisfied students, characterized by a lower average level of further cooperation with the university, an average level of extracurricular work with students and a high level of other factors.

Cluster № 4 includes dissatisfied students, characterized by a low level of all factors, except for the social work of the university (middle level).

Cluster № 5 includes satisfied students, characterized by a high level of all loyalty factors, except access to information about the educational process (medium level).

Cluster № 6 includes partially satisfied students, characterized by low level of motivation, provision of educational process, secondary level of extracurricular work with students, quality of educational process and above average level of social work of the university, further cooperation with the university and high level of access to information about the educational process.

As a result of the cluster analysis on the values of factors from 1405 questionnaires, 20% of the students surveyed fall into cluster № 1, 18% into cluster № 5, 17% to cluster № 4, 15% to cluster № 2, 14% to cluster № 0, 9% to cluster № 3, and 7% to cluster № 6.

Step 8. In the process of reviewing the results of cluster analysis, a table is compiled in which the boundaries (minimum value and maximum value) of the factor in this cluster are reflected at the intersection of the columns (the received factor) and the rows (cluster number). Considering the results of the questionnaire, it is possible to identify which students entered each cluster. Further, by allocating each student and assigning them a cluster number, it is possible to analyze the values of the initial factors.

Based on the results of the cluster analysis (Figure 4), the boundary values of the factors in each cluster are revealed. Thus, as a result of the interpretation, students are identified in one or the other cluster. Table 3 shows the results of a cluster analysis with ordering by the degree of student satisfaction.

Table 3. *Results of cluster analysis*

	C1 (cluster 1)	C2 (cluster 5)	C3 (cluster 3)	C4 (cluster 0)	C5 (cluster 6)	C6 (cluster 2)	C7 (cluster 4)
Social work of the university	high level	high level	above the average level	the average level	above the average level	above the average level	the average level
Accessibility of information on the educational process	high level	the average level	high level	the average level	high level	below-average level	below-average level
Level of motivation	high level	high level	above the average level	above the average level	below-average level	above the average level	high level
Extracurricular work with students	above the average level	above the average level	the average level	below-average level	the average level	below-average level	below-average level
Ensuring the educational process	high level	high level	above the average level	above the average level	below-average level	below-average level	low level
Further cooperation with the university	above the average level	high level	below-average level	above the average level	above the average level	the average level	below-average level
Quality of the educational process	high level	above the average level	above the average level	the average level	the average level	the average level	below-average level
Student loyalty	Satisfied	Satisfied	Partially Satisfied	Partially Satisfied	Partially Satisfied	Partially Satisfied	Unsatisfied

To identify the loyalty of students, clusters are combined into classes. Clusters № 1 and 5 fall into the class of «Satisfied Students», clusters № 0, 2, 3, 6 fall into the class « Partially Satisfied Students» and cluster № 4 falls into the «Dissatisfied Students» class.

Based on the results of cluster analysis, it is possible to identify the reasons for students' dissatisfaction. In total, the main reasons for getting into the «Satisfied Students» class are those students who do not have low values of the factors. In the class of «Partially Satisfied Students» are low values of 1– 3 factors. In the class of «Dissatisfied students» are students who have noted low indicators of the values of all factors, or all but one of the factors.

7. Conclusion

Questions of research on customer loyalty are of fundamental importance in the improvement of university education. A review of known research in the field of consumer loyalty management has substantiated the urgency of researching issues related to the specifics of forming student loyalty in relation to universities.

The data processing technique for revealing student loyalty involves processing the results of the questionnaire using factorial, correlation and cluster analyzes based on Kohonen maps. The next stage is the development of rules for decision-making for the information support system for managing student loyalty.

As a result of processing data on student questionnaires, 7 factors were identified, among which the factor «Accessibility of information on the educational process» has the greatest impact on student loyalty, and the least influence is the factor «Further cooperation with the university». Additionally,nthe eighth factor, «Loyalty of the student» was highlighted, as some comprehensive assessment of the university's satisfaction.

According to the results of the cluster analysis of data on the values of the factors, 7 clusters were identified, later combined into three classes: «Satisfied Students», «Partially Satisfied Students», «Unsatisfied Students», and reasons for dissatisfaction of students were analyzed.

The offered approach for the analysis of the questionnaire data to reveal consumer loyalty in educational sphere assumes the need for further research and development.

References

1. N. I. Yusupova, O. N. Smetanina, and L. M. Iskhakova. “Models and methods of the processed in-formation in the Alumni management”, Vestnik VGTU, vol. 8, no. 1, pp. 17-21, 2012. (in Russian)
2. Gallyamova I.R. “Information support for decision-making in the field of consumer loyalty management”. Actual problems of science and technology: materials of the XI All-Russian Winter School-Seminar of students, graduate students and young scientists (with international participation), Ufa, USATU, vol. 1, pp. 51-54, 2018. (in Russian)
3. Smetanina O.N., Gallyamova I.R. “Information support of decision-making at management of loyalty with use artificial intelligence technologies”. Problems of modern science and education. № 7 (127), Moscow, 2018. pp. 26-28. ISSN 2304–2338
4. John R. Rossiter, Larry Percy. “Advertising Communications & Promotion Management. Second edition”. New York, N.Y.: McGraw-Hill, 1997. (Also, Japanese translation, 2000; Russian translation, 2001; Chinese translation, 2004).
5. William D. Wells, Sandra Moriarty, Nancy Mitchell, Charles Wood. “Advertising & IMC: Principles and Practice (11th Edition) (What's New in Marketing)”, Pearson; 11th Edition, 2018.
6. David A. Aaker. “Building Strong Brands”, Simon & Schuster UK, 2010.
7. Gordon R. Foxall, Ronald Earl Goldsmith, Stephen Brown. “Consumer Psychology for Marketing”, Vol. 1, Cengage Learning EMEA, 1998.
8. Frederick F. Reichheld. “The Loyalty Effect: The Hidden Force Behind Growth, Profits, and Lasting Value”, Harvard Business Review Press; Revised ed. edition, 2001.
9. H. Igor Ansoff. “Strategic Management”, Springer, 2007.
10. Cherednichenko I. P. “Psychology of management”, Rostov n/D: Feniks, 2004. (in Russian)
11. Jean-Jacques Lambin, Isabelle Schuiling. “Market-Driven Management: Strategic and Operational Marketing”, Palgrave Macmillan; Third Edition, Revised edition, 2012.
12. Stephan A. Butscher “Customer Loyalty Programmes and Clubs (2nd Edition)”, Routledge; 2nd Edition, 2016.
13. N. I. Yusupova, M. V. Chumakova “Semantic analysis of information for decision-making in the management of customer loyalty in the banking sector”, Proceedings of the 5th All-Russian Conference «Information Technologies for Intelligent Decision-Making Support», Volume 1, May 16-19, Ufa, Russia, pp. 36-41, 2017.
14. Rozanova L.F., Turutina A.D., Markevich I.A. “The probabilistic-statistical model of consumers' opinions for the quality assurance system”, Proceedings of the 3rd International Conference “Intelligent Technologies for Information Processing and Management” (ITIPM' 2015), Volume 2, November 10-12, Ufa, Russia, pp. 43-47, 2015.
15. Krichevsky M.L. “Intellegent methods in management”, M.L. Krichevsky. SPb.: Piter, 2005 (in Russian).

THE STRUCTURE OF SECURE SYSTEM FOR COLLECTION, STORAGE AND PROCESSING OF TELEMETRIC INFORMATION ON THE SLATE OF AIRCRAFT SUBSYSTEMS

V.V. Berkholts, A.M. Vulfin, M.B. Guzairov, A.I. Frid, K.V. Mironov
Department of Computer Science and Robotics, Ufa State Aviation Technical University, Ufa, Russia
e-mail: torina4@yandex.ru, vulfin.alexey@gmail.com, guzairov.murat@gmail.com, frid46@mail.ru, mironovconst@gmail.com

Abstract:*The problem of aircraft subsystems state identification and data transmition to enterprise-developer is highly actual. Enterprise-developer can give recommendations for quick eliminate defects during aircraft operation and analyze the causes of faults. The intruder can interfere with process of collection, storage and processing of telemetric information. In the paper the structure of secure system for collection, storage and processing of telemetric information on the state of aircraft subsystems is proposed.*

1. Introduction (style CSIT-Title2)

The arising malfunctions and pre-failures of the on-board equipment of the aircraft can be diagnosed on the basis of the processed telemetric information (TMI). This allows specialists of ground technical services to plan repair and preventive measures based on an assessment of the current state of the equipment.

Collected and processed TMI will allow Enterprise-manufacturer (EM) experts to provide support to ground services engineers in decision-making in the event of technical failure of the aircraft units and modules.

The ability to transmit information about the actual state of the on-board subsystems in the process of operation to the manufacturing enterprises will make it possible to increase the operational efficiency of the aircraft in the regular mode. As well, it will greatly assist in the investigation of incidents, failures and cyberattacks.

The purpose of the research is to increase the protection of the automatic system for collecting, storing and processing TMI on the state of on-board airborne subsystems. The increase in security is planned through the analysis of the application of modern (including intelligent) technologies for the protection and processing of TMI. It is based on the analysis of modern technologies for the protection and processing of TMI.To achieve this goal, it is necessary to provide a structural diagram of a secure system for the collection, storage and processing of telemetric information on the state of the aircraft subsystems on the basis of the modular principle

2. Analysis of the problem of secure collection, storage and processing of TMI in distributed information system

Automated information system (AIS) of ground maintenance services is a set of software and hardware needed to receive, store and process information about the state of aboard complex technical device (CTD) on board the aircraft. AIS is a territorially distributed system. It combines the infrastructure of information systems of ground-based service stations and the information system of the EM through secure communication channels. The TMI acquisition is realized by reading the CTD status log on board the aircraft during technical inspection and maintenance at ground stations using wireless and / or wire sensor networks.

system (SAS) capable of detecting errors in the network configuration, possible routes of cyberattacks of various categories of violators determining critical network resources and ensuring the selection of an adequate security policy.he projected system of TMI collection, storage and processing is a platform for modelling possible actions of violators in the form of attack graphs. Subsequent checking of graph properties and calculation of security metrics forms one of the components of the integral indicator of the system's overall dependability.

3.Development of a structural diagram of a secure system for collecting, storing and processing TMI on the state of the on-board subsystems of an aircraft

The generalized structure of the territorially distributed hierarchical system for the collection, storage and processing of TMI is presented in Figure 1. TMI comes from the sides of aircraft based on ground service stations.

The level of TMI collection by the systems of ground service stations is the realization of the sensor network of the primary TMI collection from the output interfaces of the on-board aircraft systems using wire and wireless sensors.

Preliminary storage of accumulated data at the level of primary accumulation and preparation of TMI for transmission to a part of AIS EM is implemented.

The creation of a secure channel through global communication networks and the transfer of TMI to a part of the AIS EM is realized at the transmission level of accumulated data. Organization levels of reception and distribution of information at the enterprise are realized according to the three-layer CISCO model. [4].

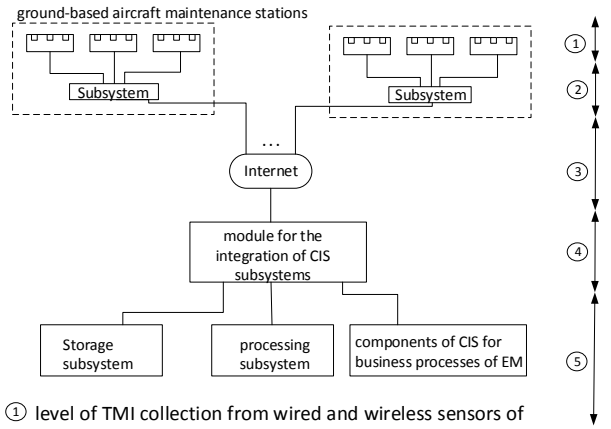
There is a level in the corporate information network of EM. It includes subsystems for storage and processing of TMI. Also, there is a segment designed to support and implement the business processes of the enterprise.

The protected system for collecting, storing and processing TMI on the state of the on-board subsystems of the aircraft is based on the modular principle and contains rather large subsystems with a high degree of connectivity of components inside and a sufficient degree of autonomy at the level of interaction of the subsystems themselves. Each level and subsystems is built on the basis of organizational principles, specific to each task and are regulated by existing documents.

4.Development of the subsystem's structure for TMI collection and storage at ground-based aircraft service stations

In modern practice industrial networks are not isolated they use common information transfer technologies. Most Industrial Ethernet protocols do not have built-in security mechanisms. As a result, any simple cyberattack can lead to negative consequences.

Therefore, the problem of the industrial networks security is actual.



① level of TMI collection from wired and wireless sensors of ground-based aircraft servicing systems
② level of primary surveillance and preparation of TMI for transmission to the AIS of the manufacturer's enterprise (AIS EM)
③ the level of data transmission over secure channels through global data networks in the AISEM
④ level of organization of reception and distribution of TMI on EM
⑤ level of storage and processing of TMI in the CIS

Fig. 1. *- Generalized structural diagram of a protected system for collecting, storing and processing TMI*

It is necessary to be guided by the normative documents of the international and federal standards to ensure the security of subsystems implementing the first two levels of the proposed structure. It is necessary to take into account the requirements [5, 6] to design the subsystem of the wireless sensor network for collecting TMI.

These documents provide detailed information about the reference system of ensuring information security and contain recommendations for identifying and analyzing network security risks, a review of measures and tools of control. Also, these papers summarize the issues related to the implementation and operation of the measures of network security monitoring and management tools.

It is necessary to implement methods of protection of industrial monitoring systems considered in [7,8,9]. These documents provide a detailed description of the information protection organization to carry out industrial activities using industrial monitoring systems.

The physical architecture of the TMI collection and storage subsystem at the ground-based aircraft maintenance stations is constructed in accordance with [10] and is shown in Fig. 2.

Mechanisms for collecting and storing a large size of TMI about the state of individual units and elements of an aircraft should take into account the actively developing concept of the industrial Internet of things (IIot). IIoT is deployed network consisting of a large number of devices provided with a set of sensors that communicate with each other by means of thin and short wireless connections. The first step is to collect data from the sensors. One of the most promising solutions is the low-power and low-transmission protocol IEEE 802.15.4 or IEEE 802.15.4e. The IEEE 802.15.4 and IEEE 802.15.4e protocols and their architecture layers are subject to IETF standards [11]. A small effective range is sufficient to transmit data within the ground-based service station.

It is proposed to use heterogeneous wired (physical interface RS-485) and wireless sensor networks (IEEE 802.15.4, IEEE 802.15.4e) for the collection of TMI based on protocols using the built-in mechanisms of Modbus over TCP to protect the transmitted data (stream encryption)

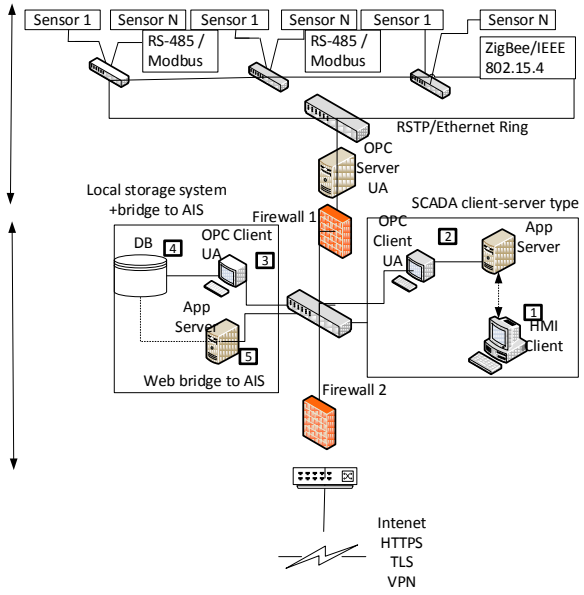


Fig. 2. *- The subsystem of data collection and storage at aircraft service stations*

The heterogeneous industrial network includes communication lines, switches, interface converters. The least reliable elements of this system are communication lines, so redundancy is necessary.

Reservation methods based on the improved RSTP protocol have too long switchover time [12]. The basis of modern industrial solutions is the use of networks with a ring topology (Ethernet ring) [12].

To organize TMI collection in conditions of the possibility of horizontal scaling and use of heterogeneous network, it is necessary to use an open standards-based OPC (Openness, Productivity and Connectivity) infrastructure. Infrastructure defines a set of programming interfaces for several logical (program) objects as well as the methods (functions) of these objects. It is proposed to use the unified architecture of OPC Unified Architecture (OPC UA) [13].

This architecture is a platform-independent service-oriented architecture that combines all the functionalities that exist in the OPC Classic specification and is compatible with OPC Classic. The choice of this architecture is due to the possibility of efficient transport of high-level structured data and the ability to securely access OPC servers through firewalls. Particular importance is the built-in implementation of TMI protection mechanisms using encryption and authentication mechanisms at the level of interacting devices in the industrial network.

Client-server implementation of the mechanism of access and operational storage of TMI implies the use of dedicated application servers and web clients that allow to administer the subsystem. The Web application deployed on the application server provides secure access to the AIS EM. To build a secure web application you need to use two-way authentication at the web server level. Currently, the most secure two-way authentication technology is authentication using the TLS protocol. Interaction at the web application level allows creating a secure channel for transferring the accumulated telemetry data to the AIS EM for further analysis. In detail these questions were considered in the previous works of the authors [14].

The security of the first two levels of the proposed structure is ensured by the introduction of protective measures and the use of architectural templates conditioned by normative documents of the international and federal standard in the field of information and industrial safety. A distinctive feature is the use of service-oriented architecture and client-server implementation of the mechanism for

accessing and storing TMI with the help of a set of web-applications.**F**

5.Development of the subsystem structure for receiving, storing and processing TMI in AIS EM

The organizations of reception and distribution of TMI at the enterprise are realized according to the three-level CISCO [4] model and correspond to core layer level and distribution layer of the network, which are a complex of network devices (routers and switches) providing redundancy of channels and high-speed Data transmission between different segments of the distribution level (Figure 3).

Protection of network infrastructure is an important part of the IS architecture.

The use of the CISCO model in the design of the Security Architecture for Enterprise (SAFE) architecture allows to take into account the modern experience of designing and deploying secure networks based on the defense in depth principle against external and internal attacks [15].

As a part of the corporate information network a level is distinguished that includes subsystems for storing and processing TMI as well as a segment designed to support and implement business processes of the EM (Figure 4).

To ensure storage of TMI and efficient access to accumulated volumes of data it is necessary to build a fault-tolerant storage system using MySQL replication mechanisms such as Master-Master and Master-Slave with the introduction of a caching proxy server and a separate server that performs cluster monitoring functions.

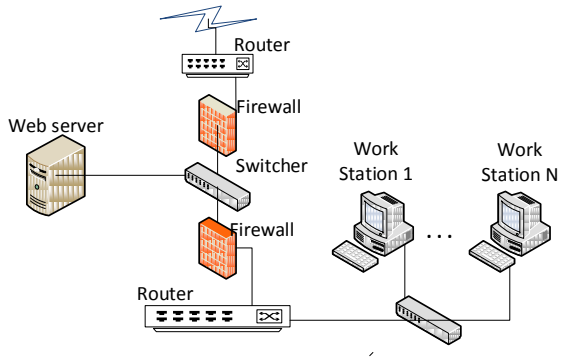
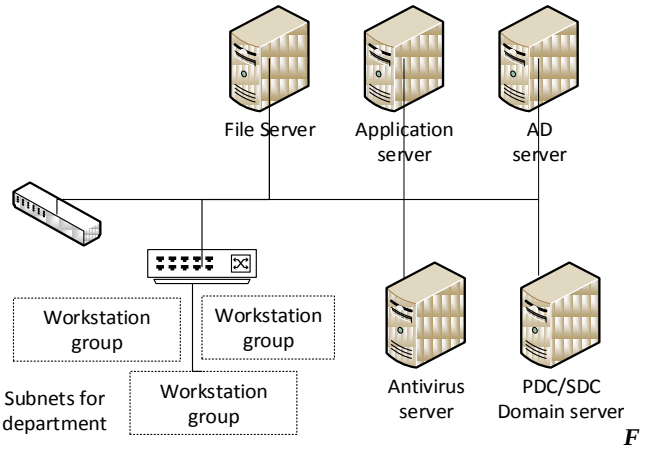


Fig. 3 - Structure of the subsystem for receiving, storing and processing TMI in AIS EM



***ig. 4** - Structure of the segment of the corporate information network support and implementation of business processes EM*

F

Such a scheme (Figure 5) allows to distribute the load on the database between several servers, increases the fault tolerance of the system as a whole and provides the deployment of a platform for efficient processing and analysis of accumulated TMI data [16].

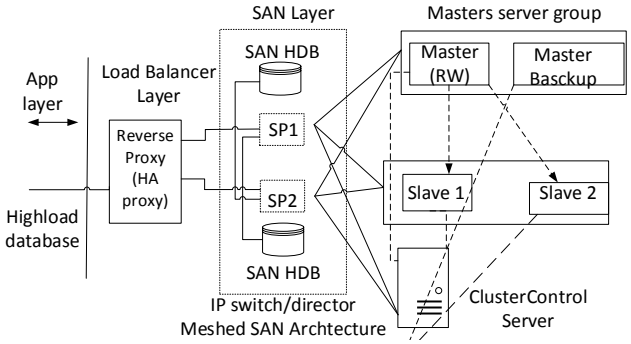


***Fig. 5** - Structure of the storage subsystem TMI with fault tolerance functions*

Accumulated volumes of TMI are proposed to be located on external storage devices (disk arrays) and to organize a storage area network (SAN) of the "meshed fabric" type. This architecture includes a set of cells where each switch is connected to all the others. If one of the inter-commutator links (ISL) fails the connectivity of the network is not violated. SANs are characterized by the provision of network block devices (via Fiber Channel protocols) [18].

Implementation of TMI processing algorithms requires solving a number of tasks related to the design and deployment of an appropriate infrastructure for the processing of accumulated data. There are many tools for distributed processing of large amounts of accumulated data (frameworks: Hadoop, Apache Spark, ClickHouse, ElasticSearch, Splunk Free) [18, 19, 20, 21, 22]. The proposed structure of the TMI distributed processing subsystem is shown in Figure 6.

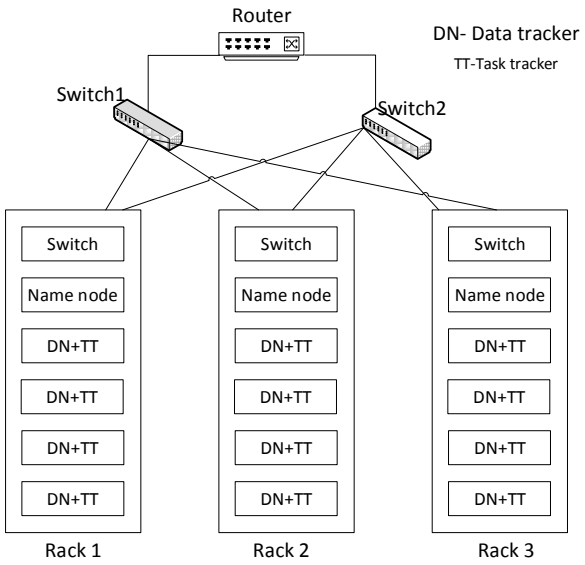


***Fig. 6** - Structure of the distributed processing subsystem TMI*

The main element of the distributed processing system of TMI is the distributed file system. The most popular is HDFS [18]. Additional measures to ensure the confidentiality of stored data is encryption at the level of individual columns of the database. To audit access to large data you must use the Database Activity Monitoring [DAM] solutions.

The next element of the large data processing system is the distributed programming and machine learning infrastructure. [18]. The structure of this infrastructure includes the machine learning

tool MLLib which allows to implement tools for intellectual analysis of accumulated TMI data.

The network infrastructure of the CIS segment and the data distribution segment are designed taking into account the requirements of the enterprise security architecture (SAFE) and allows the implementation of defense in depth from external and internal attacks.

Fault tolerance of the storage subsystem is ensured by using MySQL replication mechanisms such as Master-Master and Master-Slave. This allows you to additionally solve the task of load balancing on the database. It also provides the ability to deploy a site for efficient implementation of processing and analysis of accumulated TMI.

Conclusion

A block diagram of a protected system for the collection, storage and processing of telemetric information on the status of aircraft subsystems based on the modular principle is proposed. A distinctive feature of the proposed solution is that it contains sufficiently large subsystems with a high degree of connectivity of components inside and a sufficient degree of autonomy at the level of interaction of the subsystems themselves. Each subsystem is built on the basis of organizational principles that are specific to the problem being solved and is regulated by existing regulatory documents to ensure certain aspects of the system's dependability.

The TMI collection, storage and processing system is a platform for modeling possible actions of violators in the form of attack graphs with subsequent verification of graph properties and calculation of security metrics. This approach allows us to solve complex problems of ensuring trouble-free operation, fault tolerance, availability, security, serviceability, observability of the TMI transmission system from the aircraft.

The security of the subsystems of the first two levels is realized by the introduction of protective measures and the use of architectural templates, fixed by normative documents of the international and federal standard in the field of information and industrial safety. A distinctive feature is the use of service-oriented architecture and client-server implementation of the mechanism for accessing and storing TMI with the help of a set of web-applications.

The network infrastructure of the CIS segment and the data distribution segment is designed taking into account the requirements of the enterprise security architecture (SAFE) and allows the implementation of defense in depth from external and internal attacks. Fault tolerance of the storage subsystem is ensured through the use of MySQL replication mechanisms such as Master-Master and Master-Slave which allows additional solve the tasks of load balancing on the database and provides the ability to deploy the site for efficient processing and analysis of accumulated TMI data.

Recommendations for the development of architecture and the selection of solutions at various levels of the proposed secure system for collecting, storing and processing TMI will allow specifying the methodology for constructing protected systems of this type.

Acknowledgments

This article is supported by RFBR grant № 17-07-00351.

References (style CSIT-Title2)

1. Frid A.I., VulfinA.M., Zakharov D. Ju, Berkholts V.V., Mironov K.V.Architecture of the security access system for information on the state ofautomatic control systems of aircraft // Proceedings of the 19thInternational Workshop on Computer Science and Information Technologies (CSIT'2017), Germany, Baden-Baden, October 8-10,2017, T.№2, P. 21-27
2. Frid A.I., VulfinA.M., Zakharov D. Ju, Berkholts V.V., Mironov K.V. Architecture of the security access system for information on the state of automatic control // Proceedings of the 19thInternational Workshop on Computer Science and Information Technologies (CSIT'2017), Germany, Baden-Baden, October 8-10,2017, T. №1, P. 16-19
3. Frid A.I., VulfinA.M., , Berkholts V.V. Analysis of the methods of constructing information attack models for the system of telemetric information transmission, Proceedings of the Information Technology Intelligent Decision Support (ITIDS2018), Russia, Ufa URL: [http://itids.ugatu.su/index.php/itids/itids2018/paper/view/50]
4. Cisco 3 layer Hierarchical Network Model | Core |Distribution | Access URL [http://w7cloud.com/cisco-3-layer-hierarchical-network-model-core-distribution-access/]
5. ISO / IEC 27033-1-2011 "Information technology (IT)" Methods and means of ensuring safety. Network security. Part 1. Overview and concepts "and ISO / IEC 27033-3-2014" Information technology (IT);
6. ISO / IEC 27033-1-2011 "Information technology (IT)" Methods and means of ensuring safety. Network security. Part 3. Reference network scenarios. Threats, design methods and management issues
7. Order FSTEC of Russia of March 14, 2014 No. 31 "On the approval of the requirements for ensuring the protection of information in automated control systems of production and technological processes in critical facilities, potentially hazardous facilities, as well as objects posing an increased risk to life and health of people and environmental protection »
8. Order FSTEC of Russia of 23 March 2017 No. 49 "On Amending the Composition and Content of Organizational and Technical Measures to Ensure the Safety of Personal Data when Processing in Personal Information Systems approved by the Order of the Federal Service for Technical and Export Control of February 18, 2013 No. 21
9. Requirements to ensure the protection of information in automated control systems of production and technological processes in critical facilities, potentially hazardous facilities, as well as objects posing an increased risk to life and health of people and the environment, approved by the order of the Federal Service for Technical and Export Control March 14, 2014 No. 31 »
10. NIST 800-82 and ISA / IEC 6244
11. 802.15.4e-2012: IEEE Standard for Local and Metropolitan Area Networks – Part 15.4: Low-Rate Wireless Personal Area Networks (LRWPANs) Amendment 1: MAC Sublayer, IEEE Std., Apr. 2012.].
12. Encyclopaedia of ACS TP , Reservation of industrial networks URL [http://www.bookasutp.ru/Chapter8_3.aspx]
13. OPC Unified Architecture URL: [https://opcfoundation.org/about/opc-technologies/opc-ua/]
14. Guzairov M.B., Frid A.I., Vulfin A.M., Berholts V.V. Protected access to the database on the status of automatic control systems (ACS) by aviation GTE via the web application // Information and Security. 2017 Vol. 20 No. 3 (4). P.410-413
15. CISCO URL [http://www.cisco.com/go/ safe]
16. MySQL Repication URL: [https://severalnines.com/resources/tutorials/mysql- replication-high-availability-tutorial]
17. Data Center Storage Evolution URL [https://www.siemon.com/us/white_papers/14-07-29-data-center-storage-evolution.asp]
18. Hadoop Apache URL [http://hadoop.apache.org/]
19. Apache Spakr URL [http://spark.apache.org/]
20. ClickHouse URL [https://clickhouse.yandex/]
21. Elastic URL [https://www.elastic.co/]
22. Splunk Free URL [https://www.splunk.com/en_us/software/features-comparison-chart.html]

WEB-BASED INFORMATION SYSTEM FOR THE IMPACT ASSESSMENT OF ENERGY ON THE GEOECOLOGY OF THE REGION

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

***Abstract**: This article formulates and briefly discusses the formulation and proposed approaches to solving the problem of the impact of energy on the geoecology of the region. Applied semantic technologies are also presented, for example, such as ontological and cognitive. The main attention is paid to the Web-oriented information system - its structure, main components and approaches to implementation, as well as one of the most important components of the system –Contingency Management Language CML.*

1. Introduction

This work is being executed as a part of international project, which is supported by EAPS and RFBR, for research of impact assessment of energy on region geoecology. Aim of the project is to develop methods and technologies for assessing the impact of energy on region geoecology. Geoecology is understood as interdisciplinary scientific direction, which combines research of composition, processes, physical and geochemical fields of the Earth's geospheres as habitats of humans and other species [2]. 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 with the purpose of preserving for the present and future generations of people a productive natural environment. More details description of geoecology, quality of life, as well as proposed approaches and methods are discussed in the article by Massel L.V. "Intelligent decision-making support in energy and ecology in view of the quality of life".

For development of tools for intelligent decision-making support in areas of energy and ecology, project uses methods of geoinformation technologies based on 3D-geovisualization [3], methods for critical infrastructures research [4], methods for decision-making support, methods for knowledge engineering, methods of object approach and also authors' methods of situational management, semantic modelling and intelligent tools for its support. For this project it is proposed to develop and adopt authors' methods of creating ontological space, knowledge in area of energy; methods of semantic modelling in energy, 3D-geovisualization methods and methods of visual analytics with elements of cognitive graphics, and also methods for development intelligent systems for support strategic decision-making in energy [4]. Also, authors will use experience gained by Russian and international scientists [5, 6, 7, 8].

This article deals with Web-oriented information system (WIS) - which integrates mathematical and semantic methods and tools for impact assessment of energy on geoecology of region, database, knowledge base and geoinformation system. It is supposed to use the results of the authors of the project, carried out earlier to study the problems of energy security: tools for semantic modeling, Geocomponent, tools for working with the knowledge base and the Repository.

2. Semantic technologies

In general, semantic modelling is understood as information modelling, based on extraction of main terms (concepts) of subject area and connections between them. Thus, semantic models can be classified as ontologies, as well as semantic networks and infologic ER-models [4]. Since 2013, in articles of L.V. Massel it is shown that cognitive, event and probabilistic (based on Bayesian network) modelling can be referred as semantic technologies. Methods and tools of this modelling are actively developed by the team, represented by the authors [4].

2.1 Ontological modelling

According to Guarino, ontological modelling is understood as "specification of conceptualization" or identification of the basic terms (concepts) of the domain and links between them and their description. Ontological modelling was discussed more detailed in articles of T. Gruber [9], N. Guarino [10-11] and others, as for Russia – in articles of T.A. Gavrilova, Yu. A. Zagorulko, L.A. Kalinichenko, Kogalovsky M.R., Serebryakov V.A. and others. Works performed under the leadership of Massel L.V. in this area were presented, in particular, at the ZONT conference, which is held in Novosibirsk.

2.2. Cognitive modelling

Cognitive modelling is understood as cognitive models creating, or, in other words, cognitive maps (directed graphs) in which vertices are corresponding to factors (concepts) and edges are connections between factors (positive or negative) depending on the nature of the cause-effect relationship. The mathematical apparatus for constructing cognitive models is the graph theory. The fundamentals of cognitive modeling were developed in due time by Wang Hao (1956), R. Axelrod (1976), D. A. Pospelov (1981).

This scientific direction was developed in the works of E.A. Trakhtengerts, at present it is actively developing in the V.A. Trapeznikov Institute of Control Sciences of the Russian Academy of Sciences (Abramova N.A, Kulba V.V., Kulinich A.A., Maksimov V.I.) for the analysis of influences in managing weakly structured situations.

2.3. Event modelling

By event modelling is understood the construction of behavioural models and as objects of modelling can be considered both people and technical objects. The task of the event modelling method is to track the sequence of events on the model in the same order as they would occur in the real system. As an event modelling tool, the Joiner-Network (JN) is used - one of the varieties of algebraic networks proposed by L.N. Stolyarov [12]. The peculiarity of JN is that it provides both a graphical representation and description in the form of logical formulas, the processing of which can be automated. Works performed under the leadership of Massel L.V. in the field of cognitive and event modelling, are reflected, in particular, in [4].

2.4. Probabilistic modelling based on Bayesian network

Bayesian networks of trust are graphical models of probabilistic and cause-effect relationships in a set of variables that are described by a directed acyclic graph whose vertices are variables and the edges show conditional dependencies between them. The basis of this tool is the Bayes theorem. Fundamentals of the tool of graphical probability models (in particular, Bayesian networks) were developed by J. Perl, R.J. Cowell, and others. In Russia this apparatus was considered by V.I. Gorodetsky, A.L. Tulupyev, and others. Works done under the leadership of Massel L.V, are described, in particular, in [13].

3. Web-oriented information system

As it was said in the introduction, it is planned to create a Web-oriented information system (WIS) for the project implementation. Fig. 1 shows the architecture of this system.

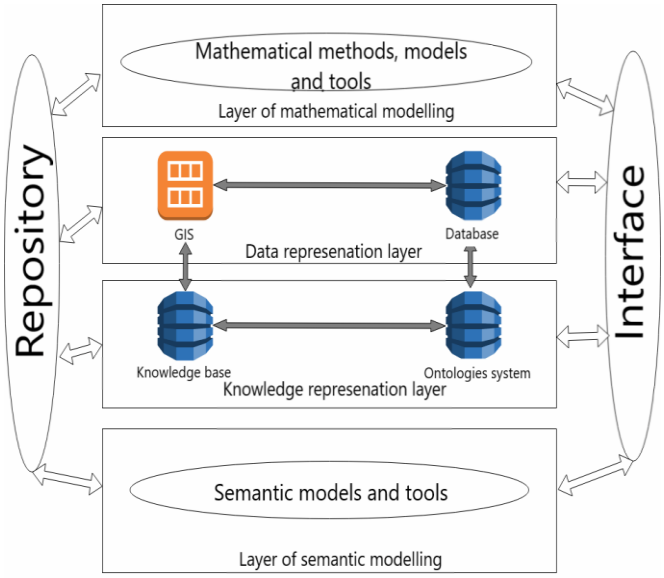


Fig. 1 *Architecture of Web-oriented information system*

This system will be based on the integration of the geo-information system, mathematical and semantic modelling tools, tools for assessing the impact of energy on the geoecology of the region, databases and knowledge base. Some components of the system will be implemented using multi-agent technology.

As it seen from the architecture mentioned above, system will consists of four layers:

- **Layer of mathematical modelling** - contains programs, developed on the basis of selected methods and models, for calculating the amount 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 (population density) in the considered territory.
- **Data representation layer** - integrates geoinformation 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
- **Knowledge representation layer** – contains a knowledge base that stores 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 designing a database.
- **Layer of semantic modelling** – contains semantic models for describing the interconnection of factors that determine the quality of life, taking into account anthropotechnical factors: energy supply and the impact of pollutants from energy companies on the environment

The metadescriptions of the information which is presented on all four levels are stored in the Repository. When implementing the user interface, it is supposed to apply the situational calculus [15] and the components of the Contingency Management Language CML, which is one of the most important components of the WIS and will be considered in the next section.

4. Contingency Management Language CML

One of the basic components of WIS is the CML – Contingency Management Language. Fig. 2 represents its metaontology.

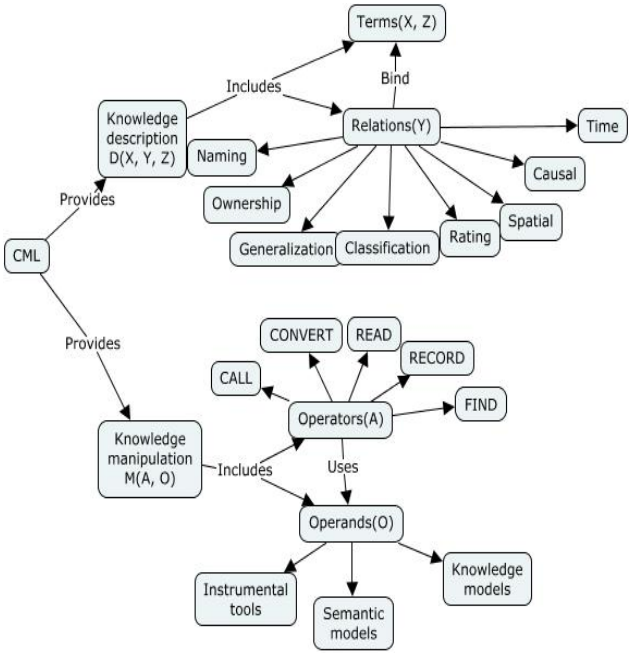


Fig. 1. *Metaontology of Contingency Management Language*

It includes two parts: (D, M): (1) tools for knowledge description D (descriptions of situations, scenarios and actions) and (2) tools for knowledge manipulation M (call of tools and usage of both semantic and knowledge models).

This language will be based on the integration of situational management and situation calculus. The possibility of such integration was considered by the authors in [16]. Now let's consider how, with the help of this integration, it is possible to describe some of the concepts of the situation description component.

- **Relations of naming:**
  <object> <has> <name>
  <object> := <physical object> | <software component> | <informational component>
  For this type of relations it's possible to use object characteristics, introduced in situation calculus.
- **Relations of classification:**
  <situation> <has type> <initial | transitional | target>
  <control action | measure> <has type> <preventive | operative | liquidation>
  Situation calculus aims to classify situations and actions. To classify threats it makes sense to use the dictionary of names and terms.
- **Spatial relations:**
  <situation> <happens at> <location>
  to describe these relations it is proposed to use the dictionary of names and terms.
- **Time relations**:
  <situation> <happens at> <time>
  When describing these relations one might stick to the concept of time, because in the situation calculus it sets the onset time of specific situation.

Tentative analysis shows that in such a way it is possible to consider all proposed relationships and to compare them with the basic concepts of situational calculus.

For the program implementation of this language, it is suggested to use the program complex Semp-TAO, developed at the A.P. Ershov Institute of Informatics Systems of the SB RAS. At the core of the complex is a knowledge representation model that combines classical and modern tools of representation and processing of knowledge: frames, production systems, semantic networks [17]. The main tool of representing declarative knowledge is the semantic network. Processes of information processing and output are specified in the form of a system of products that work on a semantic network.

Based on this model, the language of representation and processing of knowledge was developed, which is intended for specification of applications in various subject areas and has a rich set of tools for this purpose [18]. The language allows to represent the concepts of the domain in the form of classes of objects and relationships, create new data types and specify the required output and data processing processes in applications, and includes tools necessary for creating user interfaces and working with text files. Next, an example of an ontology will show how, with the help of this complex, it is possible to describe an ontology.

First, we'll create the following base classes: **commonObject** – describes common object in system. It has next fields "name" (object name), "*global_id*" (object id in database), "*owner_id*" (model's identifier to which object belongs) и "*inner_id*" (object id inside model).

commonRelationStruct – describes common relation between objects. It has following fields "*global_incoming_id*" (global identifier of the object from which the connection is made), "*inner_incoming_id*" (inner identifier of the object from which the connection is made), "*global_outer_id*" (global identifier of the object to which the link is connected), "*inner_outer_id*" (inner identifier of the object to which the link is connected) и "*value*" (value of this relation).

commonMap - describes common model object. It has following fields: "*global_id*" (model's id in database), "*name*" (model's name) и "*description*" (model's description).

Now, consider how one can use the classes above to describe ontology. Preliminarily, we'll create class, which describes ontology concept with following fields: "*incoming_relations*" (set of object incoming relations) and "*outgoing_relations*" (set of outgoing relations). This class is inherited from *commonObject* class. Then, ontology class is described, it consists from single field "*objects*", which is a set of ontology concepts. Also, this class extends *commonMap* class.

Thus, it can be seen that using the Semp-TAO language it is possible to describe the concepts of the contingency management language included in the knowledge description component.

5. Results and discussion

The article proposes an approach to solving the problem of monitoring the impact of energy on the geoecology of the region, based on the integration of mathematical models, GIS technologies and modern intelligent technologies in the framework of a Web-based information system (WIS). The proposed WIS is 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 objects and the spread of pollution.

Conclusion

- The applied technologies of semantic modelling are briefly described.
- The architecture of the Web-based information system is presented and the components of this system are described.
- One of the most important components is considered – Contingency Management Language CML and approaches to its implementation, based on the integration of concepts of situational management and situation calculus using the software package Semp-TAO, was shown.

Acknowledgments

The results were obtained with partial financial support of RFBR grants № 18-57-81001, № 16-07-00474, № 18-07-00714, № 17-07-01341.

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. Richard J. Huggett "Geoecology: An Evolutionary Approach", Routledge, 1995.
3. 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.
4. 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.
5. "Energy, Climate Change, Environment". *International Energy Agency*, 2016.
6. 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
7. Hussey, K., and J. Pittock "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
8. 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.
9. Gruber T.R. "A Translation Approach to Portable Ontologies". *Knowledge Acquisition*, 1993, 2(5):199-220, 1993.
10. Guarino N., D. Oberle, S. Staab, "What Is an Ontology?". *Handbook on Ontologies*, 2009, 1-17. 10.1007/978-3-540-92673-3_0.
11. Guarino N. "Formal Ontology in Information Systems". In: *Proceedings of FOIS'98*, Trento, Italy, 6-8 June 1998. pp. 3-15.
12. Stoljarov L.N. "The philosophy of event modeling by the example of the emergency scenario in energy sector" In: *International Conference "Information Technologies in science, education, telecommunications and business": Proceedings.* Ukraine. Gurzuf, 2010. pp. 197-200.
13. Massel L.V., E.V. Pjatkova "Application of Bayesian belief networks to intelligent support in research of energy security problems". *Bulletin of the Irkutsk State Technical University*, 2012. 2: 8-13.
14. 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
15. 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
16. Massel L.V., Kuzmin V.R. "Situation calculus as development of semiotic approach for creating intelligent decision-making system". *Bulletin of NSU:Information technologies*, 2017; 4(15): 43-52.
17. Zagorulko Yu.A., Popov I.G. "A Software Environment based on an Integrated Knowledge Representation Model". In: *Perspectives of System Informatics. - Proc. of the Andrei Ershov Second Internat. Conf. PSI'96*, Novosibirsk, Russia, June 25-28, 1996. - pp. 300-304.
18. Zagorulko Yu.A., Popov I.G. "Knowledge representation language based on the integration of production rules, frames and a subdefinite model". *Joint Bull. of the NCC & IIS. Ser.: Comput. Sci*., 1998; 8:81-100

THE METHOD OF MATHEMATICAL MODELING OF PROCESSES TO IDENTIFY, ESTABLISH CHARACTERISTICS AND RELATIONSHIPS OF THE CONNECTIONS OF SOCIALLY SIGNIFICANT PHENOMENA

V.V. Antonov[1], Ya.S. Mikhailova[1], V.A. Kolesnikov[2], L.V. Chernyshova[2]

Department of computer science and robotics, Ufa state aviation technical University, Ufa, Russia[1]
Ufa Law Institute, of the Ministry of Internal Affairs, Ufa, Russia[2]
e-mail: Antonov.V@bashkortostan.ru, ysm.act@gmail.com, kolesnikov62va@rambler.ru, lidaionova@mail.ru

Abstract: *Prospects for progress in the creation of analytical systems, the purpose of which is the statistical determination of the trends in the spread of socially significant processes and related events, are critical because of the global geopolitical crisis. His imprints are imposed on almost all states, in conditions of which the significance of the influence of negative and positive social phenomena is determined by a steady increase. In this article, the method of mathematical modeling of the process of identifying, establishing characteristics and potential relationships between various processes that make up or have an effect on socially significant phenomena is delineated.*

1. Introduction

The most important component of any socially-oriented automated system is to ensure the effectiveness of its management and use, which is the subject of many studies [1,3,8,12]. The accuracy of strategic decisions at various levels of the power vertical, the policies of financial corporations, the activities of organizations providing security and many other significant components of modern society depend on the systems of the specified designation. The proposed method is designed for initial conditions that presuppose the operation of specified and systematized information. The determination of the characteristics of a set of phenomena and the results obtained from them are the decisive condition for choosing the principles of algorithmization, as well as the methods of calculation at the stages of subsequent investigations. It should be noted that the prerequisites for taking the necessary measures, both for strengthening and easing any consequences of the phenomena of the processes of a designated character, are formed on the basis of the extent of the spread of such phenomena and the tendencies accompanying them. This conclusion can be assumed as one of the central in the context of the overall task of analysis and presented through membership functions. Depending on all interests and combinations of their manifestations, it becomes obvious that the definition of their characteristics is a fundamental step in the calculation.

2. Formal mathematical model of information system

Let us turn to an analysis of the elements of the phenomenon. When using a socially-oriented automated system containing a description of the parameters of the elements, the categorization and distribution of data occurs on the basis of confirmation of the criteria set by the evaluation system [2,4]. Each stage of gathering information and placing it into categories presupposes the rules of categorization, which are the initial and basic for the assessment of the phenomenon as a whole. For the purpose of this, a number of information objects provide for categories of translated information [8,10]. Intermediate, but inalienable, step is computational procedures for formalizing qualitative information in order to prevent a decrease in the quality of the completeness of the data coming from the experts and, as a consequence, the accuracy of the subsequent results. For this task, the fuzzy logic apparatus can be used, as well as for the task of formulating information incompleteness as the next important quality of the input data. In the future, this device can be used to unify and translate a large number of data in the whole system under consideration. For simplicity, we will designate each procedure for unification and translation as preliminary calculations for the main iterations, then the system can be divided into subsystems. A common bridge between the system and all subsystems is categorically-topological logic [2.9] formalizing the general structure of calculations in order to achieve the expediency of procedures for algorithmizing the entire mathematical apparatus. Let us dwell in more detail on the subsequent procedures. Thus, the delineation and processing of information can be carried out by means of technical means, aimed at searching for information on a group of characteristics, performed by parsers, as the most convenient and easy to use in use. During procedures for analyzing the characteristics and prevalence of an event, it is difficult to determine their potential scale as a phenomenon. The detailing of the events accompanying such phenomena predetermines the prospect of taking into account the exact number of constituent elements and allows them to be correlated with the coefficient of significance of this or that determinable factor. Such an analysis can be carried out by means of logical and statistical categorization of the information databases induction, that is, one attribute for prevalence or increase in order to compare with the prevalence of the factor or group of factors. From this it follows that, provided that the details of fuzzy positions are detailed and the conditions of accuracy are satisfied while distinguishing the principles of behavior of socially significant phenomena, the conditions for examining the boundaries of the set of events can be specified [11]. The conclusion is obvious that their group assumes a systematic repetition, which can be treated as the phenomenon mentioned earlier, and the set of processes in the phenomenon presupposes the formation of the object of investigation. Then, the system itself can be considered as a relationship in the object of investigation, which implies the definition of the properties of the isomorphism of sets. However, in order to determine the behavior of the phenomena, additional calculations are necessary, and such problems can be handled by estimation through graph theory in applied tools of the spectrum of graphs on the subsequent delineation of clear and fuzzy cognitive maps. Let's address to algorithm of the basic calculations of cumulative influence of all possible factors on the basic indicators of stability of occurrence of events [13]. Initially, the formation of the system itself is due to a situational sampling of external events that are dominant for isomorphic data. Such a step is envisaged for further accuracy of operation, as well as for the establishment of events, the prevalence of which can be regarded as unambiguous in determining the nature of the events that constitute the essence of the phenomenon itself. For the purpose of this they can be interpreted as oriented graphs and, based on connectivity criteria from data of different influence classes, are translated into matrices by structured relations in the dependency relationships, further formulation can be represented by adjacency and reachability matrices, which not only makes the data unified for further iterations in the derivation of connectivity and strong connectivity in matrices, but also operations of disjunction and conjunction [6].

So, let the sets be given starting from the properties of isomorphism, then, depending on their categories, the graphs $G_1, G_2, ..., G_n$.

For example, there are two graphs

$$G_1 = (V(G_1), \mathrm{E}(G_1)) \quad G_2 = (V(G_2), \mathrm{E}(G_2)) \tag{1}$$

and the sets $V(G_1), \mathrm{V}(G_2)$ do not intersect. Then the union of graphs G_1, G_2 is a graph with a set of vertices $V(G_1) \bigcup V(G_2)$ and a family of edges $\mathrm{E}(G_1) \bigcup \mathrm{E}(G_2)$. It is also possible to form a connection of graphs G_1, G_2 denoted by $G_1 \oplus G_2$ taking their union and joining each vertex of the graph G_1 to each vertex of the graph G_2. Since the isomorphism conditions must be described in graphs, we define connectivity expressions (which exclude incoherence of graphs). We give a connection with the separating set *R*, when the ratio $R \subset V(G_n)$ of a graph G_n can be defined as a vertex connection $k(G_n)$, and the edge connection $\lambda(G_n)$ as $R \subset E(G_n)$ a graph G_n. The conditions for given sets can be considered valid for $\lambda(G_n) \geq k$, that is, the graph *G* is edge-*k*-connected and the minimal edge separating set in the graph *G* contains at least *k* edges for all data sets. In this case, the sets of graphs are oriented and multigraphs are excluded, and the selected connections are derived from the described conditions into a matrix description of structured relations in the following dependencies: strongly connected, one-sided, weakly connected (previously excluded disconnected relations also need to be retained for possible subsequent iterations) [5].

On the basis of the above, we can define the adjacency matrix of a graph G_n in the form of a square matrix $\mathrm{A} = (\mathrm{M}_n \times \mathrm{M}_n)$, since the rows and columns are put in a one-to-one correspondence to the vertices of the set *V*. The value of the element (a_{ij}) of this matrix located at the intersection of the *i*-th row and the *j*-th column is determined by the rule:

$$a_{ij} = \begin{cases} 1, \text{ if } (\mathrm{v}_i, \mathrm{v}_j) \in G \\ 0, \text{ if } (\mathrm{v}_i, \mathrm{v}_j) \notin G \end{cases}$$

The reachability matrix of a given graph G_n is determined by the binary closure matrix with respect to the transitivity of the relation *E* (from the given adjacency matrix of the graph). In this case, information on the existence of paths between the digits of the digraph is preserved. In the formation of the reachability matrix expressions, the condition for using disjunction operations and conjunction of adjacency matrices is valid for subsequent derivation to connectivity matrices and strong connectivity. Depending on the set in digraphs, we obtain the matrix expression by attainability $M^* = M \vee M^2 \vee ... \vee M^n$. We derive matrices for strong and weak connections of the following form: in matrices of strong connection $S_1(G_n) = \{s1_{ij} : i = 1,...,n;\ j = 1,...,n\}$ and connection $S_2(G_n) = \{s2_{ij} : i = 1,...,n;\ j = 1,...,n\}$ we have

$$s1_{ij} = \begin{cases} 1, \textit{if } \mathrm{v}_i \textit{ achievable from } \mathrm{v}_j \textit{ and} \\ \quad \mathrm{v}_j \textit{ achievable from } \mathrm{v}_i \\ 0, \textit{otherwise} \end{cases} \tag{3}$$

$$s2_{ij} = \begin{cases} 1, \textit{ if } \exists \textit{ route connecting} \\ \quad \mathrm{v}_i \textit{ and } \mathrm{v}_j \textit{ and} \\ \textit{if } \exists \textit{ route connecting} \\ \quad \mathrm{v}_j \textit{ and } \mathrm{v}_i . \\ 0, \textit{otherwise} \end{cases} \tag{4}$$

Then we obtain the matrix disjunction of the choice $S = S_1(G_n) \vee S_2(G_n)$.

The subsequent choice by disjunction involves the translation of an array of digraphs of the matrix description of connectivity into matrices by weight functions: $\mathrm{A} = \{(\mathrm{a}_{ij}), \mathrm{i} = 1,...,\mathrm{n};\ \mathrm{j} = 1,...,\mathrm{n};\ \mathrm{a}_{ij} = f(V_i, V_j)\}$, where V_i, V_j are the vertices of the graph, $f(V_i, V_j)$ is the weight function.

In the future, weighted values can be translated into the cognitive map $G_n = (\mathrm{V}, \mathrm{E})$, where the set of vertices *V* - is the totality of all the elements of the event, examined by the given array and obtained by weight functions, and *E* the connection between the elements *V* taking into account the cause-effect relationships between them and their mutual influence .

2.1. Mathematical model for describing the characteristics of an event with a fuzzy cognitive map

As a result, we have a cognitive map of the event in the form of a matrix representation $\mathrm{A} = \|\mathrm{a}_{kl}\|$ whose elements a_{kl} reflect the direct influence of the *l*-th element on the *k*-th element, where the sign a_{kl} shows the "direction" of the effect, its absolute magnitude is the degree of such influence, and the zero value corresponds to the absence of influence. The essential values are determined in dynamics $y_k^t = (y_k^0, ..., y_k^t)$ - the vector of the dynamics of the *k*-th element up to the time *t*, where $k \in M$, $Y(\mathrm{t}) = (\mathrm{y}_1(\mathrm{t}), ..., \mathrm{y}_m(\mathrm{t}))$ is the vector of the values of the elements at time *t*, $Y_t = (\mathrm{Y}(0), \mathrm{Y}(1), ..., \mathrm{Y}(\mathrm{t}))$ is the trajectory dynamics of the event up to the time *t*, *t = 0,1,2…* Then we have a matrix relation of the determined influence in the dynamic characteristics of the cognitive positions sought by the conditions

$$s_{ij} = \begin{cases} 1, \textit{if } \exists \textit{ route connecting} \\ \quad \mathrm{v}_i \textit{ and } \mathrm{v}_j \textit{ and} \\ \textit{if } \exists \textit{ route connecting} \\ \quad \mathrm{v}_j \textit{ and } \mathrm{v}_i . \\ 0, \textit{otherwise} \end{cases} \tag{5}$$

Using the obtained matrix relation, we can determine the detailed weights of digraphs (as the required conditions) and in order to verify the derived characteristics, we obtain an extended interpretation by means of the F-graphs of cognitive maps: $F = \{(V, E)\}$

$$\mathrm{F}(\mathrm{x}_i, \mathrm{x}_j, e_{ij}) = \begin{cases} +1, \textit{if the increase } (\textit{decrease}) \\ \quad \textit{of } x_i \textit{ entails an increase} \\ \quad (\textit{decrease})\ x_j \\ -1, \textit{if the increase } (\textit{decrease}) \\ \textit{of } x_i \textit{ entails am decrease} \\ (\textit{increase})\ x_j \end{cases} \tag{6}$$

i,j=1,2,…,k.

In the weighted signed digraph, the values of the weight coefficients of the arcs are given, in general form $F(\mathrm{x}_i, \mathrm{x}_j, e_{ij})\omega_{ij}$. In particular, the normalized exponent of the intensity of the influence of the previously presented characteristic function with the properties can be used objectively ω_{ij} to the

situation: $1 \le \omega_{ij} \le 1$, $\omega_{ij} = 0$, if x_i it does not depend on x_j and the effect is absent, $\omega_{ij} = 1$ with the maximum positive influence x_i on x_j, ω_{ij} =-1 with the maximum negative influence x_i on x_j.

An important variety of cognitive positions for the system are formalized from clear cognitive maps with an unclear relational description. In this case, the elements of the system are perceived as unstructured or semi-structured, and the graphs are set by fuzzy parameters, and the arcs are given by the relations of their influence on the cause-effect relationships. The main purpose of implementing the fuzzy map device is to exclude the possibility of output beyond the matrix values (or ranges of values in the future). In the final combined versions of formalization, a mixed model of the matrix relation of the determined influence in the dynamic characteristics of cognitive positions is derived. Returning to the given expressions, we can define: provided that the binary relations on the set are fuzzy, the relations between the concepts (as nodes) are represented by fuzzy sets in powers of belonging. The set of relations of influence between all concepts is given by a matrix of fuzzy relations. At the same time, the fuzzy relation is considered as a fuzzy map of the fuzzy set by the influencing results, which, under system modeling, makes it possible to represent these concepts under external influence. In this case, it is fair

$$\begin{aligned} &G = G_1 \bigcup G_2, \quad G_1 = (V(G_1), \mathrm{E}(G_1)), \\ &G_2 = (V(G_2), \mathrm{E}(G_2)) \quad \Rightarrow \\ &G \ = (V(G_1) \bigcup V(G_2),\ E(G_1) \bigcup E(G_2)) \end{aligned} \tag{7}$$

For a simplified interpretation, we translate the associated set into an adjacency matrix, where the element ω_{ij} of the given matrix is at the intersection of row *i* and column *j* to estimate the effect of the elements x_i on x_j. With expert intervention, the values can be translated into percentages. For further transformations, we can configure the cognitive maps in the model, provided by the conditions for sequential variation of the set by the balance of sign digraphs in models (the formal mathematical model of balance, is described conditionally for the purpose of regrouping). This is achieved by transfer into symbolic model delay, repeated description sheet, transfer to the binomial model taking into account the potential uncertainty certain experts sets (probability space graphs with *N* vertices if $N \to \infty$, where each pair of *N* is connected with a probability *p* and the result obtained by the number of edges be a random variable with an expected value $N \to \infty$ If G_0 - graph with vertices P_1, P_2, …, P_N, and *n* edges, the probability of obtaining it by means of this process will be $P(G_0) = p^n x(1-p)^{0.5N(N-1)}$. Different interpretations of vertices, edges and weights on edges, as well as various functions that determine the influence of links on elements, lead to the conclusion of a set of grouped elements, as a result of which the graphs are localized, weighed and added to the adjacency matrices with a subsequent iteration cycle. In the future, by calculating the spectrum of graphs, we can reconstruct the entire chain of events and further identify the phenomenon itself. Transforming spectra, we are dealing with invariance, even after groups of matrix permutations and transformations. The spectrum is a finite sequence of numerical invariants, its distinctive feature is the informative quality of the content of graph structures. This indicates that the essential information for us (or a significant part of it) is contained in the spectrum and in the end we can use its spectrum instead of digraph, since the finite sequence of numbers is easier to process in automatic mode, which proves applicability for the systems discussed.

In order to verify the results, the entire structure obtained is additionally checked for isomorphism properties and the data is entered into an intermediate covariance matrix for the likelihood estimation of algorithmic processes [7] (we exclude the description, since system algorithms were not initially specified). The result of the study is the dependence of the change in cognitive maps in the fuzzy representation with further interpretation on the spectrum of graphs and subsequent processing of intermediate values of fuzzy positions.

3. Conclusion

On the basis of the research, a new approach is proposed to form the structure of a software analytical complex, processing socially significant information objects (events) based on the use of fuzzy cognitive maps. The conclusion is that a significant part of the required information is contained in the digraph spectrum and, ultimately, we can use its spectrum instead of the digraph, since the final sequence of numbers is easier to process in automatic mode, which proves applicability for the systems discussed. This approach provides the ability to describe in the language of set theory the relationship between modules (objects), programs, information system and business processes.

List of used sources

1. Antonov V.V. The method of designing an adaptive software package on the basis of the methodology of the category formal model of an open subject domain // Vestnik UGATU, 2015. Vol. 19. No. 1. P. 258-263.
2. Antonov V.V., Kulikov G.G., Antonov D.V. The set-theoretical IP model for multidimensional analytical analysis, which meets the requirements of the data warehouse // Bulletin of the USATU, 2012. Vol. 16. No. 6 (51). C. 189-201.
3. Antonov V.V., Kulikov G.G. Semantico-mathematical language for describing the structure of an intellectual system based on fuzzy logic [Text] / Antonov V.V., Kulikov G.G. // International journal "Software products and systems". Tver. 2011, No. 3 (95). Pp. 33-35.
4. Antonov V.V., Kulikov G.G. Method of constructing a mathematical model of a domain / V.V. Antonov, G.G. Kulikov // Vestnik SSEU: nauch. journal. Samarsk. state. econ. University. 2010. № 5 (67). Pp. 10-14.
5. Blokh A.Sh., Graph-schemes and algorithms: Textbook for the phys.-math. fact. ped. in-tov. - Mn: Vysh. shk., 1987. - 144s.
6. Vokuyeva, T.A., Calculation of the mutual interaction matrix of the cognitive map, Izvestiya Komi, Scientific Center, Ural Branch of the Russian Academy of Sciences. Issue 3 (11). Syktyvkar, 2012.
7. Goldblatt R. Toposes. Categorical analysis of logic. - Moscow: Mir, 1983. - 488 p.
8. Konev K.A. Ontological concept of solving problems in applied scientific research // Informatization of education and science. 2016. No. 1 (29). Pp. 147-155.
9. Konev K.A. Ontological specialization of the field of scientific knowledge // Informatization of education and science. 2017. No. 3 (36). Pp. 95-105.
10. Masel L.V. Fractal approach to the structuring of knowledge and examples of its application // Ontology of design. 2016. Vol. 6. No. 2 (20). Pp. 149-161.
11. Novikov D.A., "Cognitive Games": Linear Impulse Model, Probl. managed. 2008. № 3; Autom. Remote Control. 71: 4 (2010).
12. Feder E. Fractals. - Moscow: Mir, 1991. - 254 p.
13. Tsvetkovich D.M., Dub M., Sachs H., Spectra of graphs. Theory and application // Kiev: Nauk. dumka, 1984. - 384 p.

DEEP LEARNING BASED PERSON BIOMETRIC IDENTIFICATION

M.R. Bogdanov[1], I.N. Dumchikova [1], I.S. Dokuchaev [1], A. Dumchikov [2]
Department of Computer Science and Robotics Ufa State Aviation Technical University Ufa, Russia[1]
Department of Applied Informatics, M.Akmullah named after Bashkir State Pedagogical University, Ufa, Russia[2]
e-mail: bogdanov_marat@mail.ru, 1282342@gmail.com, kvmail@mail.ru, redfoxufa@gmail.com

Abstract: *The article is about the biometric person identification using Deep Learning methods. The issues related to the preprocessong of the signal, the extracting of biometric features and classification are considered. The implementation of the LSTM Deep Learning model, which is a kind of Recurrent neural network, is discussed. Factors affecting the accuracy of biometric identification are being studied*

1. Introduction

Modern authentication technologies are subject to hacking. It is considered that biometric identification is more reliable. Biometric authentication includes some tools to identify individuals by using their unique, measurable physiological and behavioural characteristics for example face recognition, iris, retina, voice, and fingerprint. However, biometric authentication systems are increasingly facing the risk of being tricked by biometric tools such as anti-surveillance masks, contact lenses, vocoders, or fingerprint films [1]. There are a few ways for increasing of biometric authentication reliability. First one is intended to improve an existing technology. For example, Tan and Schuckers from Clarkson University reported about vulnerability of fingerprint scanners because hackers are used artificial fingers made from Play-Doh, gelatine and silicone molds [2]. Authors proposed an anti-spoofing detection method based on ridge signal and valley noise analysis, to quantify perspiration patterns along ridges in live subjects and noise patterns along valleys in spoofs. Based on these features, separation (live/spoof) is performed using standard machine learning tools including decision trees and artificial neural networks. Authors show that the performance can reach 99.1% correct classification overall [2].

Bogdanov et al. investigated factors affecting the accuracy of biometric identification based on the using of electrocardiograms [3], [4]. PTB (290 subjects) [5] and ECG-ID (90 subjects) [6] digitized ECG databases were used under researching. Authors compared the accuracy of electrocardiogram recognition by various methods of machine learning (table 1) [4].

Table 1.

Accuracy of electrocardiograms recognition by various methods of machine learning

Database	Gaussian Mixture Models	Random Forests	Support Vector Machines	Logistic Regression
PTB	0.80	1.00	0.95	0.83
ECG-ID	0.66	0.98	0.93	0.79

Table 1 show, on the one hand, that the random forests method is more precise and on the other hand, we can see that the accuracy of recognition also depends on the registration hardware used. To increase the accuracy, you can tune the model hyperparameters. So, when using the default parameters in the SVM method, we have an accuracy of 0.93 (in the case of the ECG-ID database), however, by tuning the parameters on grid we found the best values for 'kernel' parameter is 'rbf', for 'C' parameter is 10, for 'gamma' parameter is 0.99. Using parameters found we can increase the accuracy to 0.99. The same is true for other methods [3].

Interesting results were obtained when studying the effect of sample size on the accuracy of recognition (table 2) [4].

Table 2: *The effect of sample size on recognition accuracy (PTB database)*

Sample size	Gaussian Mixture Models	Random Forests
10	0.96	1.00
30	0.82	1.00
50	0.80	1.00
70	0.82	1.00
90	0.77	1.00
290	0.80	1.00

Table 2 shows that the accuracy of recognition by the method of Gaussian Mixture Models increases with an increase in sample size on the plateau (0.80), while the accuracy of recognition by the method of random forests does not depend on the sample size.

Bogdanov et al. showed that the recognition accuracy is practically independent of the recording time [4].

2. Related work

The next approach involves the use of new biometric identification technologies. Recently there have been published papers on biometric identification using electroencephalograms [1].

Another way for improving of accuracy of recognition is using of multimodal biometric systems. The main objective of multimodal biometrics is to reduce one or more false accept rate, false reject rate and failure to enrol rate [7]. Bumbarov et al. proposed a face and ECG based multimodal system [7]. Under face recognition the challenge is face detection in scenes with complex background, different lightning conditions, changes in pose, human expressions etc. [7]. To overcome these problems author proposed a framework including Face Detection, Subspace Projection and Classification. Face Detection is intended to locate a human face in a scene and extract it as a single image. For performing of Face Detection authors used a combination of two classifiers (Haar-like features' cascade of weak classifiers, intended for fast detection of face-like objects. The second classifier is a Convolutional Neural Network (CNN) used for filtration of falsely detected faces). Subspace Projection based on Principal Component Analysis (PCA) and Spectral Regression (SR) algorithms was used for dimensionality reduction of the detected facial images, when represented as vectors in high-dimensional Euclidean space. Support Vector Machines algorithm was used for classification of detected faces [7]. The Framework was tested on image database from the Computational Vision at the California Institute of Technology. Best accuracy was 95.26%.

Bumbarov et al. when working with electrocardiograms used ECG signals collected from 28 individuals using own ECG registration hardware. The sampling rate is 512Hz and the resolution is 12bit. Authors trained the system using subsets from these signals. The testing was done two weeks later in order to be sure the time invariance of the features [7].

Authors extracted morphological features from electrocardiograms which present simultaneously amplitude and temporal characteristics of the ECG waves as well as their shape based on linear projections in subspaces [7]. Authors used Principle Component Analysis (PCA) and Linear Discriminant Analysis

(LDA) [7]. All experiments with kernel versions of PCA and LDA were made using the Gaussian kernel function. The best accuracy was 95.7% [7].

When combining of Face Recognition and ECG based Personal Identification authors used the output probabilities of Radial Basis Neural Network classifier (for ECG recognition) and LIBSVM library for calculating the output probabilities of SVM classifier (for Face Recognition) [7]. Bumbarov et al. show that combining probabilities output of ECG identification and Face Recognition framework give accuracy of 99.5% [7].

Zhang et al. proposed a multimodal biometric authentication system including three parts: an Invalid ID Filter Model to block invalid subjects, a Gait Identication Model and an EEG Identication Model [1]. The first two models realize a one-class SVM algorithm and a Recurrent Neural Network based deep learning model, respectively. The third model combines autoregressive coefficients, an RNN structure, and an SVM classifier. The system is trained with a gait dataset of 160,000 samples and an EEG dataset of 108,000 samples. Experimental results show an overall accuracy of 0.983 along with an overall false acceptance rate (FAR) of 0.0 and a false rejection rate (FRR) of 0.019.

Another approach of multimodal biometric authentication proposed Manjunathswamy et al. [8]. Authors used ECG and fingerprints as sources of biometric features. Authors extracted 12 morphological features from electrocardiograms and 2 features (bifurcations and terminations) from fingerprints. Manjunathswamy et al. reported on 82.40% accuracy score for ECG recognition, 92.60% accuracy score for fingerprint recognition and 97.50% accuracy score after fusion of two modalities [8].

3. Motivation and aim

The goal of the researches presented is develop a method of biometric identification using ECG based on Long Short-Term Memory (LSTM) network. The project includes data acquisition, signal preprocessing, feature extraction and classification.

4. Using of artificial neural networks for biometric recognition

Very interesting results may be obtained when using of Deep Learning technologies for Biometric Recognition and classification of cardiac arrhythmia [9]. Usually authors use a Multilayer Perceptron (MLP) neural networks or Recurrent Neural Networks (RNN) for biometric classification. Gawande et al. described a biometric system based on MLP intended for ECG recognition [9]. Authors used lead II for analysis. ECG samples from 8 healthy individuals are recorded regularly during 36 months. Authors extracted 7 statistical and 3 morphological features. Each electrocardiogram contains at least 5 ECG cycles. Seven statistical features have been extracted after decomposing the ECG signal by using wavelet transform. Three morphological feature using Pan Tompkins algorithm are calculated [9]. Multilayer Perceptron Neural Network with single hidden layer is trained with 3 runs and 1000 epochs. The testing results show that the MLP has accuracy of 99.76 during experimentation [9].

Page et al. presented a low-power wearable ECG based biometric authentication system [10]. The approach uses 1-lead ECG data at sample rate of 500 Hz. The main processing block consists of three major stages: 1. Preprocessing, 2. QRS Identification, and 3. User Identification. To achieve the goal of biometric authentication, the system used neural networks for both detecting QRS complex segments and performing user identification. When preprocessing it is filtering and normalization of signal was performed. To QRS complex identification a neural network model consisting of a single 307-node hidden layer with 60% dropout and a tanh activation function was built. Authors reported on 99.54% accuracy. According to Page et al. the model required 2 hidden layers with sizes 577 and 380, dropout rates 76% and 78% and a tanh activation function [10].

Tantawi et al. described ECG based Biometric Recognition system. PTB, MIT_BIH and Fantasia electrocardiogram datasets were used under researching [11]. Authors used as biometric features a RR intervals extracted and decomposed using discrete Daubechies wavelet in wavelet coefficient structures which was reduced by excluding the uninformative coefficients and they are fed into a Radial Basis Functions (RBF) neural network for classification [11]. The best accuracy achieved was 95.89% [11].

Kim et al. decided to use weighted fuzzy membership functions (NEWFM) for biometric authentication using ECG [12]. They extracted biometric features from 73 subjects in the Physionet Database using Haar-wavelet transformation coefficients a4, d3, and d4. Extracted features then were entered into the neural network with weighted fuzzy membership functions for use in order to categorize ECG waveforms in the proposed biometric authentication model [12]. Authors have reached accuracy of 98.32% using proposed method [12].

5. Data

We used a digitized ECG database Physikalisch-Technische Bundesanstalt (PTB) presented by prof. Michael Oeff to physionet.org project [13]. The warehouse includes 549 ECG samples from 290 subjects. Each ECG record included 12 standard leads (i, ii, iii, avr, avl, avf, v1, v2, v3, v4, v5, v6) and three Frank leads (vx, vy, vz). The sampling rate was 1 kHz, the bit depth was 16 bits in the voltage range ± 16.384 mV. Most subjects suffered from various cardiovascular disorders, the control group included 51 healthy subjects [13].

6. Preprocessing

The preprocessing we started with extracting of lead 1 (Fig. 1).

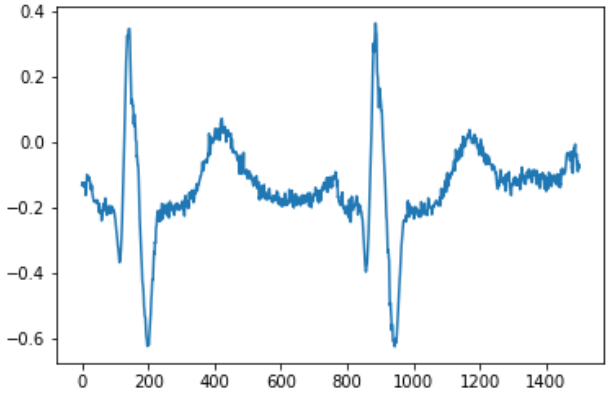


Fig 1. *The first lead of ECG*

As we can see from Figure 1, the original signal is redundant and contains artefacts. To eliminate the redundancy, we applied a low-pass filter with a cutoff frequency of 50 Hz (Fig. 2).

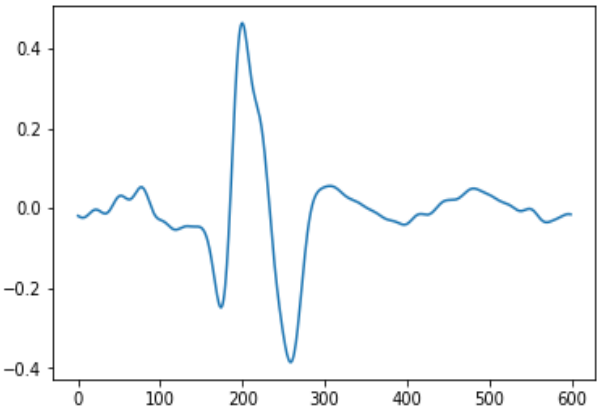


Fig 2. *The signal after low-pass filter*

Further, the signal was cut into fragments of 600 values. The fragments thus obtained were synchronized over the R peaks of the QRS complex (Fig. 3).

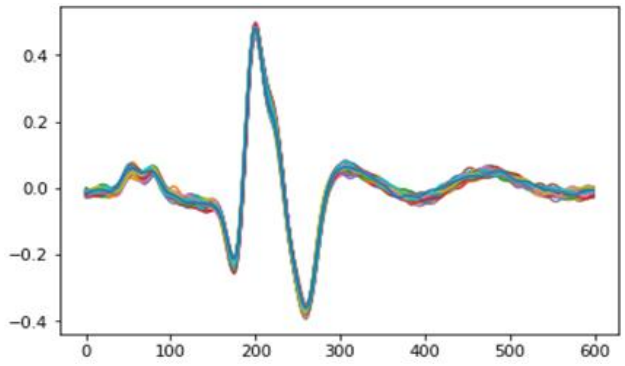


Fig 3. *The ECG signal after preprocessing*

7. Feature extraction

Under the researching we used a two types of biometric features: 1) temporal and amplitude values of Q, R and S-regions of lead 1 and amplitude values of Q,R and S-regions of lead 3 (9 features in total), 2) entire cardio cycle from lead 1 (600 values). Temporal and amplitude features were extracted according to Bogdanov et. al [4].

8. Classification

8.1. Recurrent Neural Networks

Recurrent neural networks (RNN) are a powerful and reliable type of neural networks and currently belong to the most promising algorithms, since they are the only ones that have internal memory.

Because of their internal memory, RNN is able to remember important things about what they got, which allows them to be very accurate in predicting what will happen next. It is for this reason that they are the preferred algorithm for sequential data such as time series, speech, text, financial data, audio, video, weather and much more, as they can form a much deeper understanding of the sequence and its context compared to other algorithms.

In the RNN, the information cycles through the loop. When it makes a decision, it takes into account the current input, as well as what it learned from its previously received inputs (Fig. 4).

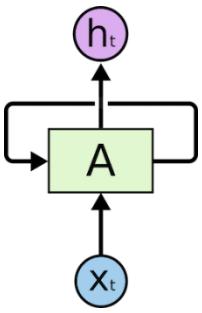


Fig 4. *This loop structure allows the neural network to take the sequence of input [14]*

The image below (Fig. 5) shows the deployed RNN. On the left you can see RNN, which unfolds after the equal sign. It should also be noted that there is no cycle after the equal sign, since different time stamps are rendered and the information is transferred from one time sign to the next. This illustration also shows why RNN can be viewed as a sequence of neural networks.

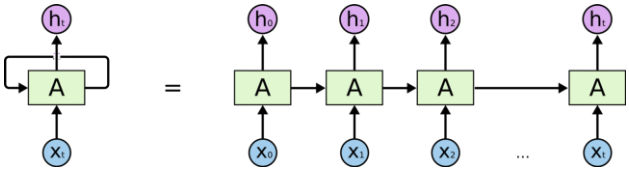


Fig 5. *Unrolled version of the net. First, it takes the $x_{(0)}$ from the sequence of input and then it outputs $h_{(0)}$ which together with $x_{(1)}$ value is the input for the next step. Thus, the $h_{(0)}$ value and $x_{(1)}$ value is the input for the next step. Analogous, $h_{(1)}$ value from the next is the input with $x_{(2)}$ value for the next step and so on. This way, it keeps remembering the context while training [14].*

There are two obstacles that RNN faces. One of the modifications of recurrent networks, which we will consider below, solves these problems.

One of the problems is the problem of "exploding" gradients. The essence of the problem lies in the fact that the algorithm can assign a large value to the scales without any particular reason. Another problem is the disappearing gradients. When gradient values become too small, the model stops learning or studies too slowly. This was a serious problem in the 1990s and it is much more difficult to solve than the first one. Fortunately, it was solved with the concept of a network of long short-term memory (LSTM).

Long long-term memory networks (LSTMs) are a modification of recurrent neural networks, which basically expand their memory. Therefore, it is well suited for studying important experiments that have a very long period of time between them.

LSTM is designed to solve the problems of long-term dependence. Remembering information for long periods of time is practically their default behavior, not what they are trying to learn. LSTMs also have the same chain as RNN, but the repeating module has a different structure. Instead of having a single layer of a neural network, there are four interacting in a specific way.

8.2. Architecture of Deep Learning Model

Architecture of Deep Learning Model is shown in Fig 6.

Depending on the extracted features, the form of the input and output data changes from the ECG. Then the data goes to the first LSTM layer. All outputs of this layer are fed to the next LSTM layer with 256 blocks. After the treatment with this layer goes a fully connected layer. This layer outputs a vector whose length depends on the number of objects to be classified in the database.

Table 3

Results of the experiment

Description of the features submitted to the model	The Model of Deep Learning (all subjects)	The Model of Deep Learning (healthy subjects)
9 features from lead 1 and 3	0.97	0.92
Entire cardio cycle (600 points)	0.99	0.98

As we can see the method using entire cardiocycle as feature table is more precise.

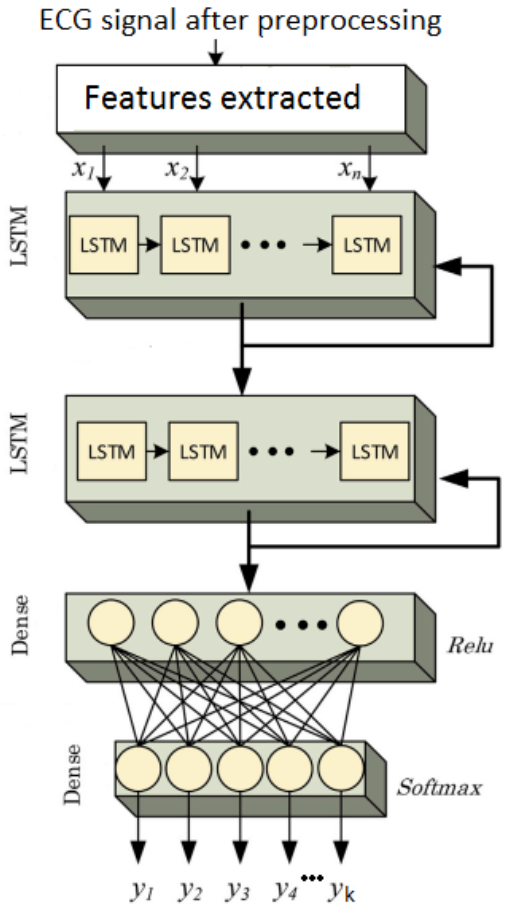


Fig 6. Architecture of Deep Learning Model

8.3. Experiment

For the final experiment, two methods of extracting features were chosen. The first way is to give our model the entire cardiocycle. The second way is to use the main points of the QRS complex of the first and third leads.

Regardless of the way in which features were extracted, the accuracy of the model was calculated using the cross-validation method. For this purpose, the data was divided into 10 parts. In total, the model goes through 10 complete training cycles. At each iteration of the cycle, the network is trained in 9 parts and tested on one. Thus, each part will be used as a test set. For the value of the final accuracy, the arithmetic mean was taken.

Also, regardless of how we extract the features, we need to encode the class labels. Since the output layer of our model is a vector whose length is equal to the number of classes, we should apply the following method: we code our class labels with a unique length vector equal to the number of classes in such a way that only one component is equal to one and the others to zero.

Since there were subjects in our database with different numbers of measurements, it was interesting to see how many records at different time intervals affect the accuracy of the model. It is also worthwhile to conduct the experiment by sampling only healthy individuals. The results of all the experiments are shown in Table 3.

9. Conclusion

The results of the work show that deep learning methods with a larger training sample allow us to find hidden patterns in data that allow us to better classify an object than using features that have been selected manually. Since a high recognition accuracy was obtained and the cardiograms were recorded at different time intervals, this leads to the conclusion that the cardiogram contains invariant features.

In the presence of a huge array of data, deep training methods should be used to detect hidden patterns, rather than using manual selection of features.

References

1. Zhang X., Yao L., Chen K.,Wang X., Sheng Q.Z., Gu T. DeepKey: An EEG and Gait Based Dual-Authentication System. ACM J. Comput. Cult. Herit., Vol. 9, No. 4, Article 39. Publication date: March 2017.
2. Tan B., Schuckers S. Spoofing Protection for Fingerprint Scanner by Fusing Ridge Signal and Valley Noise. https://www.clarkson.edu/sites/default/files/2017-11/spoofingprotection.pdf
3. Bogdanov M.R., Dumchikov A.A, Kartak V.M., Fabarisova A.I. Optimizing Factors Influencing on Accuracy of Biometrical Cardiometry. OPTA-SCL 2018, Omsk, Russia, published at http://ceur-ws.org.
4. Bogdanov M.R, Kartak V.M., Dumchikov A.A., Fabarisova A.I. Factors Influencing Accuracy of Biometrical Personal Identification Based on Cardiograms. DOI: 10.1134/S1054661818030033. ISSN 1054-6618, Pattern Recognition and Image Analysis, 2018, Vol. 28, No. 3, pp. 421–426. © Pleiades Publishing, Ltd., 2018.
5. Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov PCh, Mark RG, Mietus JE, Moody GB, Peng C-K, Stanley HE. PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals. Circulation 101(23):e215-e220 [Circulation Electronic Pages; http://circ.ahajournals.org/content/101/23/e215.full]; 2000 (June 13).
6. Nemirko A.P., Lugovaya T.S. Biometric human identification based on electrocardiogram. Proc. XII-th Russian Conference on Mathematical Methods of Pattern Recognition, Moscow, MAKS Press, 2005, pp. 387-390. ISBN 5-317-01445-X.
7. Ognian Boumbarov, Yuliyan Velchev, Krasimir Tonchev and Igor Paliy (2011). Face and ECG Based Multi-Modal Biometric Authentication, Advanced Biometric Technologies, Dr. Girija Chetty (Ed.), ISBN: 978-953-307-487-0, InTech, Available from: http://www.intechopen.com/books/advanced-biometric-technologies/face-and-ecg-based-multi-modal-biometric-authentication
8. Manjunathswamy B.E., Abhishek A.M, Thriveni J, Venugopal K R, Patnaik L.M. Multimodal Biometric Authentication using ECG and Fingerprint. International Journal of Computer Applications (0975 – 8887). Volume 111 – No 13, February 2015.
9. Gawande P.S., Ladhake S.A. Artificial Neural Network based Electrocardiogram Classification for Biometric Authentication. International Journal of Computer Applications (0975 – 8887). Volume 109 – No. 2, January 2015.
10. Page A., Kulkarni A., Mohsenin T. Utilizing Deep Neural Nets for an Embedded ECG-based Biometric Authentication System. 978-1-4799-7234-0/15/$31.00 ©2015 IEEE.
11. Tantawi M., Revett K., Salem Abd-B., Tolba M. ECG based Biometric Recognition using Wavelets and RBF Neural Network. Recent Advances in Information Science. ISBN: 978-960-474-304-9.
12. Kim Ho J., Lim J.S. Study on a Biometric Authentication Model based on ECG using a Fuzzy Neural Network. 4th International Conference on Advanced Engineering and Technology (4th ICAET). IOP Conf. Series: Materials Science and Engineering 317 (2018) 012030 doi:10.1088/1757-899X/317/1/012030
13. Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov PCh, Mark RG, Mietus JE, Moody GB, Peng C-K, Stanley HE. PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals. Circulation 101(23):e215-e220 [Circulation Electronic Pages; http://circ.ahajournals.org/content/101/23/e215.full]; 2000 (June 13).
14. Amit Shekhar. Understanding the Recurrent Neural Network. https://medium.com/mindorks/understanding-the-recurrent-neural-network-44d593f112a2
15. Nikolenko S. Deep learning. https://logic.pdmi.ras.ru/~sergey/teaching/mlhse17/35-dl2-backprop.pdf