ISSN 2587-814X (print), ISSN 2587-8158 (online)
Russian version: ISSN 1998–0663 (print), ISSN 2587–8166 (online)

Vol. 13 No 4 – 2019

BUSINESS INFORMATICS

HSE SCIENTIFIC JOURNAL

CONTENTS

Automation of management and production processes

Business processes modeling and analysis

Modeling of social and economic systems

Information security

Data analysis and intelligence systems

Publisher:
National Research University
Higher School of Economics

Subscription index
in the Rospechat catalog –
80870

The journal is published quarterly

The journal is included
into the list of peer reviewed
scientific editions established
by the Supreme Certification
Commission of the Russian Federation

Editor-in-Chief:
A. Golosov

Deputy Editor-in-Chief
S. Maltseva
Y. Koucheryavy

Computer Making-up:
O. Bogdanovich

Website Administration:
I. Khrustaleva

Address:
28/11, build. 4, Shablovka Street
Moscow 119049, Russia

Tel./fax: +7 (495) 772-9590 *26311
http://bijournal.hse.ru
E-mail: bijournal@hse.ru

Circulation:
English version – 300 copies,
Russian version – 300 copies,
online versions in English and Russian –
open access

Printed in HSE Printing House
3, Kochnovsky Proezd, Moscow,
Russia

ABOUT THE JOURNAL

Business Informatics is a peer reviewed interdisciplinary academic journal published since 2007 by National Research University Higher School of Economics (HSE), Moscow, Russian Federation. The journal is administered by School of Business Informatics. The journal is published quarterly.

The mission of the journal is to develop business informatics as a new field within both information technologies and management. It provides dissemination of latest technical and methodological developments, promotes new competences and provides a framework for discussion in the field of application of modern IT solutions in business, management and economics.

The journal publishes papers in the areas of, but not limited to:

✦ data analysis and intelligence systems

✦ information systems and technologies in business

✦ mathematical methods and algorithms of business informatics

✦ software engineering

✦ internet technologies

✦ business processes modeling and analysis

✦ standardization, certification, quality, innovations

✦ legal aspects of business informatics

✦ decision making and business intelligence

✦ modeling of social and economic systems

✦ information security.

The journal is included into the list of peer reviewed scientific editions established by the Supreme Certification Commission of the Russian Federation.

The journal is included into Web of Science Emerging Sources Citation Index (WoS ESCI) and Russian Science Citation Index on the Web of Science platform (RSCI).

International Standard Serial Number (ISSN): 2587-814X (in English), 1998-0663 (in Russian).

Editor-in-Chief: Dr. Alexey Golosov – President of FORS Development Center, Moscow, Russian Federation.

EDITORIAL BOARD

ABOUT THE HIGHER SCHOOL OF ECONOMICS

Consistently ranked as one of Russia's top universities, the Higher School of Economics (HSE) is a leader in Russian education and one of the preeminent economics and social sciences universities in Eastern Europe and Eurasia.
Having rapidly grown into a well-renowned research university over two decades, HSE sets itself apart with its international presence and cooperation.

Our faculty, researchers, and students represent over 50 countries, and are dedicated to maintaining the highest academic standards. Our newly adopted structural reforms support both HSE's drive to internationalize and the groundbreaking research of our faculty, researchers, and students.

Now a dynamic university with four campuses, HSE is a leader in combining Russian educational traditions with the best international teaching and research practices. HSE offers outstanding educational programs from secondary school to doctoral studies, with top departments and research centers in a number of international fields.

Since 2013, HSE has been a member of the 5-100 Russian Academic Excellence Project, a highly selective government program aimed at boosting the international competitiveness of Russian universities.

ABOUT THE SCHOOL OF BUSINESS INFORMATICS

The School of Business Informatics is one of the leading divisions of HSE's Faculty of Business and Management. The School offers students diverse courses taught by full-time HSE instructors and invited business practitioners. Students are also given the opportunity to carry out fundamental and applied projects at various academic centers and laboratories.

Within the undergraduate program, students participate each year in different case-competitions (PWC, E&Y, Deloitte, Cisco, Google, CIMA, Microsoft Imagine CUP, IBM Smarter Planet, GMC etc.) and some of them are usually as being best students by IBM, Microsoft, SAP, etc. Students also have an opportunity to participate in exchange programs with the University of Passau, the University of Munster, the University of Business and Economics in Vienna, the Seoul National University of Science and Technology, the Radbound University Nijmegen and various summer schools (Hong Kong, Israel etc.). Graduates successfully continue their studies in Russia and abroad, start their own businesses and are employed in high-skilled positions in IT companies.

There are four graduate programs provided by the School:

- Business Informatics
- E-Business;
- Information Security Management;
- Big Data Systems.

The School's activities are aimed at achieving greater integration into the global education and research community. A member of the European Research Center for Information Systems (ERCIS), the School cooperates with leading universities and research institutions around the world through academic exchange programs and participation in international educational and research projects.

DOI: 10.17323/1998-0663.2019.4.7.16

Developing digital twins for production enterprises

Valery L. Makarov
E-mail: makarov@cemi.rssi.ru

Albert R. Bakhtizin
E-mail: albert@cemi.rssi.ru

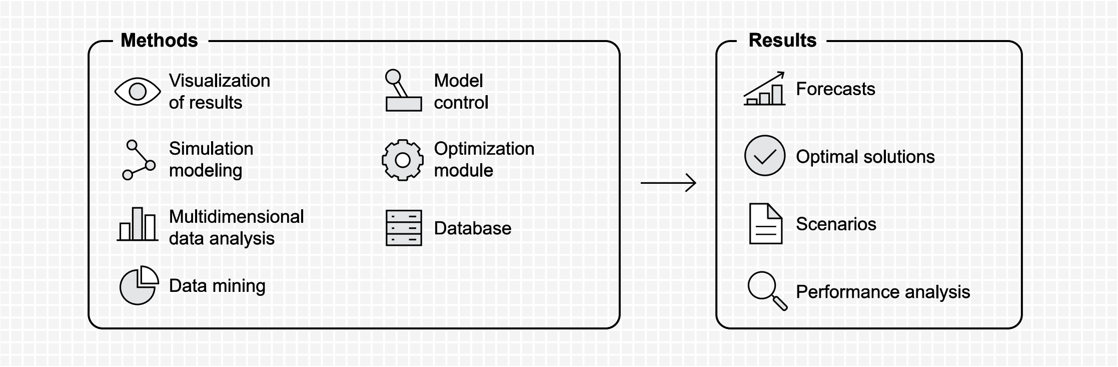
Gayane L. Beklaryan
E-mail: glbeklaryan@gmail.com

Central Economics and Mathematics Institute, Russian Academy of Sciences
Address: 47, Nakhimovsky Prospect, Moscow 117418, Russia

Abstract

This article presents a new approach to developing digital twins of production companies with the use of simulation methods. It describes the concept of digital twins as an integrated system that aggregates simulation models, databases and intelligent software modules of the class of genetic optimization algorithms, subsystems of data mining, etc. The article presents examples of simulation models of different production companies, in particular, a typical assembly plant and a typical oil production enterprise. The first company carries out activities to assembly products from individual components with its own individual characteristics. To describe the behavior of such an enterprise, methods of agent and discrete-event modeling are used. The second enterprise produces raw carbohydrate materials at existing fields with individual characteristics. The integrated simulation models thus developed are integrated with a subject-oriented database and optimization modules that facilitate providing a control of the technological and resource characteristics of the respective production enterprises. The development of these models was performed using AnyLogic and Powersim simulation systems that support agent-based modeling and system dynamics methods. We demonstrate here the possibility of creating 'digital twins' for production companies using modern simulation tools.

Graphical abstract



Key words: digital twins; agent-based modeling; system dynamics; discrete-event simulation; digital economy; production companies.

Citation: Makarov V.L., Bakhtizin A.R., Beklaryan G.L. (2019) Developing digital twins for production enterprises. *Business Informatics*, vol. 14, no 1, pp. 7–16. DOI: 10.17323/1998-0663.2019.4.7.16

Introduction

In modern times, as we transition to the digital economy, a new scientific vector is developing related to the creation of so-called "digital twins" – digital copies of real physical objects (for example, manufacturing enterprises, financial corporations, etc.) that help to optimize the effectiveness of all the main processes. The most important characteristic of digital twins is the existence of a virtual model that is supportive in an actual state, mainly due to continuous updating of data that used for the evaluation of multiple characteristics of the physical object under examination.

The concept of digital twins has been proposed recently [1, 2]. However, researchers point out that the design of digital twins should be based mainly on the use of simulation techniques providing the most realistic representation of a physical object in the virtual environment [3]. In this case, the computer model must support the ability to solve problems of optimizing the multiple characteristics of the simulated object using data updated in real time. A similar approach is applicable to many objects of life as lifeless nature. There are examples of the development of simulation models for the simulation of complex socio-economic systems [4, 5], vertically integrated petroleum companies and financial corporations [6], ecological and business systems [7, 8], modeling human crowd behavior in emergencies [9, 10] and some others. In this study, various methods of simulation modelling are used, in particular, the methods of system dynamics, agent-based and discrete-event modelling, which are supported in Powersim, AnyLogic, etc. [6].

A significant contribution to the development of system dynamics methods has been made in [11–16]. Within the research on agent-based modelling, we should highlight the works [17–20]. Important research in the field of discrete-event simulation is presented in papers [21, 22].

An important feature of modern simulation systems is the ability to integrate the developed models to databases and data warehouses (e.g. MS SQL Server, Oracle, SAP HANA), as well as the ability to integrate with external software modules, usually through a special API (application programming interface). For example, simulation models developed in AnyLogic and written in the programming Java can be integrated with the applications designed with the use of C++ and MPI (message passing interface) for parallelizing the respective computational procedures. As a result, the simulation model can be aggregated with genetic optimizing algorithms through objective functions, providing the ability to optimize characteristics of the simulated object in real time [23–25]. To ensure the software and management pool of available digital twins, it is possible to use an approach based on a service-oriented architecture (SOA).

The purpose of this article is to develop the approach to designing digital twins for a variety of production companies using simulation methods, databases, optimization modules, etc. for the implementation of the rational production management concept in real time and to support strategic and operational decision-making systems.

1. The concept of creating digital twins

The suggested concept of creating digital twins is based on the use of an integrated approach combining simulation methods, optimization modules (of a type of genetic algorithms), a database (data warehouse), a multidimensional data analysis subsystem (online analytical processing, OLAP) , etc. (*Figure 1*).

An important feature of the approach being proposed is to ensure continuous integration of the considered subsystems with support for the functioning of all software modules in real time.

Thus, the following interacting subsystems are suggested:

✦ **simulation modelling subsystem** designed to compute the values of multiple characteristics of the enterprise under given scenario conditions;

✦ **optimization module (genetic algorithm)** that is aggregated through the objective functions with simulation models of the production company and provides the possibility to seek the best (rational, suboptimal) decisions with the existing restrictions;

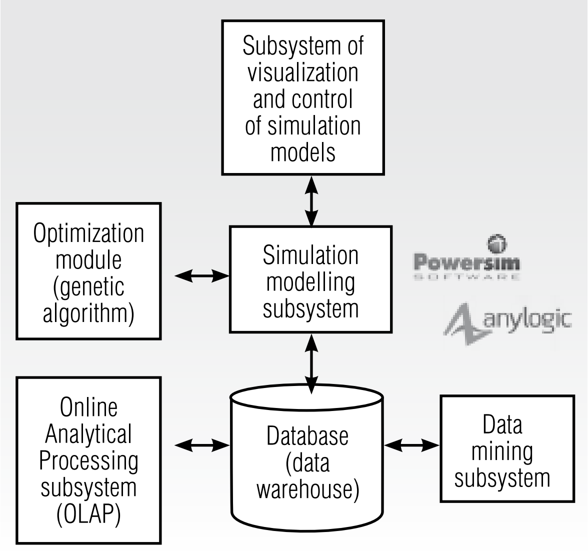

Subsystem of visualization and control of simulation models
Optimization module (genetic algorithm)
Simulation modelling subsystem
Powersim SOFTWARE
anylogic
Online Analytical Processing subsystem (OLAP)
Database (data warehouse)
Data mining subsystem

Fig. 1. The integrated system architecture of a digital twin of the manufacturing enterprise

✦ **database (data warehouse)** provides collection and processing of relevant data across the enterprise, as well as initial data for simulation models and saves the results of simulation modelling;

✦ **subsystem of visualization and control of simulation models** integrated with enterprise simulation models (through a special API or using web services technologies – SOA) and allowing access to various functions of simulation models, for example, calling optimization experiments, saving the results of scenario modeling in the database and doing other activities;

✦ **Online Analytical Processing subsystem (OLAP)**, allowing us to analyze the results of simulation and optimization with the drill-down of the corresponding aggregation data (e.g., for the enterprise as a whole, for individual business-directions, by products, customers, etc.);

✦ **data mining subsystem** that provides an analysis of the relationships between the most important characteristics of an information model of production and updates the values of all influencing factors and significance coefficients, followed by saving the results to the system database. Such an approach provides continuous updating of existing interdependencies in the enterprise simulation models.

Next we will consider some examples of simulation models developed for enterprises of various classes, in particular, a typical assembly plant and a typical oil production enterprise.

2. Simulation model of assembly plant

Consider the aggregated simulation model of a typical assembly plant implemented in the AnyLogic system (*Figure 2*).

This assembly plant produces some products using two production lines. The first production line provides the formation of an intermediate product using five separate components supplied in accordance with a given schedule to the assembly system (the element of “Assembly 1”). The second production line provides the formation of the final product (the element of “Assembly 2”) using two components, one of which is assembled using the output of the first production line.

In this model, each production component is an agent with its own individual characteristics. For example, if this is a production skeleton, then it has certain sizes and other specified technical characteristics. Based on the variation of the set of inflow agent-components, it ensures manufacture of various types of products at the output of the system (the element of “Exit”). At the same time, for components of individual types with their own individual characteristics, dynamic synchronization is performed using a special element that ensures the placement of agent-components in the queue and seeks the pairs of agents that satisfy a given correspondence criterion (the element of “Synchronizer”). As a result, agent-components corresponding to some specified criteria are being sent to the assembly, while other agent-components remain in the queue, or are forced out of it (for safekeeping).

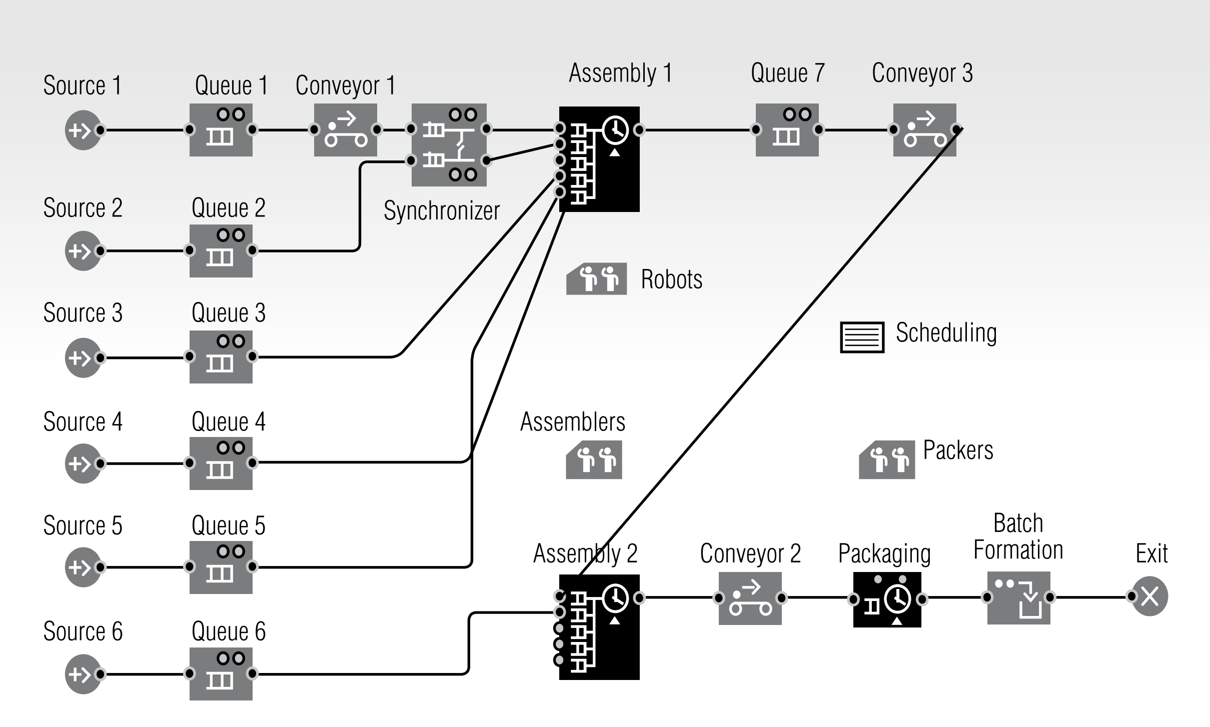


Fig. 2. Simulation model of an assembly plant in AnyLogic

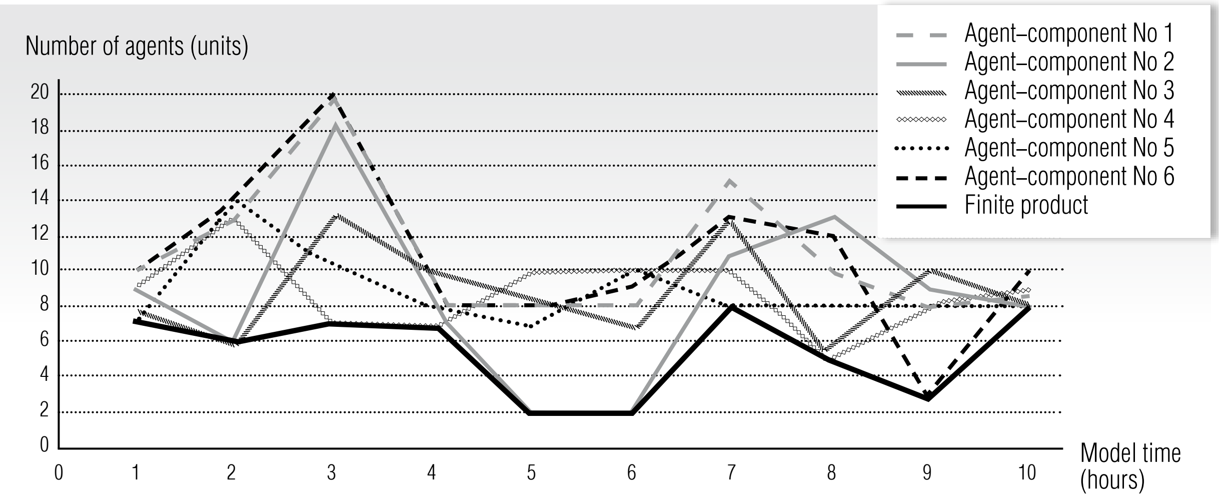


Fig. 3. The dependence of the total number of assembled products on the dynamics of the supply of agent–components

One of the possible examples of analysis of the stability of the production process using the simulation model so developed is the dependence of the total number of assembled products on the dynamics of the supply of agent-components (*Figure 3*).

In *Figure 3* we see that in conditions of the deficit of agent-components No 2 that occurs in the time interval from 5 to 6 hours, the rate of assembly of the finished product is reduced to the minimum level (two products per hour) equaling the value of the supply rate of the most deficit (the second) component.

Thus, the assembly rate of the final product is limited by the rate of supply of the most deficit component needed at the corresponding assembly stage. The rate of assembly and delivery of the final product also significantly depends on the resources involved in the corresponding production processes (elements of "Robots", "Assemblers" and "Packers" in *Figure 2*). In the case of the deficit of resources necessary for the production of a unit of production, the time to complete the corresponding task (product assembly) will be increased in proportion to the value of the availability coefficient of the corresponding resource.

3. Simulation model of an oil production company

Further, we consider the aggregated simulation model of a typical oil production company implemented in the Powersim system (*Figure 4*).

In contrast to the previous discrete-event and agent-based simulation of the assembly plant, this model operates with continuous raw and financial flows and therefore it was developed with the use of system dynamics methods [6]. At the same time, the main assets of the production enterprise, in particular, the set of new wells and the set of old wells, differentiated by fields, are important resource characteristics. The transition of wells from a new set to the old set takes place after a certain time interval (usually five years) with the associated change in a production volume, which is reflected in the production function, taking into account the different contributions of new and old wells in operation, respectively.

On the other hand, the indicators of economic efficiency of the exploited fields and wells depend not only on the volume of extracted raw materials, but also on the operating and investment expenditures differentiated by the respective fields. If a certain field is not oper-

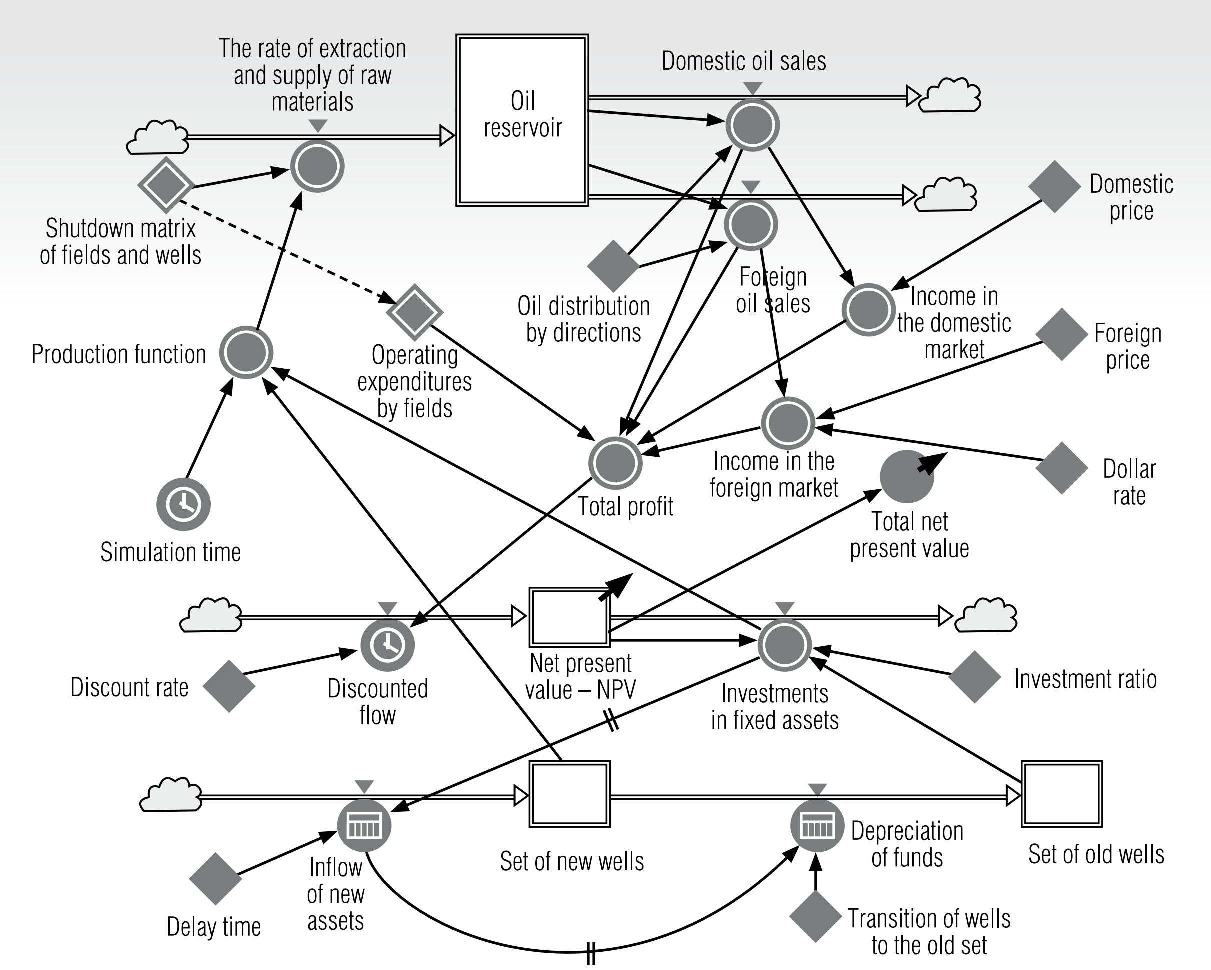


Fig. 4. Simulation model of an oil production company

ated at a certain time (for example, due to the suspension of the respective wells), then material and financial flows, as well as income and expenditure characteristics, cease to be formed on it. Further on, the most important relations of the proposed model will be described.

Here,

✦ $t = 1, 2, ..., T$ is the simulation time (by years), T is the horizon of a strategical planning;

✦ $i = 1, 2, ..., I$ is the indexes of fields, I is the total number of fields;

✦ $a_i(t)$ is the number of new wells (fixed assets of an oil production company) of the i-th field ($i = 1, 2, ..., I$) at the moment $t\,(t = 1, 2, ..., T)$;

✦ $\{b_i(t),\ b_i(1)\}$ is the number of old wells of the i-th field ($i = 1, 2, ..., I$) at the moment t ($t = 1, 2, ..., T$) and the initial number of old wells at the initial moment $t = 1$;

✦ $\chi_i(t) \in \{0, 1\}$ is the shutdown matrix the i-th fields ($i = 1, 2, ..., I$) at the moment $t\,(t = 1, 2, ..., T)$: if $\chi_i(t) = 0$ the i-th field is not operated, if $\chi_i(t) = 1$ the i-the field is operated;

✦ $c_i(t)$ is the rate of production and supply of raw materials (ton per year) of the i-th field ($i = 1, 2, ..., I$) at the moment $t\,(t = 1, 2, ..., T)$;

✦ μ is the coefficient of decreasing of old fields;

✦ $\{v(t),\ \tilde{v}(t)\}$ are average annual production volume of new and old wells, respectively at the moment$t$ $(t = 1, 2, ..., T)$;

✦ $\{p_1(t), p_2(1)\}$ are prices of raw materials supplied in domestic and foreign markets, respectively at the moment t $(t = 1, 2, ..., T)$;

✦ $\{P_{1i}(t),\ P_{2i}(t)\}$ are incomes by i-ths fields $(i = 1, 2, ..., I)$ in domestic and foreign markets, respectively at the moment$t$ $(t = 1, 2, ..., T)$;

✦ $\{\tilde{O}_i(t),\ \tilde{I}_i(t)\}$ are operation and investment expenditures by i-ths fields $(i = 1, 2, ..., I)$, respectively at the moment t $(t = 1, 2, ..., T)$;

✦ $s(t)$ is the dollar rate at the moment t $(t = 1, 2, ..., T)$;

✦ $\lambda(t)$ is s share of supply in a domestic market at the moment t $(t = 1, 2, ..., T)$, $0 \le \lambda(t) \le 1$;

✦ τ is the time period, during of which wells can be classified as news (as a rule, five years);

✦ r is discounted rate.

The number of new and old wells respectively:

$$a_i(t) = \sum_{t=1}^{T} \left(\tilde{a}_i(t) - \tilde{a}_i(t-\tau)\right), \qquad (1)$$

$$b_i(t) = b_i(1) + \sum_{t=1}^{T} \tilde{a}_i(t-\tau), \qquad (2)$$

$$i = 1, 2, ..., I;\ t = 1, 2, ..., T.$$

The rate of production (and supply) of raw materials:

$$c_i(t) = \chi_i(t)\left(a_i(t)v(t) + b_i(t)\tilde{v}(t)e^{-\mu \cdot t}\right), \qquad (3)$$

$$i = 1, 2, ..., I;\ t = 1, 2, ..., T.$$

Revenue in domestic and foreign markets respectively:

$$P_{1i}(t) = p_1(t)c_i(t)\lambda(t), \qquad (4)$$

$$P_{2i}(t) = p_2(t)s(t)c_i(t)\left(1 - \lambda(t)\right), \qquad (5)$$

$$i = 1, 2, ..., I;\ t = 1, 2, ..., T.$$

The profit of the oil production company:

$$\pi_i(t) = P_{1i}(t) + P_{2i}(t) - \chi_i(t)\tilde{O}_i(t), \qquad (6)$$

$$i = 1, 2, ..., I;\ t = 1, 2, ..., T.$$

The discounted cash flow of an oil production enterprise:

$$DCF_i(t) = \frac{\pi_i(t) - \chi_i(t)\tilde{I}_i(t)}{(1+t)^r}, \qquad (7)$$

The net accumulated discounted financial flow of the producing enterprise and the total financial flow (net present value) summarized by all fields, respectively:

$$NPV_i = \sum_{t=1}^{T} DCF_i(t), \qquad (8)$$

$$NPV^* = \sum_{i=1}^{I} NPV_i. \qquad (9)$$

Note that to predict the dynamics of the average daily production (production) rate for new and old wells $\{v(t),\ \tilde{v}(t)\}$, which affects the rate of production and supply of raw materials (3), data mining methods are currently being applied, in particular, artificial neural networks (ANN). Such ANN estimates multiple input characteristics (for example, the level of reserves, data on actually carried out and planned geological and technical measures, the selected production technology, etc.) to predict the average daily production rate of wells and annual volume. Thus, the values of the most important basic characteristics of the suggested simulation model are updated in real time taking into account the physical characteristics of the exploited fields.

One possible example of analysis of the effectiveness of a portfolio of investment projects for the set of exploited fields is the dependence of the discounted financial flow $DCF_i(t)$ on the values of the elements of the matrix of "shutdowns" of fields $\chi_i(t) \in \{0, 1\}$ $(i = 1, 2, ..., I)$ (*Figure 5*).

Figure 5 shows that the "shutdown" of the second field leads to doubling the total financial flow (summarized by all fields) — *NPV**. Such a positive effect results from the fact that the level of operating expenditure for the second field (per unit of production) significantly exceeds the level of costs of other fields with a comparable level of production volume. Nevertheless, if there is a hard restriction on the minimum needed total volume of production and supply of raw materials, then the exclusion of the second field from operation is impossible.

At the same time, if the number of simultaneously evaluated raw materials assets is large (for example, several thousand), then genetic optimization algorithms [23–25], aggregated with the simulation model of a production company, should be used to identify and "turn off" such fields.

Conclusion

The article presents a new approach to the development of digital twins for manufacturing enterprises which is based on the concept of continuous integration of a number of key subsystems, in particular, a simulation modeling subsystem, optimization module, database, data mining subsystem, etc. An important feature of digital twins is updating source data and values of influencing parameters in real time. Thus, digital twins significantly expand the functionality of traditional enterprise simulation models, mainly due to the greater realism and interactivity of the corresponding technology.

Examples of digital twins developed for manufacturing enterprises of different types are presented. In particular, these are the typical assembly plant and an oil production company. The first model, in particular, allows us to study the dependence of the total number of assembled products on the dynamics of the supply of components, taking into account their individual characteristics. The second model allows us to evaluate the impact of the matrix of "shutdowns" of oil fields on the dynamics of the discounted financial flow. It is shown that in the case of an absence of hard restrictions on the minimum required total volume of the raw material production (the production plan),

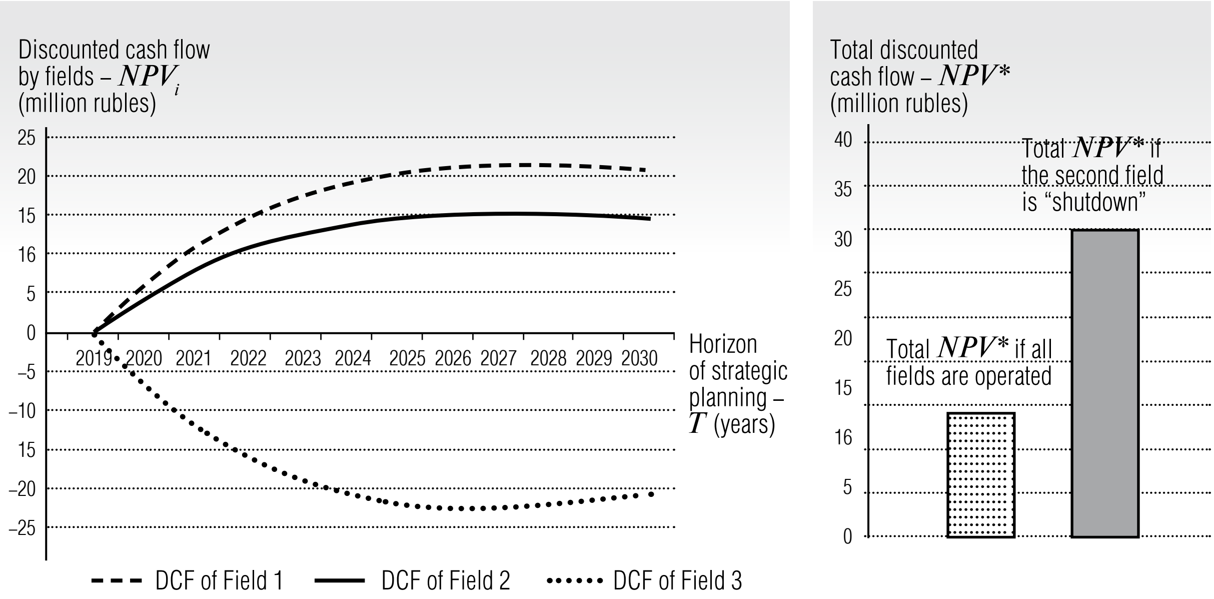


Fig. 5. The dependence of the discounted financial flow on the values of the elements of the matrix of "shutdowns" of fields

a significant increase in the net present value of the portfolio of fields is possible due to the exclusion of low-debit wells and fields with a relatively high level of operating expenditures.

If it is necessary to study the influence of multiple control parameters (for example, thousands of fields and wells, hundreds of components, dozens of production lines, etc. are simulated), then it is possible to use special optimization modules aggregated with simulation models through objective functions. Such an integrated approach is the most promising in terms of the development of the concept of digital twins and can be used in the future at a more detailed level. ■

Acknowledgments

The study was funded by the Russian Foundation for Basic Research (RFBR) under the research project No 18-29-03139.

References

1. Saddik A.E. (2018) Digital twins: the convergence of multimedia technologies. *IEEE MultiMedia*, vol. 25, no 2, pp. 87–92.
2. Yan X., Yanming S., Xiaolong L., Yonghua Z. (2019) A digital-twin-assisted fault diagnosis using deep transfer learning. *IEEE Access*, vol. 7, pp. 19990–19999. DOI: 10.1109/ACCESS.2018.2890566.
3. Talkhestani B.A., Jung T., Lindemann B., Sahlab N., Jazdi N., Schloegl W., Weyrich M. (2019) An architecture of an intelligent digital twin in a cyber-physical production system. *Automatisierungstechnik*, vol. 67, no 9, pp. 762–782. DOI: 10.1515/auto-2019-0039.
4. Makarov V.L., Bakhtizin A.R., Beklaryan G.L., Akopov A.S. (2019) Development of a software platform for large-scale agent-based modeling of complex social systems. *Program Engineering*, vol. 10, no 4, pp. 167–177 (in Russian). DOI: 10.17587/prin.10.167-177.
5. Makarov V.L., Bakhtizin A.R., Beklaryan G.L., Akopov A.S., Rovenskaya E.A., Strelkovsky N.V. (2019) Aggregated agent-based simulation model of migration flows of the European Union countries. *Economics and Mathematical Methods*, vol. 55, no 1, pp. 3–15 (in Russian). DOI: 10.31857/S042473880004044-7.
6. Akopov A.S. (2017) *Simulation modeling.* Moscow: Urait (in Russian).
7. Akopov A.S., Beklaryan L.A., Saghatelyan A.K. (2019) Agent-based modelling of interactions between air pollutants and greenery using a case study of Yerevan, Armenia. *Environmental Modelling and Software*, vol. 116, pp. 7–25. DOI: 10.1016/j.envsoft.2019.02.003.
8. Akopov A.S., Beklaryan L.A., Saghatelyan A.K. (2017) Agent-based modelling for ecological economics: A case study of the Republic of Armenia. *Ecological Modelling*, vol. 346, pp. 99–118. DOI: 10.1016/j.ecolmodel.2016.11.012.
9. Akopov A.S., Beklaryan L.A. (2015) An agent model of crowd behavior in emergencies. *Automation and Remote Control*, vol. 76, no 10, pp. 1817–1827. DOI: 10.1134/S0005117915100094.
10. Beklaryan A.L., Akopov A.S. (2016) Simulation of agent-rescuer behaviour in emergency based on modified fuzzy clustering. Proceedings of the *2016 International Conference on Autonomous Agents & Multiagent Systems (AAMAS 2016), Singapore, 9–13 May 2016*, pp. 1275–1276.
11. Forrester J. (1969) *Urban dynamics.* Waltham, MT, USA: Pegasus Communications.
12. Forrester J. (1959) Industrial dynamics: A major breakthrough for decision makers. *Harvard Business Review*, vol. 36, no 4, pp. 37–66.
13. Meadows D.H., Randers J. Meadows D.L. (2005) *Limits to growth: The 30-year update.* London: Earthscan.
14. Meadows D.H. (1972) *Limits to growth: A report for the Club of Rome's project on the predicament of mankind.* New York: Universe Books.
15. Sidorenko V.N. (1998) *System dynamics.* Moscow: MSU, TEIS (in Russian).
16. Akopov A.S., Khachatryan N.K. (2014) *System dynamics.* Moscow: CEMI RAS (in Russian).
17. Schelling T.C. (1971) Dynamic models of segregation. *Journal of Mathematical Sociology*, vol. 1, no 2, pp. 143–186.

18. Axelrod R. (1997) *The complexity of cooperation: Agent-based models of competition and collaboration*. Princeton: Princeton University Press.

19. Bakhtizin A.R. (2008) *Agent-based models of economy*. Moscow: Economics (in Russian).

20. Akopov A.S., Khachatryan N.K. (2016) *Agent-based modeling.* Moscow: CEMI RAS (in Russian).

21. MacDougall M.H. (1987) *Simulating computer systems: Techniques and tools.* MIT Press.

22. Delaney W., Vaccari E. (1989) *Dynamic models and discrete event simulation*. New York: Marcel Dekker.

23. Akopov A.S., Beklaryan L.A., Thakur M., Verma B.D. (2019) Parallel multi-agent real-coded genetic algorithm for large-scale black-box single-objective optimization. *Knowledge-Based Systems*, vol. 174, pp. 103–122. DOI: 10.1016/j.knosys.2019.03.003.

24. Akopov A.S., Beklaryan A.L., Thakur M., Verma B.D. (2019) Developing parallel real-coded genetic algorithms for decision-making systems of socio-ecological and economic planning. *Business Informatics*, vol. 13, no 1, pp. 33–44. DOI: 10.17323/1998-0663.2019.1.33.44.

25. Beklaryan G.L., Akopov A.S., Khachatryan N.K. (2019) Optimisation of system dynamics models using a real-coded genetic algorithm with fuzzy control. *Cybernetics and Information Technologies*, vol. 19, no 2, pp. 87–103. DOI: 10.2478/cait-2019-0017.

About the authors

Valery L. Makarov

Dr. Sci. (Phys.-Math.); Academician of Russian Academy of Sciences;

Academic Supervisor, Central Economics and Mathematics Institute, Russian Academy of Sciences,
47, Nakhimovsky Prospect, Moscow 117418, Russia;

E-mail: makarov@cemi.rssi.ru

ORCID: 0000-0002-2802-2100

Albert R. Bakhtizin

Dr. Sci. (Econ.); Corresponding Member of Russian Academy of Sciences;

Director, Central Economics and Mathematics Institute, Russian Academy of Sciences,
47, Nakhimovsky Prospect, Moscow 117418, Russia;

E-mail: albert@cemi.rssi.ru

ORCID: 0000-0002-9649-0168

Gayane L. Beklaryan

Cand. Sci. (Econ.);

Senior Researcher, Laboratory of Computer Modeling of Social and Economic Processes;
Central Economics and Mathematics Institute, Russian Academy of Sciences,
47, Nakhimovsky Prospect, Moscow 117418, Russia;

E-mail: glbeklaryan@gmail.com

ORCID: 0000-0002-1286-0345

DOI: 10.17323/1998-0663.2019.4.17.27

Improvement of a pharmaceutical enterprise's business processes at the stage of preclinical development of new drugs

Mikhail V. Belov[a]
E-mail: mbelov@ibs.ru

Mikhail A. Shakhmuradyan[b]
E-mail: mshakhmuradyan@econ.msu.ru

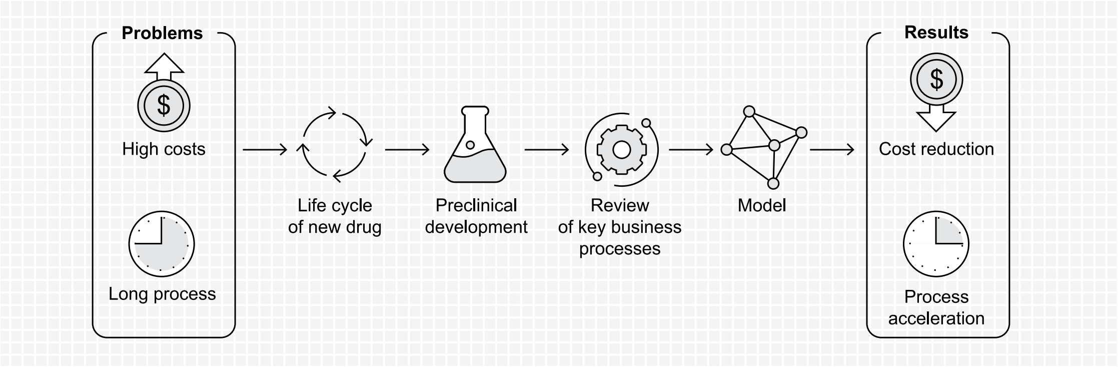
[a] IBS Group Holding Ltd.
Address: 3 build. 1, Skladochnaya Street, Moscow 127018, Russia

[b] Lomonosov Moscow State University
Address: 1 build. 46, GSP-1, Leninskie Gory, Moscow 119991, Russia

Abstract

The rapid increase of the population, as well as the high rate of urbanization, are leading to an increased need for medical supplies, as well as the need for faster release of new drugs to the market. In this article, the authors analyze and aggregate the activities of a number of pharmaceutical companies at the preclinical stage of drug development to identify optimal management models at the stage of each operational iteration. The work methodology is based on an empirical study based on two progressive stages: a qualitative analysis of the business processes of pharmaceutical enterprises, reflected in regulatory rules and company reports, as well as in-depth interviews with representatives of these commercial organizations. Based on the results obtained at the end of the first stage, the authors established the problem of the lack of a unified presentation of standardized operating procedures, as well as an aggregated representation of the stage of research work on the production of the drug. When considering each business process of this stage, the authors presented a model of the minimum unit of integrated activity (MUnIA). This model most optimally describes local operational business processes during drug development, and can also serve as an instrumental framework for guiding preclinical studies in the formation of a document on standardization of operational procedures. The results of the first stage of the empirical analysis were verified during the second part of the work — in-depth interviews with industry representatives. The findings of this study can be used by project managers at the preclinical testing stage to reduce the time spent on operating procedures.

Graphical abstract



Key words: business process; pharmaceuticals; pharmaceutical company; digitalization; project management.

Citation: Belov M.V., Shakhmuradyan M.A. (2019) Improvement of a pharmaceutical enterprise's business processes at the stage of preclinical development of new drugs. *Business Informatics*, vol. 13, no 4, pp. 17–27. DOI: 10.17323/1998-0663.2019.4.17.27

Introduction

Today the international pharmaceutical industry is undergoing significant changes. Due to the increase in the number of able-bodied people, the increase in general financial well-being, and the strengthening of the health policy of developed countries, the demand for pharmacological preparations is steadily increasing, while the costs of medicines are only going up [1]. The potential decline in the financial results of a modern pharmaceutical enterprise is explained not only by a high investment threshold for entering this market, but also by general competition, which requires the constant implementation of comprehensive innovative management strategies to maximize the product's life value by improving the current operating activities of the enterprise. Today in the world market, it is customary to single out several types of pharmaceutical organizations. The main types of pharmaceutical companies, classified by main type of activity [2], are presented in *Table 1.*

We will consider the business processes of the first type of pharmaceutical enterprise, including a full cycle of creation, testing and implementation of the product. The aim of the study is to analyze the business processes of the pharmaceutical company at the preclinical research stage, with the subsequent proposal of an informed solution in terms of their improvement.

The article has the following structure. First of all, theoretical provisions are considered by describing the full life cycle of drug development and production. What follows are the results of the analysis of business processes at the stage of preclinical trials of the drug, with the identification of a potential area of operational process management. Finally, we formulate the results of summarizing the systemic processes of pharmaceutical company research work and we present the model verified through interviews with managers of pharmaceutical companies.

Table 1.

The main types of pharmaceutical companies classified by main type of activity

No	Type of pharmaceutical company	Examples
1.	Full circle companies whose activities include partici pation in the entire product life cycle from the development and research of the molecule to the marketing	HTC «ChemRar», Roche Holding, Novartis International AG, Pfizer Inc., GlaxoSmithKline plc, Sanofi, etc.
2.	R&D enterprises	Research institute «ChemRar», Research institute of Chemical Diversity, Insilico Medicine Inc., etc.
3.	Manufacturing companies	«ChemRar» Research institute of Chemical Diversity, BioIntegrator, Sanofi–Aventis East, etc.
4.	Distribution companies	Roche Moscow, Bioritm, Dileo–pharma, PHARMA Distribution Group, etc.

1. The life cycle of drug development and production

To consider the main business processes of a modern full-cycle pharmaceutical enterprise, it is necessary to determine the subject of activity for the production of a medicinal product. In this study, the subject of activity is drug development – the process by which potential medicines are discovered or developed. The process of discovery drugs today has an established theoretical and methodological base and includes two main phases – preclinical research (on average, 3 years) and clinical trials (from 6 years or more) [3–4].

Since the demand for new methods of pharmacological treatment remains high, and the time required to bring new drugs to the market is about 10 years, pharmaceutical enterprises are looking for different ways to shorten the life cycle of drugs at different stages. On average, to develop and obtain approval for a new drug, a full-cycle commercial enterprise needs about $ 2.6 billion [5]. If we consider the financial aspects (*Figure 1*), it can be noted that business processes for the production of drugs at

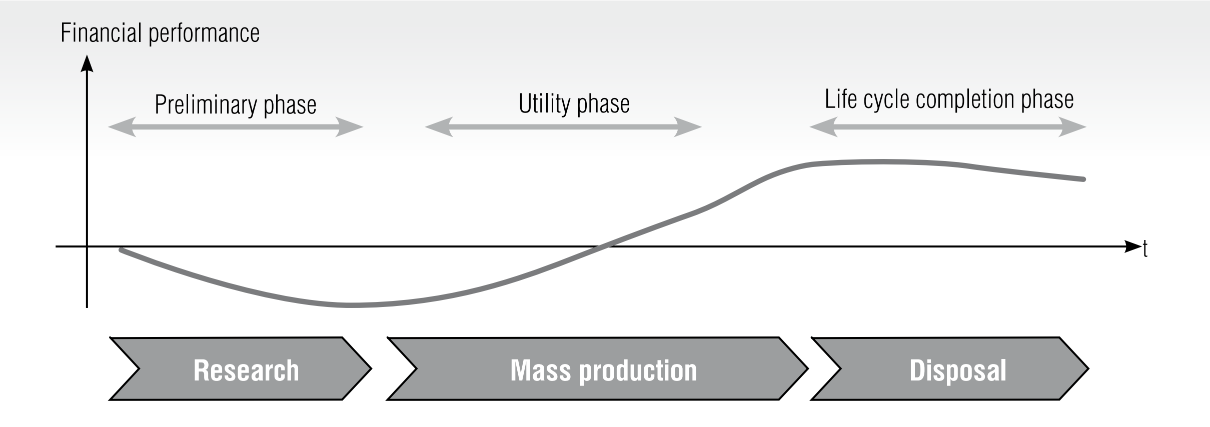


Fig. 1. Financial performance of the drug life cycle stages

the stage of research in the preliminary phase, together with the average cost of preclinical studies, characterize the period most unfavorable for the investor. In order to improve operational activities and shorten the life cycle of drug development, the authors conducted a comprehensive analysis of the business processes of pharmaceutical enterprises reflected in regulatory documents and reports of a number of international pharmaceutical companies.

In this work, the business process is defined as a set of tasks that includes systems and methods applied to create and develop the final product or service provided to the buyer [6]. Accordingly, for the possible improvement of business processes, it is necessary to decompose the entire production life cycle of the drug and determine processes that could be systematized to reduce time.

The life cycle of the production and development of drugs in the early stages involves a marketing analysis, which is carried out by the relevant division of the company. During such analysis, current and future demand is taken into account, and a proposal is formulated to create a specific drug or treat a specific disease. Before making a decision, the company's management can analyze the current production and research capacities to assess the possibility of working on the projects proposed in the plan.

At the time of approval, the ultimate goal of drug development, research leaders who organize and control the preclinical stages (target search, agent search, prototype optimization, drug candidate testing) and clinical trials (Figure 2) are appointed. The clinical trials of a drug take the largest part of the drug development time. At the same time, due to regulatory reasons, at the level of the company it is not possible to accelerate this stage.

The process of movement from preclinical trials to clinical trials can be described by the next steps. A pharmaceutical company is applying for a clinical trial authorization (CTA), which is being reviewed by a number of experts. After that, a decision is made about the possibility of conducting tests on humans.

At phase 1 of the trial, the safety and pharmacology of the candidate drug are tested on a small group of healthy volunteers (from 20 to 100 people) who are given small doses of the compound. Sometimes phase 0 or "proof-of-concept" (PoC) is performed, when a candidate for drugs is tested in a small group of patients (from 5 to 15) to determine the "mechanism of action" of the drug in the human body.

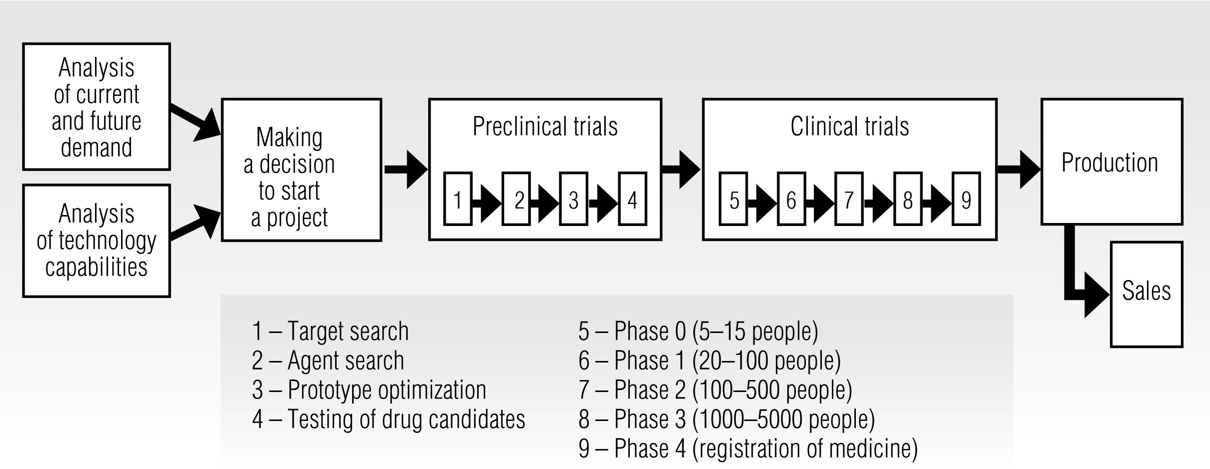

Analysis of current and future demand
Analysis of technology capabilities
Making a decision to start a project
Preclinical trials
1
2
3
4
Clinical trials
5
6
7
8
9
Production
Sales
1 – Target search
2 – Agent search
3 – Prototype optimization
4 – Testing of drug candidates
5 – Phase 0 (5–15 people)
6 – Phase 1 (20–100 people)
7 – Phase 2 (100–500 people)
8 – Phase 3 (1000–5000 people)
9 – Phase 4 (registration of medicine)

Fig. 2. Main stages of the drug life cycle of a full cycle pharmaceutical company

At phase 2, the effectiveness of the compound is studied on a group of volunteers who have a condition for the treatment of which the drug is directed. Typically, such a group includes from 100 to 500 patients whose vital indicators are constantly monitored and evaluated. The purpose of phase 2 is to determine the most effective dose and method of drug delivery (for example, orally or intravenously). Most drugs that fail during clinical trials due to their inefficiency and insecurity are detected at this phase.

At phase 3, potential drugs are tested in a significantly larger population (from 1000 to 5000 people), in several international centers. At the same time, the pharmaceutical company needs to collect sufficient data on the safety and efficacy of drugs in order to apply for licensing to the regulatory authority (for example, in the UK it is The Medicines and Healthcare Products Regulatory Agency and in the USA is The Food and Drug Administration).

At phase 4, the drug is finally approved by the national supervisor and production is launched with the simultaneous determination of distribution channels.

The listed stages are normatively defined and there are mandatory procedures which to identify the safety of the drug. In modern conditions, the pharmaceutical industry is forced to transform and improve precisely the stage of preclinical development in order to compensate for the longer wait for human trials and approval of regulatory authorities. Therefore, given the regulatory requirements of the relevant regulatory authorities, it seems important to consider the process of drug development precisely at the stage of research work.

2. The business processes at the preclinical stage of drug development

A significant part of the work on the identification and development of drugs in the early stages is carried out by universities and research institutes, the activities of which are supervised and controlled by project managers of full circle companies [7]. For example, in a university laboratory, scientists with the support of grants from research institutions or of a pharmaceutical company conduct basic research to determine the cause of the disease. The interaction of several entities in the development of drugs is an integral part of the business processes of a modern international pharmaceutical company. Under these conditions, not only the technological or qualification level of the performers, but also the coordination and regulation of their interaction have a significant impact on the progress of research.

The activity of project managers at the preclinical stage is determined not only by corporate standards, but also by external standards [8]. In particular, an internal preclinical testing processes are governed by such standards as the “good laboratory practice” (GLP) system created in the USA, Russian GOST 33647-2015, as well as GAMP (good automated manufacturing practice) and ISO / IEC 17025: 2005 (in the Russian Federation – GOST ISO / IEC 17025-2009) [9]. Based on the “principles of good laboratory practice,” the planning and implementation of key stages of research work is carried out, and the reporting form for each iteration is also regulated. The main purpose of these activities is to study the safety of a chemical. Each hierarchical level is regulated by a number of norms; however, there is an area of operational procedures, the development of which is under the supervision of research leaders. So, in accordance with GOST 33647-2015, each unit of the testing center should have its own standard operating procedures (SOP) for the type of activity regulated by the administration; however, there is no unified presentation of the document in the form of a normative act [10]. In addition, GOST ISO / IEC 17025-2009 obliges the administrative unit organizing or hiring outside research to develop and provide a management system in accordance with its field of activity. This management system should be documented in the form of a detailed description of the system, program, procedures and instructions to the extent necessary for a qualitative reflection of the research results [9]. In addition, the project manager should monitor not only the organization of security of operational processes, but also the documentation of the security of the compound – the drug candidate. In accordance with ICH Guideline M3 (R2), approved by the International Council for the Harmonization, preclinical safety studies should be sufficient to characterize potential side effects that may occur in a supported clinical trial.

Based on the foregoing, the authors concluded that there is a large load on project managers at the preclinical stage of drug development. Nevertheless, one of the most important aspects determining the increase in time for operating procedures may be the lack of a standardized SOP [11–12]. Based on this, the following is considered the problem of finding an improved way to organize operational activities in the framework of preclinical trials.

Based on a qualitative study of regulatory documents (in particular, GOST 33647-2015 and ISO / IEC 17025: 2005), as well as business processes of a number of pharmaceutical companies, the authors described a number of operational procedures that take place at the stage of drug discovery and development (*Figure 3*). It should be noted that modern research is entirely conducted at the molecular level. The stage of target search is a priority for preclinical studies. The target can be a gene, protein, or protein-protein interactions (PPIs), either contributing to the disease, or can interfere with treatment, for example, by blocking the desired receptor.

Since some diseases are associated with dysfunction of not only one molecule, but, for example, 5–10 bound proteins, the scientific team needs to study at least five times more targets. Once a potential target has been identified, the researchers will move on to finding a compound that will act on this goal and, therefore, can affect the disease. Some 10 000 or more compounds can be considered. These are usually reduced to 10–20, which, according to the forecasts of the working group, can affect the disease. Currently, the process of finding a new drug against a selected target for a specific disease usually includes molecular docking (molecular modeling to find the optimal archi-

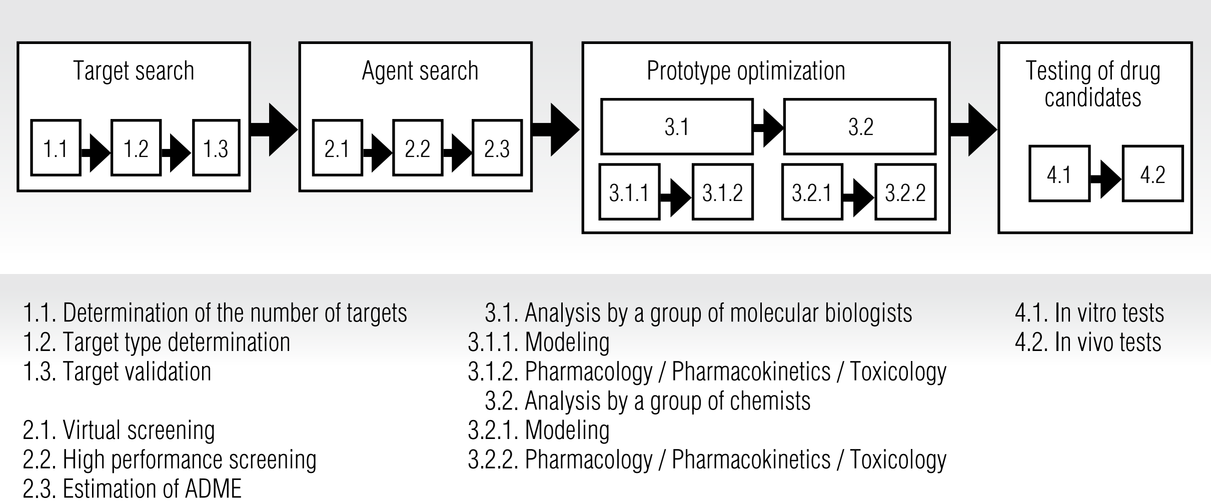

Target search
1.1
1.2
1.3
Agent search
2.1
2.2
2.3
Prototype optimization
3.1
3.2
3.1.1
3.1.2
3.2.1
3.2.2
Testing of drug candidates
4.1
4.2
1.1. Determination of the number of targets
1.2. Target type determination
1.3. Target validation
2.1. Virtual screening
2.2. High performance screening
2.3. Estimation of ADME
3.1. Analysis by a group of molecular biologists
3.1.1. Modeling
3.1.2. Pharmacology / Pharmacokinetics / Toxicology
3.2. Analysis by a group of chemists
3.2.1. Modeling
3.2.2. Pharmacology / Pharmacokinetics / Toxicology
4.1. In vitro tests
4.2. In vivo tests

Fig. 3. Aggregated scheme of the research phase

tecture in the target protein binding center), virtual screening (automated selection of the necessary chemical compounds from specialized databases), as well as high-throughput screening (high-throughput screening), in which large combinatorial libraries of chemicals are tested for their ability to inhibit the target.

After determining the relevant compounds, their pharmacokinetic verification (ADME) is performed. The next step is the optimization of prototypes, which is carried out in parallel by a group of molecular biologists, as well as a team of specialists in the field of medical chemistry. The activities of these two groups are also supervised by the project manager. After successful selection and verification of a number of compounds, in vitro and in vivo experiments are carried out. In order to gain access to the clinical phase, the project team needs to carry out the following studies: pharmacological (primary pharmacodynamics, secondary pharmacodynamics, pharmacological safety, pharmacodynamic drug interactions) and pharmacokinetic (absorption, distribution, metabolism, excretion and toxicological properties of the drug, pharmacokinetic drug interaction).

During the study of general toxic properties (assessment of acute and chronic toxicity), on the one hand, toxic doses that can cause the death of animals are revealed, and thereby determine the dose limit that cannot be exceeded. On the other hand, the safety of therapeutic doses with prolonged exposure to animals is assessed. For this, the drug is administered daily both in therapeutic doses and in doses ten times higher than therapeutic. The duration of the experiment depends on the duration of the intended course in clinical practice. During the experiment, the functional state of all body systems, the weight and behavior of animals are analyzed. After completing the course, a histological examination of all organs and tissues is performed. Also, in the framework of a safety assessment at the preclinical stage, experiments are conducted on the so-called specific toxicity. In general, these studies make it possible to assess the possible effect of the studied drug on the immune status of the body, the reproductive system and offspring, the genetic apparatus, as well as the possibility of allergic reactions and provoking malignant tumors.

However, these studies cannot provide reliable information on the effect of the studied drugs on humans, since the organism of laboratory animals differs from the human one both in pharmacokinetic characteristics and in the response of organs and systems to drugs. That is why it is necessary to conduct clinical trials of drugs. Thus, the main result of preclinical studies of a new drug is a prediction of its safety for humans.

3. The model of the minimum unit of integrated activity (MUnIA)

The aspects we have considered explain the importance of a detailed analysis of processes at all stages of research work. As already noted, industry actors are looking for more effective approaches to launching new products on the market that can accelerate product development while reducing operating costs. A detailed analysis of the processes of preclinical studies allows us to draw conclusions about the possibility of a unified and improved presentation of business processes by introducing the concept of a minimum unit of integrated activity (MUnIA). Such a presentation of the MUnIA in the context of the analysis of pharmaceutical enterprises business processes reflects the cyclical nature of the relevant work (*Figure 4*).

Any operational process at the preclinical stage begins with the formation of many hypothetical constructs that need to be verified. Verification of the generated hypotheses is a mandatory and logical continuation of the systematic process of operational activity. If it is impossible to move from one stage to another, validation is carried out at the stage of hypothesis testing (small cycle), and in case of repeated failure of the transition, a return to the stage of formation of many hypotheses for repeated test-

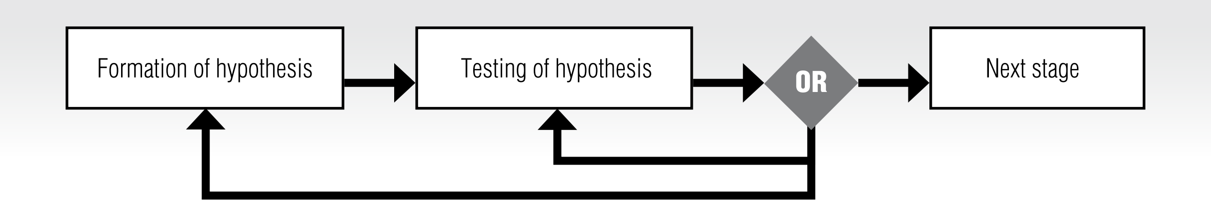


Fig. 4. The scheme of the minimum unit of integrated activity (MUnIA)

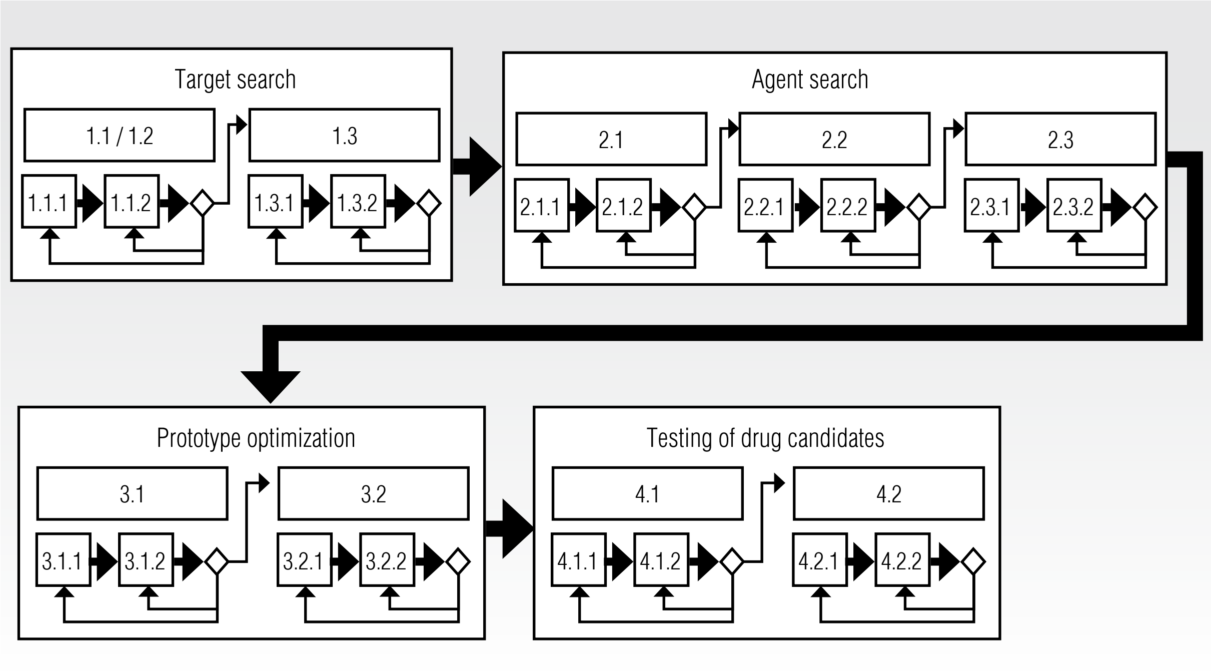
ing (large cycle). This cyclical design not only reflects the minimum systematic process of integrated research activities for the creation and development of drugs, but also associates entities with added value, which can worsen or otherwise prevent the fulfillment of the intended function of the business process. The main algorithm for working with a minimum unit of integrated activity is the complete decomposition of the business processes of a particular stage of research to an appropriate scale, the determination of the place of current activity in the model, the rationalization of the stage and subsequent improvement (adaptation to MUnIA cycles).

To demonstrate MUnIA applicability, the framework was simulated using the complex of preclinical stages of drug development as an example. So, at figure 5 each iteration of the stage of research work is schematically presented in accordance with the detailed decomposition and approbation of the model. The main methodology for verifying the applicability of the MUnIA was in-depth interviews with representatives of the pharmaceutical industry. At the same time, an in-depth interview is understood as an informal personal conversation with respondents, within the framework of previously agreed topics and structure.

In forming the sample of companies, the following criteria were used. First, the enterprise should be engaged in research and development of medicines. Secondly, the enterprise should be international and at the same time operate on the territory of Russia. Thirdly, the number of employees must be at least 1000 people. Thus, four companies were included in the study sample. The research design involved the participation of one to three managers from each enterprise.

The in-depth interviews on model verification and its application were conducted with five managers of Russian pharmaceutical companies. Respondents were selected taking into account their high awareness of the research activities of the organization. Interviews were conducted from August 1, 2019 to October 14, 2019. The duration of each interview ranged from 20 minutes to one hour; the average interview time was 35 minutes. Each respondent was asked 12 questions with specific answer options (“yes”, “rather yes than no,” “rather no than yes,” “no”) and two open questions. The first five questions clarified the activities and operational processes of the pharmaceutical company. The following three questions served to verify the main stages of the drug development life cycle. The last four questions clarified the concept of the proposed model and the possibility of its use in the operational process, namely: “Does the MUnIA model describe the minimum life system in the framework of preclinical studies?”, “Is it possible to optimize operational procedures for preclinical tests using the MUnIA model?”, “Is it possible to implement the MUnIA model in the preparation of SOPs? ”, “Is it possible to implement the MUnIA model in the content of the SOP of your company?” Two open questions were asked to determine directions for finalizing the approach to the analysis of business processes in general and the model: “what else should be taken into account in this approach to the analysis of business processes?” and “how can the MUnIA model be instrumentally integrated into the current information ecosystem of pharmaceutical enterprises?”

The MUnIA model was verified by all respondents (with certain conditions and improvements) as possible and recommended for use. A diagram was also formed (*Figure 5*), which stipulated the technical applicability and possibility of MUnIA implementation in operational business processes of the drug life cycle stage. For example, when considering such a sub stage of preclinical drug development as target search, it is necessary to consider that the initial goal is to determine the targets that will be affected by the compound, their number and type. In practice, the project’s working group considers a specific disease by analyzing a cascade of proteins. This procedure is a stage in the formation of many hypotheses about specific interactions that determine deviations. This is followed by a direct verification of scientific assumptions about the potential types and number of candidates in the target, using appropriate software tools. With successful validation, a transition from processes 1.1 / 1.2 (determination of the number of targets / determination of types



1.1. Determination of the number of targets
1.1.1. Protein cascade analysis
1.1.2. Testing protein interaction methods
1.2. Target type determination
1.3. Target validation
1.3.1. The formation of hypotheses about the relationship of the disease and targets
1.3.2. Proof of concept (PoC)
2.1. Virtual screening
2.1.1. Defining databases for library search
2.1.2. Selection of chemical libraries
2.2. High performance screening
2.2.1. Defining of test systems
2.2.2. Series testing
2.3. Estimation of ADME
2.3.1 Testing of hypothesis
2.3.2 Compounds testing
3.1. Analysis by a group of molecular biologists
3.1.1. Modeling
3.1.2. Pharmacology / Pharmacokinetics / Toxicology
3.2. Analysis by a group of chemists
3.2.1. Modeling
3.2.2. Pharmacology / Pharmacokinetics / Toxicology
4.1. In vitro tests
4.1.1. Formation of hypothesis
4.1.2. Testing of hypothesis
4.2. In vivo tests
4.2.1. Formation of hypothesis
4.2.2. Testing of hypothesis

Fig. 5. Preclinical operating procedures in accordance with the MUnIA

of targets) to process 1.3 (validation of targets) occurs. If the outcome is negative, the study is revised as part of operation 1.1.2 (verification of protein interaction methods), as part of the small cycle of the MUnIA model. In the case of a repeated negative outcome, the analysis is redirected along a large cycle to process 1.1.1, where a new set of hypotheses is formed for further verification.

Similar testing of the model is possible at any of the sub stages of preclinical trials. Thus, MUnIA is a management framework that can not only affect the main operational and intermediate procedures, but also provide systematization when monitoring the current activity of drug product development by project managers. In general, respondents confirmed the possibility of using the MUnIA model to manage and improve the preclinical stage of the study, as well as the possibility of using the above scheme as a recommendation for a unified presentation of SOPs.

Conclusion

This article analyzes the complex business processes of some full-cycle pharmaceutical enterprises at the preclinical stage for their further improvement. Qualitative analysis is based on data presented in industry regulations, open information from a number of commercial pharmaceutical companies, and in-depth interviews with industry representatives. The authors aggregated and schematically presented the main business processes of a pharmaceutical enterprise of this type. Based on the compilation, description and subsequent decomposition of the drug's life cycle, the problem of the lack of a standard unified form of processes representation at each operational iteration of preclinical trials has been identified which is a potential source of increased production time. After that, the model of the minimum unit of integrated activity (MUnIA) was proposed; this can be considered as a fundamental unit of standard operating procedures that describe the sequence of actions for developing drugs. The model can be used by both drug discovery research managers and project managers.

Thus, the priorities for further research may be the quantitative confirmation of the MUnIA efficiency at the preclinical stage, as well as the analysis of options for potential scaling of the model to other stages of the life cycle of drug production. ■

References

1. Lichtenberg F.R. (2008) Have newer cardiovascular drugs reduced hospitalization? Evidence from longitudinal country-level data on 20 OECD Countries, 1995–2003. *The National Bureau of Economic Research*. Available at: https://www.nber.org/papers/w14008 (accessed 30 June 2019).
2. Pilnikova E.G. (2016) Features of the activities of pharmaceutical manufacturing companies in modern conditions. *Business Education in the Knowledge Economy*, no 1, pp. 61–64 (in Russian).
3. Golovko A.S., Golovko Yu.S., Ivashkevich O.A. (2012) Modern methods of searching for new drugs. Bulletin of BSU. Series 2: Chemistry, *Biology, Geography*, no 1, pp. 7–15 (in Russian).
4. PhRMA (2015) *Biopharmaceutical research & development: The process behind new medicines*. Available at: http://phrma-docs.phrma.org/sites/default/files/pdf/rd_brochure_022307.pdf (accessed 23 June 2019).
5. Tufts University (2016) *Outlook 2016. Tufts Center for the Study of Drug Development*. Available at: https://static1.squarespace.com/static/5a9eb0c8e2ccd1158288d8dc/t/5aa2fc9d0852297555747051/1520630944033/Outlook-2016.pdf (accessed 23 June 2019).
6. Locuson C. (2016) Project management in drug discovery: Current practices and opportunities. *Project Management*. Available at: https://www.projectmanagement.com/articles/331399/Project-Management-in-Drug-Discovery--Current-Practices-and-Opportunities (accessed 23.06.2019).

7. Frearson J., Wyatt P. (2019) Drug discovery in academia – the third way? *Expert Opinion on Drug Discovery*, vol. 5, no 10, pp. 909–919. DOI: 10.1517/17460441.2010.506508.

8. Raman A., Tok W.H. (2018) *A developer's guide to building AI applications. Create your first intelligent bot with Microsoft AI*. Sebastopol, CA: O'Reilly Media.

9. GOST ISO/IEC 17025-2009. *General requirements for the competence of testing and calibration laboratories*. Available at: http://docs.cntd.ru/document/gost-iso-mek-17025-2009 (accessed 31 July 2019) (in Russian).

10. GOST 33647-2015. *Principles of good laboratory practice (GLP). Terms and definitions*. Available at: http://docs.cntd.ru/document/1200129061 (accessed 31 July 2019) (in Russian).

11. Chen J., Luo X., Qiu H., Mackey V., Sun L., Ouyang X. (2018) Drug discovery and drug marketing with the critical roles of modern administration. *American Journal of Translational Research*, vol. 10, no 12, pp. 4302–4312.

12. Pattanaik A. (2014) Complexity of project management in the pharmaceutical industry. Proceedings of the *PMI Global Congress 2014, Dubai, United Arab Emirates, 5–7 May 2014*. Available at: https://www.pmi.org/learning/library/project-management-complexity-pharmaceutical-industry-1487 (accessed 23 June 2019).

About the authors

Mikhail V. Belov

Cand. Sci. (Tech.);

Deputy CEO, IBS Group Holding Ltd., 3 build. 1, Skladochnaya Street, Moscow 127018, Russia;

Head of the Department of Information Business Systems, NUST MISIS,
8/11, Maly Tolmachevsky Lane, Moscow 119017, Russia;

E-mail: mbelov@ibs.ru

Mikhail A. Shakhmuradyan

Doctoral Student, Department of Economics of Innovation, Faculty of Economics,
Lomonosov Moscow State University,
1 build. 46, GSP-1, Leninskie Gory, Moscow 119991, Russia;

E-mail: mshakhmuradyan@econ.msu.ru

DOI: 10.17323/1998-0663.2019.4.28.38

Cognitive analysis and choice of an enterprise's strategic goals

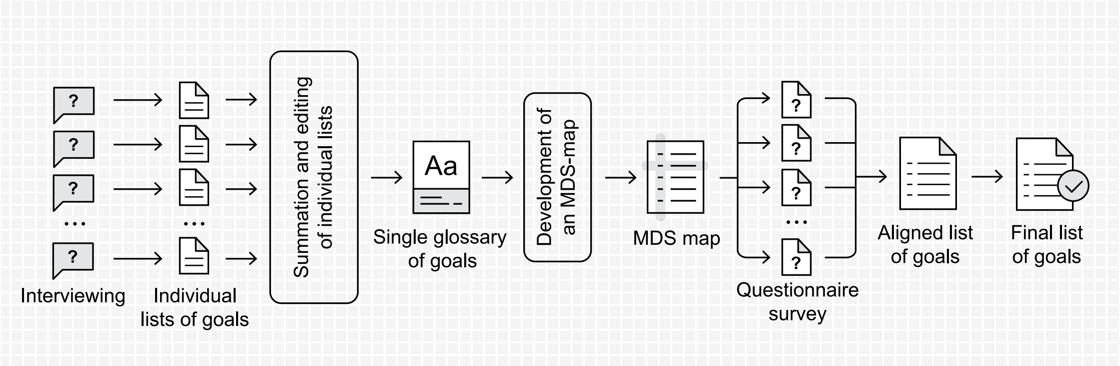
Robert A. Karayev
E-mail: karayevr@rambler.ru

Institute of Control Systems, Azerbaijan National Academy of Sciences
Address: 9, B. Vahabzade Street, Baku AZ1141, Azerbaijan

Abstract

One of the most crucial and vulnerable stages of strategic management is the cognitive stage associated with the transformation of the strategic vision and of the enterprise's mission into its strategic goals. At this stage, management is faced with the problem of developing a coordinated collective opinion on the content of the goals being formed and with the problem of objective assessment of their effectiveness. The difficulties here are due to the phenomenological features of the stage, such as the informal nature of the transformation procedure, the multi-criteria nature of goals, numerous uncertainties and risks exacerbated by the increased variability of business environments, cognitive barriers caused by linguistic discrepancies and differences in the professional experience of strategy developers. Such features of the stage ultimately lead to ambiguous decisions regarding the content of goals and ambiguous assessments of their effectiveness. In these circumstances, traditional support tools (numerous versions of expert methods, brainstorming, Norton and Kaplan's BSC, SMART technology, etc.) face serious limitations. This paper proposes a cognitive technology for forming a coordinated set of the enterprise's business goals that to a large extent takes into account the features of the given stage. The technology is a single procedure integrating the capabilities of traditional support tools and expanding the creative potential of support based on psychosemantic models and nonmetric multidimensional scaling methods. The results of a real study conducted at a number of enterprises show that cognitive technologies open up new prospects for goal analysis. They can serve as a useful complement to existing support tools and contribute to the design of more effective and realistic business strategies.

Graphical abstract



Key words: strategic goals; analysis and choice technology; cognitive approach.

Citation: Karayev R.A. (2019) Cognitive analysis and choice of an enterprise's strategic goals. *Business Informatics*, vol. 13, no 4, pp. 28–38. DOI: 10.17323/1998-0663.2019.4.28.38

Introduction

The problem of forming an enterprise's strategic goals (hereinafter referred to as "the goals") is one of the critical problems of strategic management [1–3]. The success of a strategic project largely depends on its solution. The complex, informal, heuristic nature of the problem has long been at the root of the issue of creating effective support tools that adequately reflect the phenomenology of the problem itself, the accumulated positive experience of using traditional support tools, as well as new features of modern intelligent technologies.

In the following sections, we list the most popular support tools, giving a brief critical analysis of these tools. Based on the results of our analysis, the basic principles that can be used as the foundation for an improved support technology are formulated, and the prospects of the cognitive approach in this context are substantiated. We consider the challenges of the implementation of the cognitive approach and ways to overcome them by means of models of experimental psychosemantics.

The basic principles, the diagram and operation algorithms of the proposed support technology are provided. A demonstration example of its use is provided and the possibilities and prospects of its application are discussed.

1. Brief description of the traditional methods

The most popular support tools currently used in strategic management in the selection of an enterprise's strategic goals include expert methods (survey, interviewing, Delphi method), brainstorming, SMART technology, and Norton and Kaplan's strategic cards (BSC).

Expert methods [4]. The conceptual basis of these methods is the "general evaluation scheme" proposed in [4]:

$$E = <\Omega, \Omega e, L, Q, \varphi>,$$

where Ω – the initial set of permissible goal assessments according to the accepted criterion;

Ωe – the set of the permissible goal assessments made by various experts ($\Omega e \subseteq \Omega$);

L – the rules of interaction between experts;

Q – evaluation feedback;

φ – the method for processing the assessments of various experts in order to determine the resulting assessment.

The expert methods widely used to formulate goals are, in fact, various modifications of this general scheme.

Brainstorming technique [5]. The once widely advertised "brainstorming" method was regarded as a way of "enhancing" the creative capabilities of a team of analysts in solving ill-structured problems. This is a group creativity technique by which efforts are made to find a conclusion for a specific problem by gathering a list of ideas spontaneously contributed by its members.

SMART technology [6]. This is a declarative approach that positions selected goals according to the following criteria: specific (S), measurable (M), achievable (A), relevant (R), time bound (T).

BSC strategic cards (balanced scorecards) [7]. The BSC concept proposed by Kaplan and Norton involves a hierarchical structuring of an enterprise's performance. The balanced scorecard system uses strategy at four interconnected levels – the financial level, customer level, process level, and learning and growth level. This structuring focuses on groups of goals, allowing one to deal with the well-known "dimensionality problem" [7]. Here, the number of goals in each of the layers is usually taken as equal to 5–7, which is associated with limited human psychophysiological capabilities [8].

2. Critical analysis of the traditional methods

An analysis of the aforementioned methods and the practice of their application allows us to formulate a number of significant limitations that arise in their practical use and to lay out the main ways to improve the support procedure.

As noted above, the primary task to be solved at the cognitive stage is forming the set of goals. They can be established by:

1) survey method [9], using a set of goals predetermined by experts;

2) interviewing method [10], when experts offer their own individual opinion regarding the set of goals;

3) Delphi method [11], which is a group survey technique. The procedures used in the Delphi method are characterized by three key features: anonymity, regulated feedback and group response. Feedback is provided through several survey rounds, with the results of each round being processed by statistical methods and communicated to the experts. In the second and subsequent rounds, experts argue their answers. Thus, in subsequent rounds, experts can revise their initial answers. From round to round, experts' answers become increasingly stable until they stop changing, which stops the survey. Practice shows that three or four survey rounds are usually conducted, since the estimates no longer change after that.

The common feature of these expert methods is that they are questionnaire-based. The difference is that their implementation can be carried out in closed form (experts are offered ready questionnaires) or in open form (experts form questionnaires themselves and fill them in during the interview process). Depending on the form of questionnaires, various difficulties may arise. Thus, the use of an open form can result in different sets of goals being proposed by each expert and an ambiguous interpretation of goals, since the same goals can be referred to by different experts in different terms, or goals with the same name can have different semantics and pragmatics. Using a closed form of questionnaires can significantly limit experts' ability to express their opinions on a set of goals and their verbal wording.

This problem is most clearly manifested in another popular method, the "brainstorming" technique. Even in the early days of this method, publications appeared that cast doubt on its effectiveness in generating ideas [12]. Studies of the brainstorming technique revealed three groups of processes that reduce its effectiveness [13]:

✦ social loafing, which allows managers to hide behind the backs of colleagues;

✦ evaluation apprehension – the fear of being judged by colleagues for possibly "silly" ideas;

✦ production blocking due to lowered criticality, since according to the technique, any member of the group can support any idea at any time.

Studies over the past decade show that individual methods of goal generation are still more effective than collective ones [14].

Among the support tools, the BSC approach deserves special attention. In addition to highlighting an enterprise's goals, the structural organization of this approach involves setting their mutual influences both within each of the layers and between the layers. For instance, the staff of an enterprise, even given ideal quality of its employees, can achieve results in customer relations only if the business processes of management are properly organized (the influence of the goals of the "business processes" layer on the goals of the "customer" layer). Properly organized (efficient and rational) business processes makes it possible to achieve maximum performance defined in the customer layer (market share, service satisfaction), which, in turn, allows one to achieve the desired financial results.

However, the BSC approach is only attractive up to the moment the management encounters the question: where can we get the necessary and sufficient set of business goals for a specific enterprise? Obviously, the answer here is unambiguous: only from the heads of the managers developing the strategy of the enterprise. However, the task of explicating this knowledge and its practical use is far from trivial. Its simplicity without the use of special support tools is illusory and deceptive.

3. Cognitive technology for generating and analyzing goals

3.1. Key principles

The considerations set out above allow us to formulate the key principles that can form the basis of an improved goal formation technology. First, the technology should be based on an individual rather than a group method of evaluation. Second, the technology should provide tools for solving the inevitable problem associated with the explication of managers' internal representations, analysis and coordination of their individual representations and the formation of a single collective opinion. The peculiarity of this problem lies in the fact that the explication procedure is informal, while the goal assessments themselves are multi-criteria (SMART technology, for instance, sets goals according to five criteria) and often non-metric (qualitative, linguistic) in nature, which excludes the possibility of quantitative processing of the source expert material.

Attempts have been made for a long time to use the cognitive approach to solve this type of problems [15–17].

One of the central points of the cognitive approach is the assertion that individual human behavior which is formed in response to external stimuli is determined by the structure of a person's representations in the subject area to which a specific external stimulus belongs. Such structures of human internal representations are called "mental models" [18]. Mental models are internal representations of causal relationships within a system, allowing a person to understand, predict and solve practical problems associated with this system. Mental models are based on human experience and expectations. They control our behavior in various situations and are dynamic constructs that change under the influence of learning, new information or a person's state. A person can mentally manipulate mental models, "starting" them in the form of an internal experiment and evaluating its results under different conditions and a different sequence of steps that form such a model. The described manipulations are the internal basis for the formation of all the main components of an enterprise's strategy.

Thus, managers have internal, mental models of competition [19], and entrepreneurs have mental models of the industry [20]. Both of these mental models help to assess the current situation and make justified decisions.

However, the use of the cognitive approach in the formation of goals is complicated for a number of reasons. The main such reasons are the following.

First, all of an enterprise's goals are products of managers' thoughts, sitting in their heads in the form of metal models and should be extracted in the clearest form possible. However, the form of human internal representations hinders the solution of this problem. "Brain languages" [21], "non-disjunctive" human logic [22] cannot be directly translated by means of the traditional disjunctive means of modern mathematics.

Secondly, the complexity of translation is compounded by the fact that an enterprise's goals are the product of thinking of many people – the enterprise's owner, and top and middle management involved in the strategy development. This results in the problem of coordinating many goals and moving from individual mental models to collective, team knowledge.

Third, the difficulties of application of the cognitive approach are also due to the fact that the possibilities of this approach are limited by human psychophysiological capabilities [23].

Accordingly, the number of goals included in the task of goal analysis should be limited by these capabilities. On the other hand, the set of goals being analyzed should cover almost all aspects of an enterprise's operation. The BSC approach is best suited to solve the "dimensionality problem" that arises in this case. However, the aforementioned difficulties of explication impose serious limitations on its practical use. The issue of reasonable detailing of the problem field and the issue of reduction of goals by excluding less significant goals from consideration become unavoidable here.

Our analysis has shown that the listed difficulties of applying the cognitive approach are surmountable. Models of experimental psychosemantics open up broad prospects for overcoming these difficulties [24, 25]. In recent years, models of experimental psychosemantics find increasingly wide application in social and economic research.

Experimental psychosemantics is an area of cognitive science that studies various forms of representing objects of the world in the individual human consciousness (images, symbols, verbal forms). The main method of experimental psychosemantics is the construction of so-called "subjective semantic spaces" (an individual's model representations of objects of the world) by nonmetric multi-dimensional scaling [26–28]. This is a method of making subjective assessments when the subject (expert) is asked to evaluate an object by a set of attributes, using scales based on verbal gradations. Scaling in this context differs from a single measurement in that it allows for individual observations to recreate a holistic image of the analyzed object. An important advantage of the method is that it allows us to identify the presence of different points of view on the object being analyzed by experts and to coordinate their opinions on the syntactic and semantic levels. At the same time, differences between expert assessments are not considered experimental errors but are important in themselves.

The use of psychosemantic models provides an important advantage. It makes it possible to apply a mixed approach to the formation of goals, which effectively combines the positive aspects of the interviewing method (that maximizes the goal searching space), the survey method (that involves describing goals in common terms), SMART technology (that takes into account the multi-criteria nature of goals) and the BSC concept (that structures the problem field of target analysis).

Our technology for generating and analyzing goals takes into account these key points.

In the following paragraphs, we give a general description of the technology and a demonstration of its practical application. The possibilities and prospects of the cognitive selection of goals are discussed below.

3.2. The diagram of the technology

The block diagram of the technology presented in *Figure 1* shows the sequence of steps that implement the process of analysis and formation of goals.

3.3. Goal alignment method

The goals alignment method is based on remote nonmetric multidimensional scaling models based on the Euclidean metric [27, 28] that meet the conditions of the problem we are investigating. The goal alignment diagram in this case is as follows:

$$\mathbf{C} = <\mathbf{I}, \mathbf{P}, \mathbf{Z}, \mathbf{E}, \varphi>,$$

where **I** – the name of the goal;

P – the set of attributes (dimensions) of the goal;

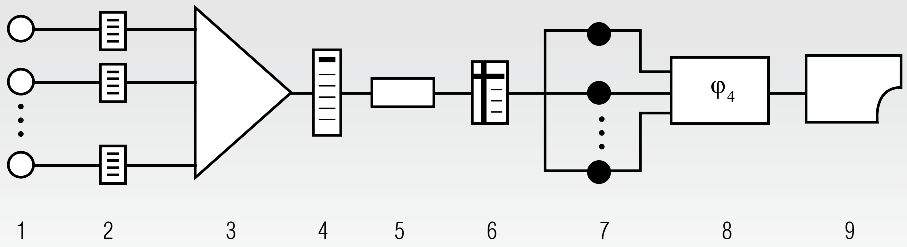
Z – the nonmetric attribute assessment scales;

E – the individual goal assessments for each of the attributes made by various experts;

φ – MDS algorithm for processing individual assessments and determining the degree of consistency of goals.

The algorithm of the method can be demonstrated through the following example.

Assume we have a questionnaire card (MDS map), which is (for clarity and simplicity of presentation) a matrix of two-dimensional scaling of goals by the attributes: “Significance of the business goal” and “Attainability of the business goal” from the management’s point of view. Each of the attributes is evaluated on a linguistic scale: “high,” “medium,” “low.” Assume also that the number of experts is M. Each of the experts places goals from the single glossary of goals (SGG) in a cell of the matrix in accordance with his or her ideas about the significance and attainability of the goal. Obvi-



1 – interviewing experts (open survey);
2 – individual lists of goals (ILG) received from various experts;
3 – simple summation of individual lists and their editing;
4 – a single glossary of enterprise goals (SGG);
5 – development of an MDS (multidimensional scaling) map – multidimensional matrix of goal assessment criteria [26, 27];
6 – MDS map;
7 – questionnaire survey of experts (closed survey) based on the SGG;
8 – processing of the results of the questionnaire survey by MDS methods, ranking and selection of the aligned list of goals (ALG);
9 – final list of goals (FLG).

Fig. 1. The block diagram of the technology of data generating and analysis

ously, some cells of the matrix can end up with several goals and some cells without a single one. Denote by m_{ij} the number of experts who placed the same goal in the cell (i, j). It is clear that $\sum m_{ij} = M$.

After each of the experts has filled the matrix, the aggregation of the obtained assessments is carried out. To this end, each cell is assigned with respect to each target ($c \in C$) a weight $\lambda_{ij}^{c} = m_{ij} / M$, $\lambda_{ij}^{c} = [0, 1]$, which, in fact, is a two-dimensional function of the distribution density of expert opinions. To take into account the misalignment of opinions, a measure called the coefficient of inconsistency (K^c) is introduced, which characterizes the degree of inconsistency of the management's opinions regarding this goal *c*.

If the opinions of the experts regarding the goal *c* coincided and they filled, for instance, cell (2, 3), then the coefficient K^c would be equal to zero, and the coordinates of this opinion would be equal to (2, 3). But since the experts' opinions are scattered across the matrix, in order to assess the degree of inconsistency, it is necessary to find the mean abscissa I^c, the mean ordinate J^c and the coefficient of inconsistency K^c. Since the function of the distribution density of opinions has already been determined, these values are calculated from the following formulas:

Mean inconsistency abscissa:

$$I^c = \sum_{ij} \lambda_{ij}^{c} \cdot i \, ;$$

Mean inconsistency ordinate: ;

$$J^c = \sum_{ij} \lambda_{ij}^{c} \cdot j \, ;$$

Inconsistency coefficient:

$$K^c = \sum_{ij} \lambda_{ij}^{c} \cdot \sqrt{(i-I)^2 + (j-J)^2} \, .$$

3.4. Rules of goal ranking

The following rules of goal ranking can be established on the basis of the coefficient K^c.

Rule 1. If $K^c < 0.5$ then the degree of inconsistency of the management's views on the goal *c* is low and the goal c can be included in the FLG;

Rule 2. If $K^c = [0.5; 0.75]$ then we have an uncertainty and further elaboration is required regarding the inclusion of the goal *c* in the FLG;

Rule 3. If $K^c > 0.75$ then the degree of inconsistency of opinions on this goal is high and the goal cannot be included in the FLG (the management does not perceive this goal as a goal and will actually exclude it from the strategy in any case).

4. Technology application example

The technology presented has been used in a number of real-life projects. Table 1 shows the results of the application of the technology to the selection of business goals for a poultry enterprise at an agricultural holding in Baku. The following business goals of the enterprise were considered (the appropriate inconsistency coefficients K^c are presented in the parenthesis):

1. Increasing production of broilers to 4500 tons per year, and of hatching eggs – to 10 million pcs per year ($K^c = 0.3$);
2. Raising the market share of the enterprise in 2013–2017 in the domestic market to 15 % ($K^c = 0.32$);
3. Achieving annual sales of up to 17.5 million AZN ($K^c = 0.33$);
4. Implementing a new marketing strategy and creating a wide marketing chain ($K^c = 0.58$);
5. Further training of production personnel ($K^c = 0.32$);
6. Modernization of technological equipment ($K^c = 0.59$);
7. Using new productive breeds of poultry ($K^c = 0.69$);

8. Involving international strategic management experts ($K^c = 0.53$);

9. Development and introduction of new feeding diets ($K^c = 0.67$).

Notation in the table: X (Y), where X is the goal number, Y is the number of experts who placed the goal in the appropriate cell.

The table shows that testing the single glossary of goals made it possible, based on the analysis of the coefficients K^c, to reveal the existence of a consistent opinion regarding only a portion of the goals (4 out of 9). These are goals 1, 2, 3 and 5: they all have values of the inconsistency coefficient K^c from 0.3 to 0.33, and this allows us to consider inconsistency of the managers' opinions as low. On the other hand, the management cannot make a decision on the remaining goals, regarding either their significance or the possibility of achieving them. This means either that their inclusion in the business strategy is ineffective or that the goal analysis procedure requires an additional iteration.

5. Discussion

5.1. Inconsistency of the managers' opinions

The practice of applying the proposed technology has shown that different outcomes are possible when assessing the degree of inconsistency of management's opinions:

1) management's opinions are completely inconsistent (the assessment of inconsistency is high, and it is impossible to distinguish groups of managers with close opinions). In this case, the results of the evaluation are obviously not suitable for decision-making. Depending on the specific situation, one should either consider the evaluation unsuccessful and choose not to conduct the study, or conduct a second evaluation. Re-evaluation should account for the possible reasons of the failure, such as, e.g. the goals were incorrectly formulated, the goal assessment scales were selected poorly, it was impossible to create the right psychological and material environment, the managers have hidden personal and group interests, etc.;

2) managers' assessments are divided into several groups, within each the consistency is quite high, but it is low in the whole by the team of managers. Therefore, it is logical to assume that this is a case of different methodological approaches or different social groups. In this case, managers' opinions cannot always be brought to consistency among themselves even through a lengthy discussion. Therefore, it is advisable to supply the decision maker with several group assessments with appropriate comments;

3) group assessment is highly consistent. Such

Table 1.

Assessment and ranking of business goals of a poultry enterprise

		Significance of the business goal		
		High	Medium	Low
Attainability of the business goal	High	1 (4); 5 (4); 6 (2); 7 (2); 8 (2); 9 (1)	4 (1); 6 (1); 9 (2)	
	Medium	1 (1); 2 (4); 3 (4); 4 (3); 5 (1); 6 (1); 8 (1); 9 (2)	3 (1); 4 (1); 7 (2)	
	Low	2 (1); 6 (1)		

an assessment can be presented to the decision maker, but in any case, it makes sense to analyze the presence of extreme opinions (to find out what percentage of managers have them, how they substantiate their points of view, what the assessment of inconsistency will be if they are not taken into account).

Thus, the situations that management faces when forming a set of goals for an enterprise are quite diverse. Therefore, depending on the specific situation, the support technology described above can be supplemented by all available sources of theoretical and reference information for calculations and additional analysis.

5.2. Goal selection

The problem field of the goal selection problem is very large. It will suffice to look at the structure of the BSC card to know that it is almost impossible to solve the problem. Here strategy developers are confronted with the "dimensionality problem," the inevitable companion of all complex projects. To solve the problem within the ideology of the cognitive approach, the "camera metaphor" was proposed [29], in which a "camera" glides across the "picture of the world" and, by the operator's will, selectively captures fragments of this world, zooming in and out on a fragment of interest. This metaphor is certainly productive, but it must be supplemented by a very important circumstance. When exploring a fragment of the world, the analyst and only the analyst alone can exercise intuitive control of the entire problem field as a whole, investigating all aspects of the problem and understanding how they relate to the problem being solved within the context of the selected fragment. This circumstance is fundamental and it indicates that the cognitive approach should be interactive in nature and it is incorrect to use it as a local computer program, as is the case in expert systems [30] and in numerous studies on cognitive modeling[1].

Cognitive models should be used as a research tool and be open to contextual analysis by human beings.

5.3. BSC layers

The proposed goal selection technology can be carried out for all four layers of the balanced scorecard (BSC). However, solving each separate local problem, managers must monitor changes in its parameters in the context of possible changes in other "related" fragments of the problem field.

5.4. Context analysis

Today, in the growing complexity of economic relations, it is customary to talk about the increasing role of context [31]. Contextual analysis in the process of goal formation is the challenge and requirement of today's management practice. The ability to correctly take into account the context based on the knowledge of cognitive technologies places increased demands not only on the basic training of managers, but also on their intuition. If mastering the technologies of cognitive analysis puts forward special tests for the professional training of strategy developers, then the ability to "embed" these technologies in a specific context requires well-developed intuition. This is already a new quality expected from strategy developers which is difficult to achieve on the basis of the traditional training system. This essentially refers to the recently discussed issue of revising business education programs.

Conclusion

Practice shows that when developing management strategies, the stage of the transition from the strategic vision and mission of the

[1] International Conference on Cognitive Modelling (https://iccm-conference.github.io/previous.html)

enterprise to the formulation of its business goals is often the most vulnerable one from the point of view of loss of effectiveness.

The greatest difficulties here are due to the transition from the individual opinions of top managers to a coordinated (team) opinion.

This strategic management issue currently does not have sufficiently effective support tools. The support technology proposed in this paper, based on psychosemantic models and the nonmetric multidimensional scaling method, implements an end-to-end procedure for objectivizing management's subjective opinions on the enterprise's goals and makes it possible to move from individual knowledge to coordinated team knowledge. In real conditions (insufficient and varied practical experience, insufficiently high qualifications, conflicting opinions, personal and group interests), the technology can serve as a tool to complement well-known support tools and, consequently, be very useful for the appropriate selection of an enterprise's development strategy in complex modern economic conditions. ■

References

1. Drucker P.F. (2001) *Management challenges for the 21st century*. N.Y.: Harper Business.
2. Kleiner G.B. (2008) *Enterprise strategy*. Moscow: Delo (in Russian).
3. Isaev D.V. (2008) The Rational model of strategic management. *Financial Weekly*, no 25, pp. 14–15; no 26, p. 12 (in Russian).
4. Makarov I.M., Vinogradskaya T.M., Rubchinsky A.A., Sokolov V.B. (1982) *Theory of choice and decision making*. Moscow: Nauka (in Russian).
5. Leskov S.L. (2012) *Brainstorming*. Moscow: MSU (in Russian).
6. Bogue R.L. (2005) *Use S.M.A.R.T. goals to launch management by objectives plan*. Available at: https://www.techrepublic.com/article/use-smart-goals-to-launch-management-by-objectives-plan/ (accessed 20 November 2013).
7. Kaplan R.S., Norton D.P. (1996). *The balanced scorecard: Translating strategy into action*. Boston: Harvard Business School Press.
8. Miller G. (1955) The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, vol. 101, no 2, pp. 343–352.
9. Walsh J. (1988) Selectivity and selective perception: An investigation of managers' belief structures and information processing. *Academy of Management Journal*, vol. 31, no 4, pp. 873–893. DOI: 10.5465/256343.
10. Markiczy L., Goldberg J. (1995) A method for eliciting and comparing causal maps. *Journal of Management*, vol. 21, no 2, pp. 305–333. DOI: 10.1177/014920639502100207.
11. Loo R. (2002) The Delphi method: a powerful tool for strategic management. *Policing: An International Journal of Police Strategies & Management*, vol. 25, no 4, pp. 762–769. DOI: 10.1108/13639510210450677.
12. Isaksen S.G. (1988) *A review of brainstorming research: Six critical issues for inquiry*. Available at: https://www.semanticscholar.org/paper/A-Review-of-Brainstorming-Research%3A-Six-Critical-Isaksen/4abc961cb62e8b230f9683125e984eec3550caa4 (accessed 22 June 2016).
13. Furnham A. (2000) The brainstorming myth. *Business Strategy Review*, vol. 11, no 4, pp. 21–28. DOI: 10.1111/1467-8616.00154.
14. Paulus P., Dzindolet M. (1993) Social influence processes in group brainstorming. *Journal Personality and Social Psychology*, vol. 64, no 4, pp. 575–586. DOI: 10.1037/0022-3514.64.4.575.
15. Schwenk C.R. (1988) The cognitive perspective on strategic decision making. *Journal of Management Studies*, vol. 25, no 1, pp. 41–55. DOI: 10.1111/j.1467-6486.1988.tb00021.x.

16. Narayanan V.K., Zane L.K., Kemmerer B. (2011) The cognitive perspective in strategy: An integrative review. *Journal of Management*, vol. 37, no 1, pp. 305–323. DOI: 10.1177/0149206310383986.

17. Karayev R.A. (2015) Cognitive approach and its application to the modeling of strategic management of enterprises. *Knowledge engineering: Principles, methods and applications* (Ed. Alfonso Perez Gama). N.Y.: Nova Science, pp. 79–101.

18. Johnson-Laird P.N. (1980) Mental models in cognitive science. *Cognitive Science*, vol. 4, no 1, pp. 71–115. DOI: 10.1207/s15516709cog0401_4.

19. Johnson P., Daniels K., Asch R. (1998) Mental models of competition. *Managerial and organizational cognition: Theory, methods and research* (Eds. C. Eden, J.-C. Spender). London: SAGE Publishing, pp. 130–146.

20. Gary M.S., Prietula M.J., Feltovich P. (2017) Mental models as the interface between the business environment and strategic decisions. *Academy of Management Proceedings*, no 1. DOI: 10.5465/AMBPP.2017.14588abstract.

21. Pribram K.H. (1971) *Languages of the brain. Experimental paradoxes and principles in neuropsychology*. Englewood Cliffs, N.J.: Prentice-Hall.

22. Brushlinsky A.V. (1979) *Thinking and prediction*. Moscow: Myisl (in Russian).

23. Bays P.M., Husain M. (2008) Dynamic shifts of limited working memory resources in human vision. *Science*, vol. 321, no 5890, pp. 851–854. DOI: 10.1126/science.1158023.

24. Petrenko V.F. (2005) *The basics of psychosemantics*. Saint-Petersburg: Piter (in Russian).

25. Petrenko V.F. (2013) *Multidimensional consciousness: A psychosemantic paradigm*. Moscow: Eksmo (in Russian).

26. Young F.W. (1970) Nonmetric multidimensional scaling: Recovery of metric information. *Psychometrica*, vol. 35, no 4, pp. 455–473.

27. Green P.E., Carmone F.J., Smith S.M. (1989) *Multidimensional scaling: Concepts and applications*. London: Allyn and Bacon.

28. Tolstova Yu.N. (2006) *Basics of multidimensional scaling*. Moscow: University (in Russian).

29. Mintzberg H., Lampel J., Ahlstrand B. (2005) *Strategy safari: A guide tour through the wilds of strategic management*. N.Y.: Free Press.

30. Waterman D.A. (1986) *Guide on expert systems*. Reading, MA: Addison-Wesley.

31. Balatsky E. (2006) The dialectic of cognition and the new paradigm of economic science. *World Economy and International Relations*, no 7 (in Russian).

About the author

Robert A. Karayev

Dr. Sci. (Tech.);

Professor, Head of Ecosystems Modeling Laboratory,
Institute of Control Systems, Azerbaijan National Academy of Sciences,
9, B. Vahabzade Street, Baku AZ1141, Azerbaijan;

E-mail: karayevr@rambler.ru

DOI: 10.17323/1998-0663.2019.4.39.48

Simulation of artefact detection in Viber and Telegram instant messengers in Windows operating systems

Alexander I. Borodin [a]
E-mail: aib-2004@yandex.ru

Roman R. Veynberg [a]
E-mail: veynberg@gmail.com

Dmitry V. Pisarev [b]
E-mail: d.pisarev@warwick.ac.uk

Oleg V. Litvishko [a]
E-mail: Litvishko.OV@rea.ru

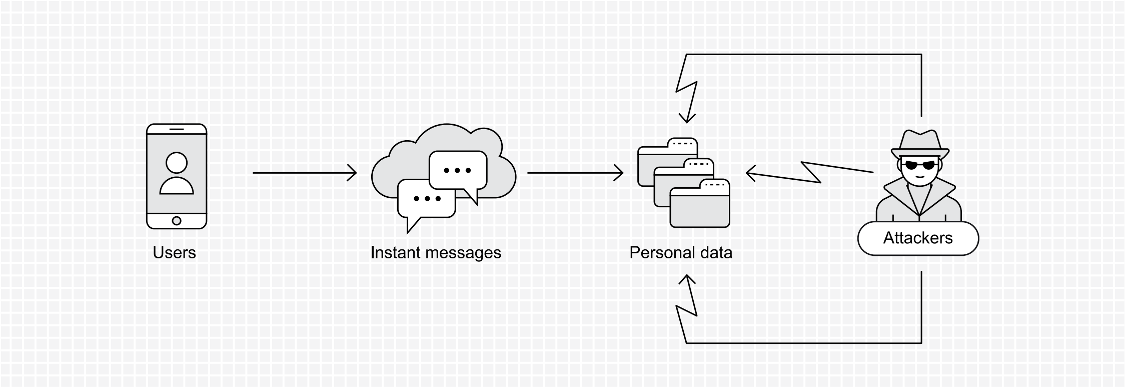
[a] Plekhanov Russian University of Economics
Address: 36, Stremyanny Lane, Moscow 117997, Russia

[b] University of Warwick
Address: Coventry CV4 7AL, United Kingdom

Abstract

Messengers are popular today on mobile devices and traditional computers. Starting as a small text messaging service, they have turned into effective communication channels for both private and corporate users, becoming more than just an SMS replacement. Users entrust to them a huge amount of information, such as a time-based map of activity, photos and other personal data. Messengers changed the way communication is done; they reduce the distance to the user and along with social networks become tools for fraud, spam or blackmail and terrorism. In this regard, it is vital to study instant messengers from a forensic point of view. This research explores and compares two popular messengers: Viber and Telegram, which is rapidly gaining popularity in the criminal world and the darknet as secure message tools. The main purpose of the research is to investigate and analyze potential artefacts remaining during the installation and use of instant messengers, as well as after their uninstallation. The authors have done several experiments to investigate the artefacts in different environments and provide clear explanation of the results. The experiments showed that even though Telegram is considered to be one of the most secure instant messengers, important and useful material on a hard drive and registry remain after complete uninstallation of the application. Exploring Viber artefacts showed up information that helps to restore the whole history of a communication. Moreover, the study confirmed that artefacts are still accessible in Windows after removal of the application.

Graphical abstract



Key words: instant messenger; computer science; simulation; exploit; artefact; Viber; Telegram.

Citation: Borodin A.I., Veynberg R.R., Pisarev D.V., Litvishko O.V. (2019) Simulation of artefact detection in Viber and Telegram instant messengers in Windows operating systems. *Business Informatics*, vol. 13, no 4, pp. 39–48. DOI: 10.17323/1998-0663.2019.4.39.48

Introduction

In recent years, instant messaging (IM) applications have gained in popularity because they are free of charge and easy to use. Nowadays IM is one of the most convenient ways to text messages, share files and videos, as well as make audio and visual calls. According to the research [1], worldwide IM user accounts are expected to grow to over 3.8 billion by year-end 2019.

The growing popularity of instant messengers relates also to various criminal activities such as fraud and terrorism [2]. They attract criminals by the opportunity the afford to simplify communication with victims or accomplices, as well as the availability of end-to-end encryption and other ways to secure or illuminate information that might be required by the authorities during an investigation.

However, despite the increased level of encryption and security, IM applications for Windows OS can provide to a potential researcher a lot of useful material. The artefacts can show information about the last date of launch, an SSID of the wireless network connected to the PC, outgoing connections, geolocations and other helpful information.

This research performs a forensic examination of popular Viber and Telegram applications by looking at the artefacts produced by IM applications. The interest in instant messengers grew with their popularity and IM applications have became the subject of various digital forensic studies.

Grispos et al. [3] tested user behavior from residual data in cloud-based synchronized applications. Communication between an attacker and victim were simulated, such as a file transfer and dialogs. The results of the study showed that artefacts remaining in the registry can link the criminal and the victim, such as traces of the file transfer between users and registry entries related to contact details. Moreover, fragments of the conversation can be recovered from memory dump.

Grispos et al. [3] also analysed residual data, simulating the conversation and file transfer between a suspect and a victim. Fragments of the conversation were found within the Windows 7 swap file, but the study of the mobile device did not provide much useful information. A broad list of artefacts that can be useful for forensic researchers such as references to the URLs and last access times was presented in the conclusion.

Cheng et al. [4] tested Windows Live Messenger installed on Windows 7. The results suggest that remaining artefacts allow one to restore the whole picture of a communication. Moreover, a user must be very competent to hide them.

Levendoski et al. [5] released information about the Yahoo messenger. Windows Vista and Windows 7 operating systems were used as platforms and comparisons conducted between OS artefacts remained after de-installation. The research showed that the structure of changes in the Windows 7 registry was modified inconsiderably compared to Windows XP.

Social network messengers have received attention from researchers because of their increased popularity. Al Mutawa et al. [6] studied Facebook chat based on web technology as a source of potential evidence for investigations. This article gives detailed information about possible artefacts, but their location depends on the browser and encoding. The study outlined a method for investigating Arabic string artefacts, but searching and converting them to readable view can take a lot of time to complete. However, the study is only limited to web-based Facebook chat.

Yasin and Abulaish [7] studied the Digsby IM aggregator to retrieve user sessions for use in investigations, despite attempts to hide information from a researcher. Results showed that they were similar to traditional IM applications. Despite the relatively recent date of the study, the messenger is not developing and supporting.

Karpisek et al. [8] studied an opportunity to decrypt traffic during WhatsApp communications and retrieve the details of calls. The current study presented a new approach to decryption of the information and found that calls can be decrypted. However, end-to-end encryption was changed by WhatsApp in 2016 and made the proposed method irrelevant.

The focus in research has recently shifted to social networks and cross-platform messengers.

Majeed et al. [9] studied three different applications: Facebook, Viber and Skype on the Windows 10 platform and the possibility to find artefacts. The result of the research showed that many artefacts are stored in one folder \AppData\Local\Packages\ for all the third-party applications. Moreover, for all applications they found artefacts saved as plain text files. The most important forensically relevant finding was common artefacts remaining for tested applications.

Dehghantanha et al. [10] studied Facebook and Skype messengers. The results indicated that artefacts could be recovered from a PC because of use of the Windows Store. IM applications installed using Windows Store leave elements valuable or critical to an investigation on the hard drive, in memory dumps and network captures.

To the best of the authors' knowledge, the number of studies that are focused on comparison of secure messengers such as Telegram and another widely used IM application such Viber in the Windows environment is limited. Telegram was investigated by Cahyani et al. [11] and Carvey and Hull [12] as a tool for terrorist-related activities. Results of the study can be of great value for forensic analysts, but the research was strictly limited on mobile devices only. As a result, it is necessary to fill the gap and study artefacts that Telegram application leaves in the Windows environment compare to Viber – another well-known messenger.

1. Methods

This section gives information about tests that were provided with Viber and Telegram messengers. The experiment was performed on Windows 10 installed in a virtual machine environment. For the research we created: a windows user with administrative rights ("user_a") and two new IM accounts (one for Viber and one for Telegram). Each of the IM apps was installed on a Windows installation. Interactions during the experiment were made using the author's personal account.

The artefacts were investigated through a series of research controlled experiments. All configuration changes were selected equally for both messengers. The detailed outline of the scenario and environment will be provided below.

1.1. Experimental environment

This study is based on the artefacts produced by two IM applications: Viber 6.9.6 and Telegram 1.1.23. The experiment was implemented on the following hardware platform: HP Z620 Workstation, CPU – Intel(R) Xeon(R) 2x E5-2660 2.20GHz, 16 Gb DIMM DDR3 (1866 MHz), 2TB Hard Drive using Ubuntu v.16.04.6 as the software operating system (OS).

Oracle Virtual Box (5.1.30 r118389 Qt5.6.3) containing Windows 10 Education (64bit, build 15063) was chosen as a platform for the experiment. The virtual workstation was configured with 4 GB RAM and 20 GB HDD space. Use of the virtual machine helped to make a considerable amount of snapshots and revert to a restore point quickly. As a result, this approach leaves researchers room for errors.

Registry and file data were collected using Regshot Portable v.1.9.0 which allow one to make a registry snapshot before and after a user activity and compare results.

The open source tool SQLite DB Browser v3.10.1-win64 was used for exploring details of databases. It helps to search, analyze and edit data and metadata in *.db files.

RegRipper v2.8 was used as a tool that helps to indicate user activity through analysis of the NTUSER.DAT file. The file provides very useful information (including key LastWrite times and data derived from binary and string values), indications of user actions. RegRipper userassist.pl plugin handles a translation UserAssist key which includes a 64-bit time stamp as well as a counter (referred to as a "run count") that appears to indicate how many times the user has interacted with the shell in the manner in which these values would be created or modified.

All software applications were installed with default setting and removed using standard Windows uninstaller.

1.2. Experiment procedure

The first step of the experiment was virtual machine creation, using Virtual Box containing Windows 10. The system was installed with default configuration and windows update service was disabled on the workstation for decreasing the number of artefacts not related to the experiment. Finally, we installed forensic tools and created a snapshot by Virtual Box. The snapshot was used as the "starting point" of the research for each IM application.

The second step was the IM app installation to collect and compare registry and file data using Regshot. The snapshots were performed on each IM application listed below in chronological order.

1. Immediately prior to installation of IM application;

2. Immediately after installation of IM application;

3. Before and after changing configurations such as:
   - ✦ switch language to German;
   - ✦ disable all automatic media downloading;
   - ✦ enable auto startup;
   - ✦ change default background;
   - ✦ deactivation/logout from the IM application;

4. Immediately prior to removal of the IM application;

5. Immediately after removal of the IM application.

A communication between an attacker and a victim was emulated by sending a simple image file. A registry snapshot was made before and after the activity.

Local databases of the messengers were stored after the aforementioned activities for further research by SQLite DB Browse.

During the configuration changing experiments, many values of the registry and files were changed and modified. Table 1 contains the most significant changes for each type of operation.

All reports were stored in a plan text file and isolated for further investigation.

The final step of the research was analyzing reports and datasets. The search of required registry values was carried out by the standard application regedit.exe. Databases of IM applications were investigated using SQLite DB Browser for data and potential artefacts and messages stored on the computer. Files containing messages were transferred and we attempted to open them without access to the owner's account.

The users' and application activities were examined through file analysis NTUSER.DAT file by RegRipper v.2.8 application.

The experiment was repeated twice in order to be sure of consistent results.

2. Results

All reports and datasets were examined in this section. The finding for each application is provided below. Further details of registry keys and paths are listed in *Table 1.*

2.1. Telegram artefacts left after installation, file structure and database

The researcher was able to find the full path to the related IM application, installation date, version and user login who installed the application as highlighted by key (*Table 1*, no 1).

During the installation, folders were created that contain the database and file structure for the Telegram application.

Files of the IM can be found in the folder: \AppData\Roaming\Telegram Desktop\. The database of the Telegram is presented in the folder: %\tdata\D877F783D5D3EF8C. However, the database is stored as separated and encrypted files and not human-readable. Attempts to open the content of the database on using another Telegram account or read with SQLite DB were unsuccessful because the files are encrypted.

It is interesting to note that image or video files saved by a user during the communication can be found in the folder unencrypted and readable in the following folder: %UserName%\Downloads\Telegram Desktop.

During the installation, several folders and registry keys (*Table 1*, no 2) were created for interaction with the AI assistant Cortana.

2.2. Telegram configuration artefacts

Further evidence shows that language configuration added and modified the following file: \AppData\Roaming\Telegram Desktop \ tdata\settings0.

Recent changes are written to the log file AppData\Roaming\Telegram Desktop \log.txt. Unfortunately, the file is updating every time the application was restarted.

The following key (*Table 1*, no 3) was constantly modified after disabling the automatic download. Changing the startup mode of the application can be traced by detecting the following key: AppData\Roaming\Microsoft\Windows \StartMenu\Programs\Startup\ Telegram.lnk.

The application applies settings by modifying each dialog file in the folder after changing background: \tdata\D877F783D5D3EF8C\. However, deactivating the application has removed all message files from IM database and created the folder: tdata\D877F783D5D3EF8C1.

The last launch of Telegram can be found in the following registry key (*Table 1*, no 4). The value of the key LastAccessedTime is stored in hexadecimal or binary format and it is necessary to use a converter to translate them into a readable form. Keys value LoggedOnSAMUser

Table 1.

Registry and file information

No	Description
1	[HKLM\SOFTWARE\Microsoft\Windows\CurrentVersion\Installer\UserData\S–1–5–21–92284784–4191497677–2105538262–1001\Products\47A4A0DF1FC991646A19B825E007A0D6\ InstallProperties] “InstallLocation”=”C:\\Users\\user_a\\AppData\\Roaming\\Telegram Desktop\\” “InstallDate”=”20171017”
2	HKU\S–1–5–21–92284784–4191497677–2105538262–1001\Software\Microsoft\Windows\CurrentVersion\Search\Microsoft.Windows.Cortana_cw5n1h2txyewy\AppsConstraintIndex\LatestConstraintIndexFolder: “C:\Users\user_a\AppData\Local\Packages\Microsoft.Windows.Cortana_cw5n1h2txyewy\LocalState\ConstraintIndex\Apps_{8adcf8d1–d1f5–43e9–805d–af5466e37b69}”
3	HKU\S–1–5–21–92284784–4191497677–2105538262–1001\Software\Microsoft\Windows\CurrentVersion\Explorer\UserAssist\{CEBFF5CD–ACE2–4F4F–9178–9926F41749EA}\Count\HRZR_PGYFRFFVBA
4	[HKCU\Software\Microsoft\Windows\CurrentVersion\Search\RecentApps\{DC6BD851–959F–45DA–BD7B–87FD4EBF9648}] “AppId”=”C:\\Users\\user_a\\AppData\\Roaming\\Telegram Desktop\\Telegram.exe” “LastAccessedTime”=hex(b):20,b5,3a,31,bc,55,d3,01 “LaunchCount”=dword:00000015
5	[HKLM\SOFTWARE\Microsoft\Windows\CurrentVersion\Authentication\LogonUI\SessionData\1] “LoggedOnSAMUser”=”test_pc\\user_a” “LoggedOnUser”=” test_pc\\user_a”
6	[HU\S–1–5–21–92284784–4191497677–2105538262–1001\ Software\ Classes \ tg] “URL Protocol”=”” @=URL:Telegram Link
7	[HKU\S–1–5–21–92284784–4191497677–2105538262–1001\Software\Classes\tdesktop.tg\DefaultIcon] @=”\”C:\\Users\\user_a\\AppData\\Roaming\\Telegram Desktop\\Telegram.exe,1\””
8	[HKU\S–1–5–21–92284784–4191497677–2105538262–1001\Software\Microsoft\Windows NT\CurrentVersion\AppCompatFlags\Compatibility Assistant\Store] “C:\\Users\\user_a\\AppData\\Roaming\\Telegram Desktop\\unins000.exe”=hex:53,
9	HU\S–1–5–21–92284784–4191497677–2105538262–1001\ Software\ Classes \tdesktop.tg\DefaultIcon] @=”\”C:\\Users\\user_a\\AppData\\Roaming\\Telegram Desktop\\Telegram.exe,1\””
10	HLM\SOFTWARE\Microsoft\Windows\CurrentVersion\Installer\UserData\S–1–5–21–92284784–4191497677–2105538262–1001\Products\47A4A0DF1FC991646A19B825E007A0D6\InstallProperties “InstallDate”=”20171027” “DisplayVersion”=”6.9.6.16” “DisplayName”=”Viber”
11	HKU\S–1–5–21–92284784–4191497677–2105538262–1001\Software\Microsoft\Windows\CurrentVersion\Search\Microsoft.Windows.Cortana_cw5n1h2txyewy\AppsConstraintIndex\LatestConstraintIndexFolder: «C:\Users\user_a\AppData\Local\Packages\Microsoft. Windows.Cortana_cw5n1h2txyewy\LocalState\ConstraintIndex\Apps_{87f4a862–0157–4db6–927a–464474baefcd}»
12	HKU\S–1–5–21–92284784–4191497677–2105538262–1001\Software\Microsoft\Windows\CurrentVersion\Explorer\SessionInfo\1\ApplicationViewManagement\W32:00000000001104F8
13	HKU\S–1–5–21–92284784–4191497677–2105538262–1001\Software\ Microsoft\Windows\CurrentVersion\Run\Viber: “”C:\Users\user_a\AppData\Local\Viber\Viber.exe” StartMinimized”
14	%User%\AppData\Roaming\ViberPC\%phone№%\Backgrounds\3\10000403.jpg
15	[HKEY_CURRENT_USER\Software\Microsoft\Windows\CurrentVersion\Uninstall\{cbbefdcb–c7ee–4854–a1bc–c96d22b9d367}] “DisplayVersion”=”6.9.6.16” “Publisher”=”Viber Media Inc.”
16	[HKEY_CLASSES_ROOT\Local Settings\Software\Microsoft\Windows\CurrentVersion\AppModel\SystemAppData\Microsoft.Windows.Photos_8wekyb3d8bbwe\PersistedStorageItemTable\ManagedByApp\{1653CDC0–15E2–4885–A58A–E21C803F0BAA}] “Metadata”=”C:\\Users\\user_a\\AppData\\Roaming\\ViberPC\\447718905468\\Thumbnails\\thumb–c06ce8612230f51f80144f7077213b68.png” “LastUpdatedTime”=hex:04,a2,9e,1d,92,4d,d3,01
17	[HKEY_CLASSES_ROOT\Local Settings\Software\Microsoft\Windows\Shell\MuiCache] “C:\\Users\\user_a\\AppData\\Local\\Viber\\Viber.exe.FriendlyAppName”=”Viber”
18	[HKCU\Software\Microsoft\Windows\CurrentVersion\Search\RecentApps\{33D886D6–91BA–419C–A151–C9D0D31EEE34}] “LastAccessedTime”=hex(b):e0,b8,bb,3d,43,50,d3,01 “AppId”=”C:\\Users\\user_a\\AppData\\Local\\Viber\\Viber.exe”
19	Uninstall: Software\Microsoft\Windows\CurrentVersion\Uninstall Fri Oct 27 14:48:48 2017 (UTC) Viber v.6.9.6.16

and LoggedOnUser in the following registry key (*Table 1*, no 5) help to understand the details of application users.

2.3. Artefacts remaining after removal of Telegram

Despite difficulties with reading files containing messages and their removal after deactivating or uninstalling the application, many artefacts remain in the registry that researchers can find for understanding the directory structure and menu link as it is presented in the keys (Table 1, no 6, 7 and 8).

Some keys (*Table 1*, no 9) provide complete information about the installation path of the program despite removal.

It is interesting to note that a huge amount of useful information can be extracted from the NTUSER.DAT file. For example "most recently used" list, or "MRU report" from RegRipper shows the time of the last Telegram database record.

2.4. Viber artefacts left after installation, file structure and database

Examination of the registry determined that the following registry key (*Table 1*, no 10) was created by Windows specifying installation date and Viber version. The installer created different changes in the file structure and registry during the installation process, for example, interaction keys for AI Cortana (*Table 1*, no 11).

✦ The following folders contain most files of the Viber application:

✦ Database: %user%\AppData\Roaming\ViberPC\;

✦ Application: %user%\AppData\Local\Viber\;

QML caching: %user%\AppData\Local\Viber Media S.a r.l.

A folder named as a user phone number that contained the main database viber.db was created after installation and activation of the Viber.

The database is unencrypted and most messages and information are readable through the SQLite DB Browser. Nevertheless, messages are presented in an unstructured form, but the contacts table gives full information about names and phone numbers.

The messages can be opened in a user-friendly form by simply replacing the viber.db file on the PC with the installed Viber application. In this case, there is no way to respond and receive messages on behalf of the owner of the database, but a researcher has full access to the messages history.

2.5. Viber configuration changes artefacts

Changes of the language settings application modify the following file: %user%\AppData\Roaming\ViberPC\%phone% \QmlWebCache\data8\7\1tt95mf7.d.

Further evidence shows that all automatic media downloads have been disabled. This can be seen in the presence of a new registry key (Table 1, no 12). This value (*Table 1*, no 13) shows that a startup mode has been changed for the Viber application.

Changes of the default background for the application can be traced by adding a new file presented in *Table 1*, no 14. All content was deleted in the database folder \ViberPC\%phone№% after deactivation of the Viber account. However, the database file config.db containing settings of the application was available in the folder. The researcher can retrieve information about the phone number and previous IM account from the "Accounts" table of the config.db file using SQLite DB Browser.

2.6. Artefacts remaining after removal of Viber

The application left the key (*Table 1*, no 15) in the registry that provides information about de-installation of the program from Windows. The artefacts remaining in the registry allow us to restore a folder structure, location and history of file transfers via messenger (*Table 1*, no 16).

The HKEY_CLASSES_ROOT\viber branch record values were added by Viber installation and are still available in the system after removal. These keys (*Table 1*, no 17 and 18) specifying the path, last accessed time to the application remain in the registry after the application has been uninstalled. Moreover, Windows created a record in the NTUSER.DAT file that indicated the date and time when the IM application was uninstalled (*Table 1*, no 19).

3. Discussion

This study investigated Windows 10 for a location of Telegram and Viber artefacts. The results indicated that use of the messenger applications leaves registry artefacts which contain material that might be useful for investigation.

Even though Telegram is considered to be one of the most secure instant messengers, this study shows that useful material such as time-based artefacts and traces of user application on a hard drive and registry have remained.

Exploring Viber artefacts showed that the researcher is able to find very interesting information that helps to restore the whole history of a communication. Moreover, the study confirmed that artefacts are still available in Windows after removal of the application. Experts can unveil information about a user who installed the software and the account which used it.

In the future, research will include exploring system processes of the IM applications in Windows 10 for further deep forensic analysis of the IM behavior and cooperation with other system applications and software.

Conclusion

Messengers are popular today on mobile devices and traditional computers. Starting as a small text messaging service, they have turned into effective communication channels for both private and corporate users, becoming more than just an SMS replacement.

Users entrust to them a huge amount of information, such as a time-based map of activity, photos and other personal data. Messengers have changed the way communication is done; they reduce the distance to the user and along with social networks become tools for fraud, spam or blackmail and terrorism.

In this regard, it is vital to study IM from a forensic point of view. This research explores and compares two popular messengers: Viber and Telegram, which is rapidly gaining in popularity in the criminal world and the darknet as secure message tools. The main purpose of the research is to investigate and analyze potential artefacts remaining during the installation and use of instant messengers, as well as after their uninstallation.

The authors have done several experiments to investigate the artefacts in different environments, with clear explanation of the results. The experiments showed that even though Telegram is considered to be one of the most secure instant messengers, important useful material on a hard drive and registry have remained after complete uninstallation of the application.

Exploring Viber artefacts showed up information that helps to restore the whole history of communication. Moreover, the study confirmed that artefacts are still available in Windows after removal of the application. ■

Acknowledgments

This research was funded within the internal research of the Plekhanov Russian University of Economics under the name “Development of methods to forecast prices for financial instruments based on neural networks.”

References

1. The Radicati Group (2015) *Instant messaging market*, 2015–2019. Available at: http://www.radicati.com/wp/wp-content/uploads/2015/02/Instant-Messaging-Market-2014-2018-Executive-Summary.pdf (accessed 25 September 2018).
2. Roberts J.J. (2017) *Here are the most popular apps for secure messages*. Available at: http://fortune.com/2017/01/17/most-popular-secure-apps/ (accessed 27 September 2018).
3. Grispos G., Glisson W.B., Pardue H., Dickson M. (2014) Identifying user behavior from residual data in cloud-based synchronized apps. Proceedings of the *Conference on Information Systems Applied Research (CONISAR 2014), Baltimore, MD, USA, 6–9 November 2014*, no 3310. Available at: http://proc.conisar.org/2014/pdf/3310.pdf (accessed 27 September 2018).
4. Cheng L., van Dongen B.F., van der Aalst W.M.P. (2019) Scalable discovery of hybrid process models in a cloud computing environment. *IEEE Transactions on Services Computing* (Early Access Article). DOI: 10.1109/TSC.2019.2906203.
5. Levendoski M., Datar T., Rogers M. (2014) Yahoo! Messenger forensics on Windows Vista and Windows 7. *Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering*, vol. 88, pp. 172–179. DOI: 10.1007/978-3-642-35515-8_14.
6. Al Mutawa N., Al Awadhi I., Baggili I., Marrington A. (2011) Forensic artefacts of Facebook's instant messaging service. Proceedings of the *6th International Conference for Internet Technology and Secured Transactions (ICITST 2011), Abu Dhabi, United Arab Emirates, 11–14 December 2011*, pp. 771–776.
7. Yasin M., Abulaish M. (2014) DigLA – A Digsby log analysis tool to identify forensic artefacts, *Digital Investigation*, vol. 9, no 3–4, pp. 222–234. DOI: 10.1016/j.diin.2012.11.003.
8. Karpisek F., Baggili I., Breitinger F. (2015) WhatsApp network forensics: Decrypting and understanding the WhatsApp call signaling messages. *Digital Investigation*, vol. 15, pp. 110–118. DOI: 10.1016/j.diin.2015.09.002.
9. Majeed A., Zia H., Imran R., Saleem S. (2015) Forensic analysis of three social media apps in Windows 10. Proceedings of the *2015 12th International Conference on High-capacity Optical Networks and Enabling/Emerging Technologies (HONET), Islamabad, Pakistan, 21–23 December 2015*, pp. 1–5. DOI: 10.1109/HONET.2015.7395419.
10. Dehghantanha A., Choo K.-K.R., Muda Z. (2016) Windows instant messaging app forensics: Facebook and Skype as case studies. *PloS One*, vol. 11, no 3, pp. e0150300. DOI: 10.1371/journal.pone.0150300.
11. Cahyani N.D.W., Ab Rahman N.H., Glisson W.B., Choo K.-K.R. (2017) The role of mobile forensics in terrorism investigations involving the use of cloud storage service and communication apps. *Mobile Networks and Applications*, vol. 22, no 2, pp. 240–254. DOI: 10.1007/s11036-016-0791-8.

12 Carvey H., Hull D. (2014) *Windows registry forensics*. Elsevier. DOI: 10.1016/C2009-0-63856-3.

About the authors

Alexander I. Borodin

Dr. Sci. (Econ.);

Professor, Department of Financial Management, Plekhanov Russian University of Economics,
36, Stremyanny Lane, Moscow 117997, Russia;

E-mail: aib-2004@yandex.ru

ORCID: 0000-0002-2872-1008

Roman R. Veynberg
Cand. Sci. (Econ.);
Associate Professor, Department of Informatics, Plekhanov Russian University of Economics,
36, Stremyanny Lane, Moscow 117997, Russia;
E-mail: veynberg@gmail.com
ORCID: 0000-0001-8021-5738

Dmitry V. Pisarev
Master of Science in Cyber Security and Management;
University of Warwick, Coventry CV4 7AL, United Kingdom;
E-mail: d.pisarev@warwick.ac.uk

Oleg V. Litvishko
Cand. Sci. (Econ.);
Associate Professor, Department of Financial Management, Plekhanov Russian University of Economics,
36, Stremyanny Lane, Moscow 117997, Russia;
E-mail: Litvishko.OV@rea.ru

DOI: 10.17323/1998-0663.2019.4.49.59

Image processing of concentrated and scattered objects

Vladimir V. Alekseev
E-mail: vvalex1961@mail.ru

Denis V. Lakomov
E-mail: LaDenV@yandex.ru

Artem A. Shishkin
E-mail: 68region333@mail.ru

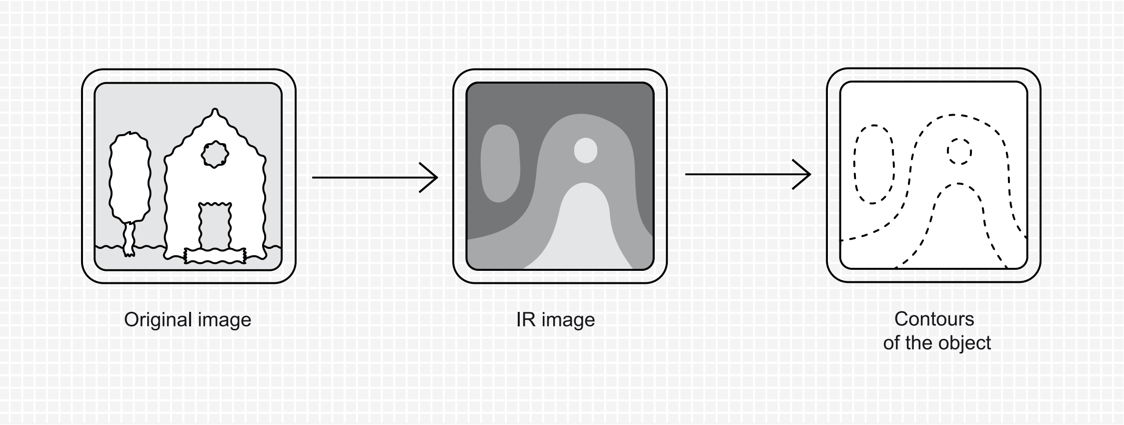
Ghassan Al Maamari
E-mail: ghassan.almaamari@gmail.com

Tambov State Technical University
Address: 106, Sovetskaya Street, Tambov 392000, Russia

Abstract

In modern control systems and information processing, the recognition of objects in the image is complicated by the fact that the impact of negative factors introduces uncertainty into this process, leading to blurring of images. In this regard, it is necessary to develop models and algorithms that would reduce the degree of uncertainty in image processing. These models are necessary, for example, when monitoring environmentally hazardous objects, for search and detection of unauthorized burial of household waste, in the field of information security, in the analysis of x-rays and thermograms, in the actions of unmanned aerial vehicles of law enforcement agencies in autonomous mode. This article presents a description of information technology for recognition in the automated mode of objects in images. The basis of this technology is the algorithm of contour analysis of images. The main distinguishing feature of the algorithm is the use of convolution of the image in four directions, as well as the tracing procedure. The aim of the study was to develop algorithms for high-speed automated visualization of external objects. We present the results of the study of the algorithm of contour analysis in the processing of various images in the visible and infrared wavelengths. Recommendations are formulated for the choice of parameters of the contour analysis algorithm, such as the mean square deviation in image blur, minimum and maximum thresholds for filtering. The results of the study can be used in production management systems, life support of the city, technical vision, environmental conditions, monitoring of business processes, as well as in the creation of simulators for training operators of complex systems, etc. In addition, we show the expediency of applying the algorithm we developed in decision support systems.

Graphical abstract



Key words: detection; image; picture; contour analysis; algorithm; operator; uncertainty; object.

Citation: Alekseev V.V., Lakomov D.V., Shishkin A.A., Al Maamari G. (2019) Image processing of concentrated and scattered objects. *Business Informatics*, vol. 13, no 4, pp. 49–59.
DOI: 10.17323/1998-0663.2019.4.49.59

Introduction

Amid the acceleration of technological progress, management issues have become major concerns in terms of science, economy and public development. At the same time, it is clear that introducing new technologies into various areas of society is hard to overestimate. The expansion of management functions and problems in different walks of science, technology, economy, and health care outlines a transition towards multi-purpose management systems. Managing and making decisions in such systems can be implemented, although they are accompanied by enormous uncertainty related to undetermined properties of a complex management object or process. It is also related to uncertainty that comes from an external object, uncertainty of interconnection between management subsystems, uncertainty of management objectives, uncertainty of quality criteria, etc.

The aforementioned aspects of system development and management processes, especially when it comes to complex management systems, make it necessary to improve existing techniques and develop new ones. These new techniques will serve as tools for object detection in images and integrity violation detection purposes as well. Alternatively, they can be used to find damage in an object during photo or video analysis. Principles and practices of decision support systems have proven that not only are they effective tools for analyzing and monitoring complex and large systems (including business flow); they can also successfully serve as software for scalar replacement. Hence, all around the world research is being conducted aimed at improving and developing new methods and algorithms for decreasing the influence of uncertainty during image analysis.

A subject area analysis has concluded that the contour analysis algorithm developed by the authors on the basis of the Canny algorithm can be applied in decision support systems. It can be used in such systems to verify the presence or absence of damage when objects, especially underground objects, are being monitored. That is why it is recommended to study the stages of the algorithm.

1. Stages of the contour analysis algorithm

A set of image pixels forms a matrix **P**:

$$\mathbf{P} = \begin{pmatrix} p_{11} & \cdots & p_{X1} \\ \vdots & \ddots & \vdots \\ p_{1Y} & \cdots & p_{XY} \end{pmatrix},$$

where p_{ij} – the color value of an image pixel;

X – the width of the image;

Y – the height of the image.

Let $\mathbf{P}_u$ be a set specified by all elements of the pixel matrix ($\mathbf{P} \to \mathbf{P}_u$).

Stage 1. Preliminary stage: converting to a grayscale image [1]. In order to convert an image to a grayscale image, the YUV model is needed. The model represents color in three variables: Y is the brightness component and U and V are the auxiliary components for color restoration. A transition from the RGB model to the YUV model is obtained by the following formula:

$$Y = 0.299 \cdot R + 0.587 \cdot G + 0.114 \cdot B.$$

$$U = -0.14713 \cdot R - 0.28886 \cdot G + 0.426 \cdot B.$$

$$Y = 0.615 \cdot R - 0.51499 \cdot G - 0.10001 \cdot B + 128,$$

where R, G, B – the intensities of red, green, and blue, respectively [2].

Stage 2. Noise reduction. Noise reduction is a necessary step to remove side effects which occurred after image blurring. The developed algorithm uses image smoothing. A Gaussian function is applied in the algorithm as a smoothing filter.

Stage 3. Calculation of variables and defining the gradient directions. The gradient values can be calculated in pixels using a four-way method (vertical, horizontal, two diagonal ways). As a result, a one-to-one correspondence between a pixel and its value of the gradient brightness (g_{ij}) is obtained.

The value of the gradient angle is rounded up in a way that the rounded-up value is multiple of 45, thus it may be equal to 0, 45, 90 or 135 degrees. The Robinson operator is used in the algorithm to compute the gradient value [3].

Stage 4. Pixel reduction using the gradient value that is different from the maximum value. Edge pixels are pixels that have the maximum gradient value (the local maximum) with respect to the neighboring pixels in accordance with the direction of the vector. Other pixels should be reduced. Their color characteristics in terms of RGB take the value of white color (255, 255, 255) [4]. Further work deals with edge pixels. Let **C** be a set of edge pixels:

$$\mathbf{C} = \{\{p_{ij}\} | p_{ij} \neq (255, 255, 255)\},$$
$$1 \le i \le X, 1 \le j \le Y,$$

All other pixels should be grouped in a set W:

$$\mathbf{W} = \{\{p_{ij}\} | p_{ij} = (255, 255, 255)\},$$
$$1 \le i \le X, 1 \le j \le Y,$$
$$\mathbf{C} \cup \mathbf{W} = \mathbf{P}_u \qquad (1)$$

Stage 5. Double detection of edges. In stage 4 all pixels, which are also the local maximums, must be verified to be consistent with two thresholds, $T_{\min}$ and $T_{\max}$. The threshold values are set by the user. If the gradient value of the pixel is lower than the minimum threshold, then the pixel belongs to set **W** and takes the white color value. Otherwise, if it is greater than the maximum threshold, then this pixel is contour and belongs to set **C**. All other pixels that are left in-between the thresholds are going to be processed at the final stage [5].

The following formula yields a representation of this stage:

$$\begin{cases} p_{ij} \in \mathbf{C},\ npu\ g_{ij} > T_{\max}, \\ p_{ij} \in \mathbf{W},\ npu\ g_{ij} < T_{\min}, \\ p_{ij} \in \mathbf{N},\ npu\ T_{\min} < g_{ij} < T_{\max}, \end{cases}$$

where **N** – a set of pixels, the gradient values of which are located in-between the thresholds.

Correspondingly, expression (1) takes the following form:

$$\mathbf{C} \cup \mathbf{W} \cup \mathbf{N} = \mathbf{P}_u \qquad (2)$$

Stage 6. Defining a stack of pixels between the threshold values. At the final stage, pixels belonging to set N are processed. The following condition must be satisfied for the pixels: if any neighboring pixel is an edge pixel, then the pixel that is being verified is considered an edge pixel as well. The main problem of this stage is to analyze all the pixels from set N in a way that groups of pixels which belong to the set and include edge pixels are also examined [6]. Pixels that don't turn out to be edge pixels take the value of white color. The distinctive characteristics of this stage are the following:

1. If the matrix element value of the current index is greater than the maximum threshold, then the pixel is applied to the stack of pixels. Hence, the vertex of the stack is then equal to 1;

2. The top element of the stack is removed. This entails subtracting 1 from its value;

3. Eight stack-derived indices, which are neighbors with respect to the index, are obtained;

4. The following condition for the neighboring indices must be verified: if the gradient value is greater than the minimum threshold and it hasn't been tagged 'considered,' then the index is tagged 'considered' and placed atop the stack;

Let us consider the most important stages in implementing the contour analysis algorithm.

2. Noise reduction

Image smoothing is the first step of the contour analysis algorithm. A Gaussian filter is applied to execute a procedure of smoothing, which blurs an image and allows us to reduce the influence of noise on the image analysis process. The main component of the Gaussian filter is an image kernel, also known as the convolution matrix, or mask. It is a matrix of odd size (3×3, 5×5, 7×7, etc) for finding the central pixel of the matrix, where the Gaussian function is also applied. The function is called a weight function. A weight multiplier is a value that corresponds to its element of the convolution matrix. The weight function is the sum of all weight multipliers of the convolution matrix [7].

Image blurring is achieved by moving the convolution matrix along the pixel matrix of the image. The image is thereby convoluted in every position of the window or, in other words, the computation of a new value of every pixel using the values of the neighboring pixels takes place. The weight function is constant throughout the convolution process. The computation is performed as follows: the value of the respective pixel is multiplied by the coefficient of the weight function; after that the obtained products are summed. The final value is assigned to the pixel which is placed in the center of the convolution matrix [8].

The Gaussian function is also used to calculate the coefficients of the weight function per each coefficient. Eventually, the coefficient value is then placed in the respective element of the convolution matrix:

$$f(x,y) = \frac{e^{-\frac{x^2+y^2}{2\sigma^2}}}{2\pi\sigma^2},$$

where x, y – the pixel's distance from the central pixel of the convolution matrix in the horizontal and vertical axis, respectively;

σ – the spread of the Gaussian function.

The values obtained for every pixel are then fixed in the respective elements of an intermediate matrix of the same size as the image. This additional intermediate matrix is necessary to avoid miscalculations that might be caused by the processed image pixels. It is important to mention threshold conditions. The conditions state that the corner pixels, unlike all other pixels, are deprived of any neighbors. To solve this problem, a temporary image should be created with the following size:

$$x+\frac{R+1}{2},\ y+\frac{R+1}{2},$$

where x, y – the length and heights of the image;

R – the size of the convolution matrix.

The procedure to calculate values of the image analyzed is as follows:

1) A temporary image is created. The empty image is overlaid by the initial picture, while its corners are overlaid by the corner pixels of the initial image;

2) The temporary image is blurred;

3) An image from the temporary one is retrieved. It is to be of the same size as the processed image with regards to the center;

4) The color of the central pixel is achieved by summing the weight coefficients of the convolution kernel and the values of colors of the neighboring pixels of the image.

The values of the processed image pixels change and should be in consistency with the neighboring values.

Image smoothing decreases the noise level by aligning the pixel values with the neighboring dots. At this point we should also specify valid values for σ. The results of [9–13] allow us to draw the following inferences:

1) If we deal with the range where $0{,}1 < \sigma < 1$, then smoothing will be insignificant because the obtained pixel values, besides the central ones, are small enough to affect their colors;

2) If $1 \le \sigma < 10$, then the neighboring pixel values will be adjusted to other pixels. The convolution matrix has been applied to the other pixels before, therefore a major part of the noise will be removed;

3) Two-digit values of σ enable blurring. This kind of blurring makes the noise and some pixels disappear from the image;

4) If σ is much smaller than 1, then there will be only the central pixel of the matrix that differs from zero. Blurring also will not occur.

3. Finding the gradient values

An operator for finding the gradient value in image pixels is the main feature of the contour analysis algorithm for concentrated and scattered objects. For further work, the direction and value of the gradient are required.

Let us consider a 3×3 matrix **D** containing the brightness values of the surroundings of the pixel:

$$\mathbf{D}=\begin{bmatrix} d_1 & d_2 & d_3 \\ d_4 & d_5 & d_6 \\ d_7 & d_8 & d_9 \end{bmatrix}.$$

The following formula yields the gradient value in every pixel in an image comprising concentrated or scattered objects [14]:

$$G=\sqrt{G_x^2+G_y^2},$$

where G – the gradient value in a pixel d_5;

G_x, G_y – approximate derivatives of the gradient value in terms of x and y, respectively (take forms of two matrices).

Let us review existing math operators.

The Roberts cross operator. This operator facilitates high-speed calculations, although it has one disadvantage, i.e. noise sensitivity. Contour lines that can be obtained using such operator are thinner than those obtained by other operators.

A matrix of approximate derivatives in accordance with the Roberts cross operator in d_5 for G_x and G_y can be computed in the following way:

$$G_x = d_9 - d_5,\ G_y = d_8 - d_6.$$

As the formula points out, the process of convolution can be performed in two possible ways, in a vertical or horizontal direction [15]:

$$Rb_0 = \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix},\ Rb_1 = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}.$$

The Prewitt operator. Unlike the previous one, this operator is based on the use of 3×3 masks and also considers 8 directions of the gradient. However, only the straight directions are used, since they bring about the best results:

$$p_0 = \begin{bmatrix} -1 & -1 & -1 \\ 0 & 0 & 0 \\ 1 & 1 & 1 \end{bmatrix},\ p_1 = \begin{bmatrix} -1 & 0 & 1 \\ -1 & 0 & 1 \\ -1 & 0 & 1 \end{bmatrix}.$$

The matrix of approximate derivatives in accordance with the Prewitt operator for G_x and G_y can be calculated as follows:

$$G_x = (d_7 + d_8 + d_9) - (d_1 + d_2 + d_3),$$

$$G_y = (d_3 + d_6 + d_9) - (d_1 + d_4 + d_7).$$

The primary benefit of the Prewitt operator is an ability to reduce sensitivity to noise by using an augmented mask [16]:

The Sobel operator. This operator works with values that are approximate to their derivatives. This aids in detecting contours in areas where the gradient is a large number. The operator comprises two matrices. The difference between the second and first matrices is that the former is rotated by 90 degrees, so is it in the Prewitt operator. In order to reduce the blurring effect, a weight coefficient that is equal to 2 is used for the matrix elements. It is achieved by raising the influence of middle pixels to avoid losing the contours of images.

This operator resembles the Roberts cross operator and the Prewitt operator. The masks of the Sobel operator have the following form:

$$s_0 = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix},\ s_1 = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix}.$$

The matrix of approximate derivatives in accordance with the Sobel operator for G_x and G_y can be calculated as follows:

$$G_x = (d_7 + 2\,d_8 + d_9) - (d_1 + 2\,d_2 + d_3),$$

$$G_y = (d_3 + 2\,d_6 + d_9) - (d_1 + 2\,d_4 + d_7).$$

The fact that the Sobel operator has a rough approximation of the gradient, which leads to a great deal of contours being lost, serves as a drawback [17].

The Kirsch and Robinson operators. These operators consist of 8 symmetric masks. For instance, the center of the symmetry of the Robinson operator is a central axis containing only zeros. It takes just computing the values of the first four masks, meanwhile the others can be obtained by inversing the first ones. The operators are quite similar; however they differ in their complexities. The Kirsch operator uses weight coefficients, such as –3 and 5, while the Robinson operator deals with –2, –1, 0, 1 and 2.

The masks of the Kirsch operator are as follows:

$$k_0 = \begin{bmatrix} -3 & -3 & 5 \\ -3 & 0 & 5 \\ -3 & -3 & 5 \end{bmatrix},\ k_1 = \begin{bmatrix} -3 & 5 & 5 \\ -3 & 0 & 5 \\ -3 & -3 & -3 \end{bmatrix},$$

$$k_2 = \begin{bmatrix} 5 & 5 & 5 \\ -3 & 0 & -3 \\ -3 & -3 & -3 \end{bmatrix},\ k_3 = \begin{bmatrix} 5 & 5 & -3 \\ 5 & 0 & -3 \\ -3 & -3 & -3 \end{bmatrix},$$

$$k_4 = \begin{bmatrix} 5 & -3 & -3 \\ 5 & 0 & -3 \\ 5 & -3 & -3 \end{bmatrix},\ k_5 = \begin{bmatrix} -3 & -3 & -3 \\ 5 & 0 & -3 \\ 5 & 5 & -3 \end{bmatrix},$$

$$k_6 = \begin{bmatrix} -3 & -3 & -3 \\ -3 & 0 & -3 \\ 5 & 5 & 5 \end{bmatrix},\ k_7 = \begin{bmatrix} -3 & -3 & -3 \\ -3 & 0 & 5 \\ -3 & 5 & 5 \end{bmatrix}.$$

Masks k_2 and k_6 correspond to the horizontal edge, masks k_0 and k_4 correspond to the vertical one, masks k_1, k_3, k_5 and k_7 correspond to the diagonal edge.

The main mask is the one that assists with obtaining the maximum value of d_5. This value is the gradient value of the pixel [18].

The Kirsch and Robinson operators are based on using just one mask which is rotated in the 8 compass directions: North, North West, West, South West, South, South East, East, North East. The masks of the Robinson operator have the following forms:

$$r_0 = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix}, \; r_1 = \begin{bmatrix} 0 & 1 & 2 \\ -1 & 0 & 1 \\ -2 & -1 & 0 \end{bmatrix},$$

$$r_2 = \begin{bmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix}, \; r_3 = \begin{bmatrix} 2 & 1 & 0 \\ 1 & 0 & -1 \\ 0 & -1 & -2 \end{bmatrix},$$

$$r_4 = \begin{bmatrix} 1 & 0 & -1 \\ 2 & 0 & -2 \\ 1 & 0 & -1 \end{bmatrix}, \; r_5 = \begin{bmatrix} 0 & -1 & -2 \\ 1 & 0 & -1 \\ 2 & 1 & 0 \end{bmatrix},$$

$$r_6 = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix}, \; r_7 = \begin{bmatrix} -2 & -1 & 0 \\ -1 & 0 & 1 \\ 0 & 1 & 2 \end{bmatrix}.$$

The gradient value in each pixel is rounded up and then set to the maximum value, which is calculated by using the masks. The angle between the rows of zeros in the mask turns out to be the angle of the gradient vector. The mask allows us to obtain the maximum gradient value [19].

The Laplace operator. It was first published in 1982. This algorithm facilitates the calculation of the second derivative using the formula:

$$\Delta^2 f = \frac{d^2 f}{dx^2} + \frac{d^2 f}{dy^2}.$$

The operator is performed in two stages. In the first one, image smoothing takes place using the Gaussian filter. The second one involves finding the Laplace operator, which results in the appearance of double contours. The contours can be detected by approximating their values. Such approximation implies finding zeros at the point of intersection of the double contours. The masks of the Laplace operator have the following form:

$$l_0 = \begin{bmatrix} 0 & -1 & 0 \\ -1 & 4 & -1 \\ 0 & -1 & 0 \end{bmatrix}, \; l_1 = \begin{bmatrix} -1 & -1 & -1 \\ -1 & 8 & -1 \\ -1 & -1 & -1 \end{bmatrix}.$$

The primary feature of the Laplace operator is the ability to find every pixel within or outside the double contour obtained by the Gaussian filter and Laplace operator [20].

Reference [21] incorporates the results of using all the aforementioned operators for analyzing an image. The authors have conducted an in-depth analysis of image processing results. The results indicate that the Prewitt, Robinson, Sobel operators bring out different results, yet they also complement one another. The Laplacian of Gaussian has demonstrated almost identical results for main contours, although false objects appeared as well, such as walls and a lawn in front of a building. In addition, the Kirsch and Robinson operators, unlike the others, have detected the greatest number of true contours in an image [4, 12, 14].

4. Test runs of the contour analysis algorithm

The suggested information system for image processing of concentrated and scattered objects incorporates an operating system, database of images, the contour analysis algorithm, a linguistic and logical model to choose the algorithm parameters, image processing results comparison model, and cross-platform software to ensure interpretation of results.

A series of test runs of the software has been carried out.

The Robinson operator was used to determine the gradient. The values are $T_{\min}$ = 20,

T_{max} = 45. These values (stage 5) enable us to obtain the best results in terms of the signal-noise ratio. *Figures 1–3* depict the results of image processing on the basis of the algorithm so developed.

The conducted research shown in *Figures 4–5* demonstrates that the contour analysis algorithm shows a 10–15% increase in detecting edge pixels compared to the standard Canny algorithm when it comes to image processing of concentrated and scattered objects. It is worth noting that the time needed to execute the algorithm has not increased dramatically. The number of test-run materials is 110 images.

5. Recommendations on choosing the algorithm parameters

The algorithm software allows us to detect objects in blurred or sharply-defined images and recognize them easily. The effectiveness of such research may be affected by the quality of an image captured by an IR device, distance range, camera angle, weather conditions. Unwanted sources of heat, i.e. residential buildings, technical gadgets, living beings, may hamper the analysis of an image.

The recommended parameters for an anal-

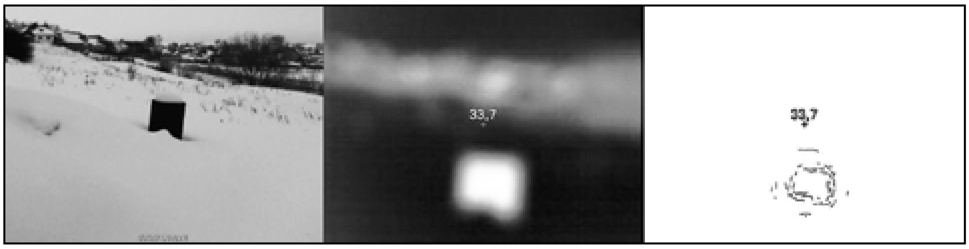


Fig. 1. Image processing of a vent unit

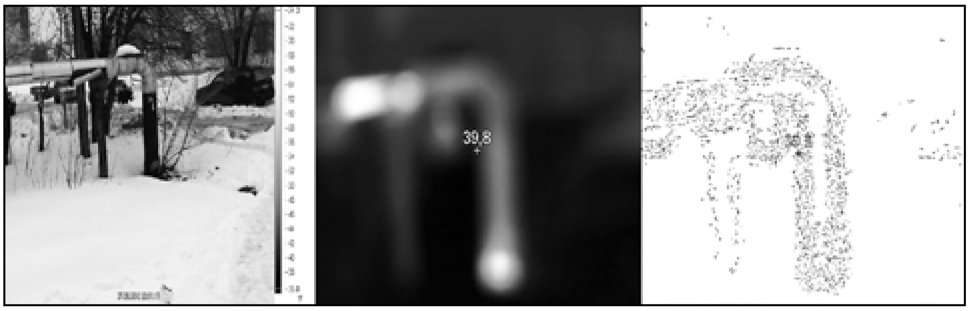


Fig. 2. Image processing of a heat pipeline

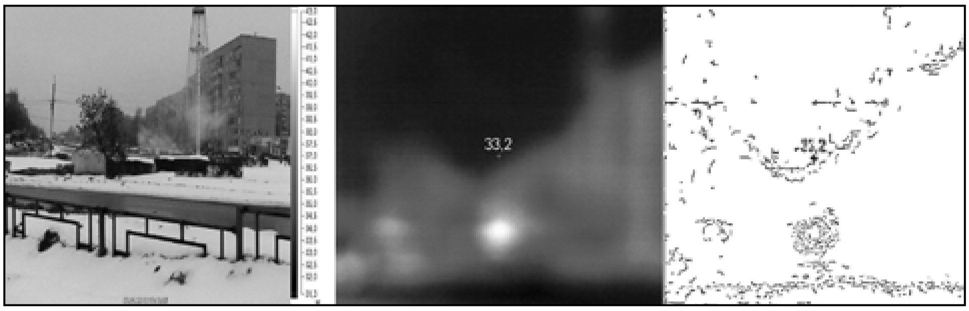


Fig. 3. Image processing of a manhole

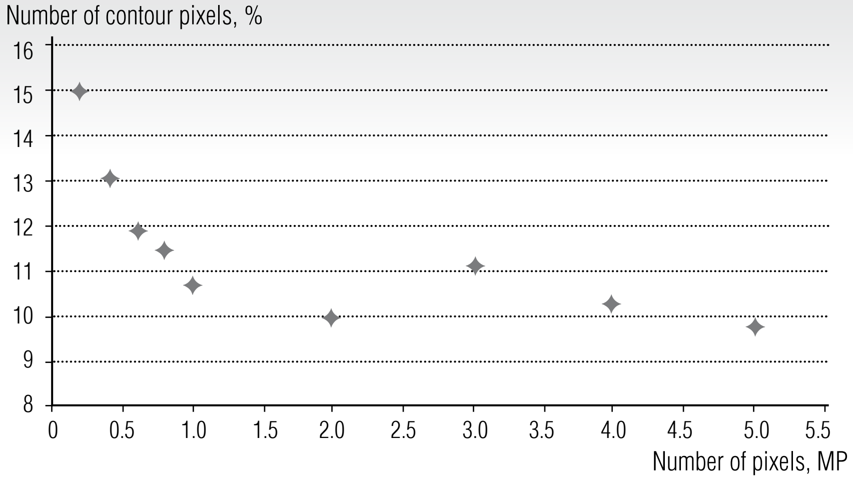


Fig. 4. The increase of the number of detected pixels if the algorithm is used in comparison with the Canny algorithm

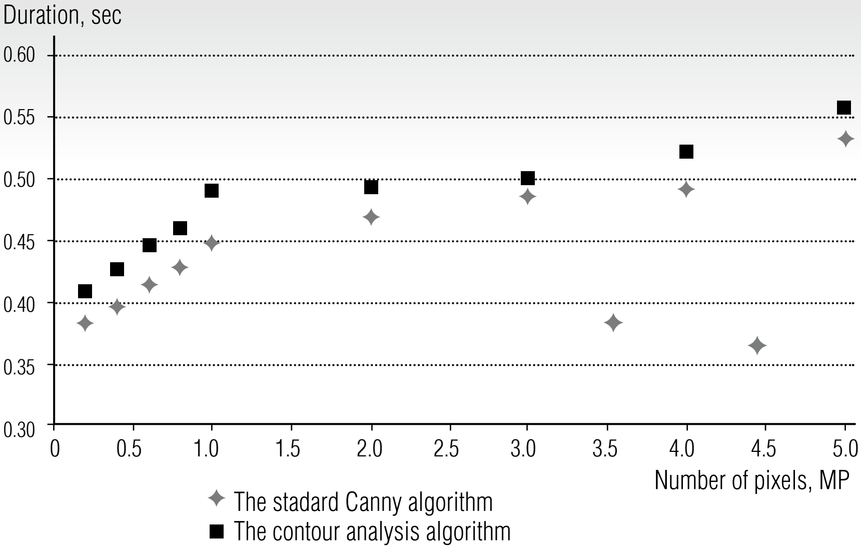


Fig. 5. The elapsed time of the algorithms to process images

ysis of a sharply-defined image are $T_{min} = 20$, $T_{max} = 45$, $\sigma = 1$.

The recommended parameters for an analysis of a blurred image are $T_{min} = 5$, $T_{max} = 15$, $\sigma = 0.01$.

The recommended parameters are applicable for the considered and similar types of images. The linguistic and logical model is being developed to choose the algorithm parameters that depend on image specifications.

The advantages of the algorithm are the following:

✦ an increase in the number of detected edge pixels;

✦ minimization of contour fragmentation;

✦ a decrease of image-noise-based uncertainty by virtue of the Gaussian filter.

The lack of well-defined criteria of choosing the threshold values which leads to the distortion of true contours and the appearance of false ones is the main drawback of the algorithm.

The research results suggest that the algorithm software should be improved to make the image corners less rounded since this causes the edges to be removed.

Conclusion

This paper showcases the description of an image processing technology for concentrated and scattered objects on the basis of the contour analysis algorithm.

The algorithm operates using the Robinson operator and the process of convolution applied to an object in four directions. It ensures that contour fragmentation is minimized.

We have provided the research results of the contour analysis algorithm for image processing of different objects in the infrared range. Recommendations on choosing the algorithm parameters are suggested on the basis of the research.

The research results may find application in technical vision systems designed to detect damage and monitor analyzed objects. Additionally, sharply-defined and scattered objects monitoring systems may see a benefit of the algorithm by training new operators. Furthermore, it can be highly useful for decision support systems and detecting integrity violations in sharply-defined and scattered objects. ■

References

1. Huo X.Q., Zheng W.L., Lu B.L. (2016) Driving fatigue detection with fusion of EEG and forehead EOG. Proceedings of the *International Joint Conference on Neural Networks (IJCNN), Vancouver, BC, Canada, 24–29 July 2016*, pp. 897–904. DOI: 10.1109/IJCNN.2016.7727294.
2. Alekseev V.V., Gromov Yu.Yu., Gubskov Yu.A., Ishchuk I.N. (2014) *Methodology of remote assessment of spatial distributions of optical-thermal parameters of the objects disguised under the surface.* Moscow: Nauchtehlitizdat (in Russian).
3. Karasev P.I., Gubskov Yu.A. (2015) Processing of graphic images in video surveillance systems. *Herald of the Voronezh Institute FSIN of Russia*, no 2, pp. 35–37 (in Russian).
4. Alekseev V.V., Lakomov D.V. (2017) Robinson Operator and its application in the canny algorithm for image recognition under uncertainty. Proceedings of the *XV All-Russian Scientific Conference “Neurocomputers and Their Application”, 14 March 2017.* Moscow: MSUPE, pp. 89-90 (in Russian).
5. Gonzalez R., Woods R. (2005) *Digital image processing.* Moscow: M: Technosphere (in Russian).
6. Pytyev Yu.P. (2010) *Methods of morphological image analysis.* Moscow: FIZMATLIT (in Russian).
7. Potapov A.A. (2008) *The latest methods of image processing.* Moscow: FIZMATLIT (in Russian).
8. Srivastava S., Delp E.J. (2003) Video-based real-time surveillance of vehicles. *Journal of Electronic Imaging*, vol. 22, no 4, 041103. DOI: 10.1117/1.JEI.22.4.041103.
9. Furman Ya.A. (2003) *Introduction to contour analysis. Applications to image and signal processing.* Moscow: FIZMATLIT (in Russian).
10. Basarab M.A., Volosyuk V.K., Goryachkin O.V., Zelensky A.A. (2007) *Digital signal and image processing in radiophysical applications.* Moscow: FIZMATLIT (in Russian).
11. Canny J. (1986) A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-8, no 6, pp. 679–698.
12. Kim N.V., Krylov I.G. (2012) Group application of unmanned aerial vehicle in surveillance tasks. *Trudy MAI*, no 62. Available at: https://mai.ru/upload/iblock/bbb/gruppovoe-primenenie-bespilotnogo-letatelnogo-apparata-v-zadachakh-nablyudeniya.pdf (accessed 15 April 2019) (in Russian).
13. Alekseev V.V., Karasev P.I., Lakomov D.V. (2016) Analysis of image processing methods applicable in the conditions of uncertainty. Proceedings of the *XVI International Conference “Informatics: problems, methodology, technologies,” Voronezh, 11–12 February 2016*, pp. 37–41 (in Russian).

14. Alekseev V.V., Lakomov D.V. (2016) Analysis of the applicability of blur in image recognition under uncertainty. Proceedings of the *III International Scientific and Practical Conference "Virtual modeling, prototyping and industrial design," Tambov, Russia, 17–19 November 2015*, vol. 2, pp. 138–141 (in Russian).

15. Kim N.V., Kuznetsov A.G., Krylov I.G. (2010) Application of vision systems on unmanned aerial vehicles in the tasks of orientation on the ground. *Aerospace MAI Journal*, vol. 17, no 3, pp. 46–49 (in Russian).

16. Kanishka Madusanka D.G., Gopura R.A.R.C., Amarasinghe Y.W.R.; Mann G.K.I. (2017) Hybrid vision based reach-to-grasp task planning method for trans-humeral prostheses. *IEEE Access*, vol. 5, no 99, pp. 16149 –16161. DOI: 10.1109/access.2017.2727502.

17. Jain A., Abbas B., Farooq O., Garg S.K. (2016) Fatigue detection and estimation using auto-regression analysis in EEG. Proceedings of the *2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Jaipur, India, 21–24 September 2016*, pp. 1092–1095. DOI: 10.1109/ICACCI.2016.7732190.

18. Yang *K*.F., Li C.-Y., Li Y.-J. (2014) Multifeature-based surround inhibition improves contour detection in natural images. *IEEE Transaction on Image Processing*, vol. 23, no 12, pp. 5020–5032. DOI: 10.1109/TIP.2014.2361210.

19. Yang K.-F., Gao S.-B., Guo C.-F., Li C.-Y., Li Y.-J. (2015) Boundary detection using double-opponency and spatial sparseness constraint. *IEEE Transaction on Image Processing*, vol. 24, no 8. pp. 2565–2578. DOI: 10.1109/TIP.2015.2425538.

20. Muthukrishnan R., Radha M. (2012) Edge detection techniques for image segmentation. *International Journal of Computer Science & Information Technology*, vol. 3, no 6, pp. 259–267. DOI: 10.5121/ijcsit.2011.3620.

21. Guan T., Wang Y., Duan L., Ji R. (2015) On-device mobile landmark recognition using binarized descriptor and multifeature fusion. *ACM Transactions on Intelligence Systems Technology*, vol. 7, no 1, article 12. DOI: 10.1145/2795234.

About the authors

Vladimir V. Alekseev

Dr. Sci. (Tech.), Professor;

Head of the Department "Information systems and information security",
Tambov State Technical University,
106, Sovetskaya Street, Tambov 392000, Russia;

E-mail: vvalex1961@mail.ru

Denis V. Lakomov

Doctoral Student, Department "Information systems and information security",
Tambov State Technical University,
106, Sovetskaya Street, Tambov 392000, Russia;

E-mail: LaDenV@yandex.ru

Artem A. Shishkin

Doctoral Student, Department "Information systems and information security",
Tambov State Technical University,
106, Sovetskaya Street, Tambov 392000, Russia;

E-mail: 68region333@mail.ru

Ghassan Al Maamari

Doctoral Student, Department "Information systems and information security",
Tambov State Technical University,
106, Sovetskaya Street, Tambov 392000, Russia;

E-mail: ghassan.almaamari@gmail.com

DOI: 10.17323/1998-0663.2019.4.60.72

The design of the structure of the software system for processing text document corpus

Vladimir B. Barakhnin [a,b]
E-mail: bar@ict.nsc.ru

Olga Yu. Kozhemyakina [a]
E-mail: olgakozhemyakina@mail.ru

Ravil I. Mukhamediev [c,d,e]
E-mail: ravil.muhamedyev@gmail.com

Yulia S. Borzilova [a]
E-mail: i.borzilova@alumni.nsu.ru

Kirill O. Yakunin [c,d]
E-mail: yakunin.k@mail.ru

[a] Institute of Computational Technologies, Siberian Branch of the Russian Academy of Sciences
Address: 6, Academician M.A. Lavrentiev Avenue, Novosibirsk 630090, Russia

[b] Novosibirsk State University
Address: 1, Pirogova Street, Novosibirsk 630090, Russia

[c] Satbayev University
Address: 22a, Satbayev Street, Almaty 050013, Kazakhstan

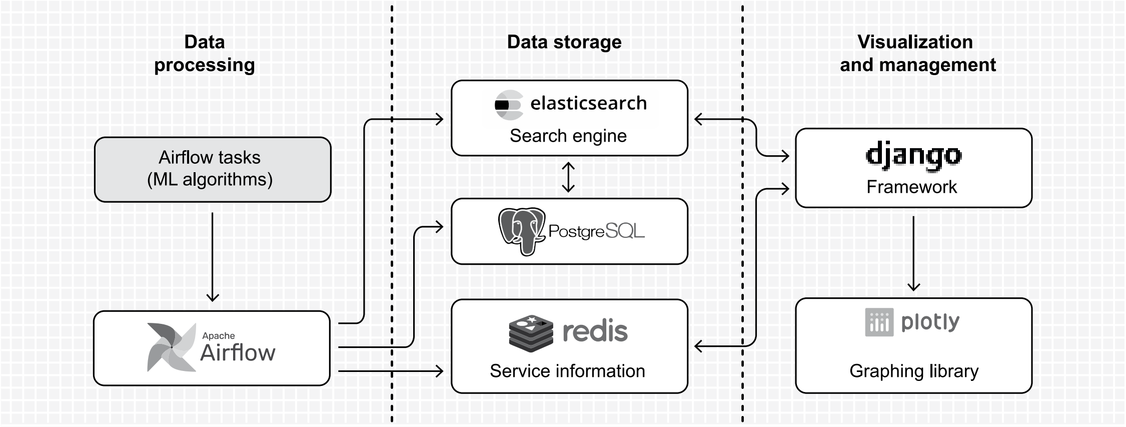
[d] Institute of Information and Computational Technologies
Address: 125, Pushkin Street, Almaty 050010, Kazakhstan

[e] ISMA University
Address: 1, Lomonosova Street, Riga LV-1019, Latvia

Abstract

One of the most difficult tasks in the field of data mining is the development of universal tools for the analysis of texts written in the literary and business styles. A popular path in the development of algorithms for processing text document corpus is the use of machine learning methods that allow one to solve NLP (natural language processing) tasks. The basis for research in the field of natural language processing is to be found in the following factors: the specificity of the structure of literary and business style texts (all of which requires the formation of separate datasets and, in the case of machine learning methods, the additional feature selection) and the lack of complete systems of mass processing of text documents for the Russian language (in relation to the scientific community-in the commercial environment, there are some systems of smaller scale, which are solving highly specialized tasks, for example, the definition of the tonality of the text). The aim of the current study is to design and further develop the structure of a text document corpus processing system. The design took into account the requirements for large-scale systems: modularity, the ability to scale components, the conditional independence of components. The system we designed is a set of components, each of which is formed and used in the form of Docker-containers. The levels of the system are: the data processing level, the data storage level, the visualization and management of the results of data processing (visualization and management level). At the data processing level, the text documents (for example, news events) are collected (scrapped) and further processed using an ensemble of machine learning methods, each of which is implemented in the system as a separate Airflow-task. The results are placed for storage in a relational database; ElasticSearch is used to increase the speed of data search (more than 1 million units). The visualization of statistics which is obtained as a result of the algorithms is carried out using the Plotly plugin. The administration and the viewing of processed texts are available through a web-interface using the Django framework. The general scheme of the interaction of components is organized on the principle of ETL (extract, transform, load). Currently the system is used to analyze the corpus of news texts in order to identify information of a destructive nature. In the future, we expect to improve the system and to publish the components in the open repository GitHub for access by the scientific community.

Graphical abstract



Key words: natural language processing; streaming word processing; text analysis information system; development of a text corpus processing system.

Citation: Barakhnin V.B., Kozhemyakina O.Yu., Mukhamediev R.I., Borzilova Yu.S., Yakunin K.O. (2019) The design of the structure of the software system for processing text document corpus. *Business Informatics*, vol. 13, no 4, pp. 60–72. DOI: 10.17323/1998-0663.2019.4.60.72.

Introduction

Modern methods of data mining allow us to process a large corpus of text documents (with a volume of more than one million documents) in order to identify certain properties of individual documents included in the corpus, as well as to identify rules characterizing their combination. Since these algorithms involve extracting a wide range of diverse characteristics from texts (which is often a complex task in itself, involving the use of complex but not always high-speed algorithms), it becomes necessary to store the extracted characteristics (along with the documents themselves) in a reference-and-information fund of the software system created. At the same time, the information model of the document repository and their characteristics will largely depend on the type of research text corpus and the nature of the tasks to be solved. For example, systems for processing news messages in order to identify destructive information [1] are significantly different from the systems for processing scientific information [2] and, especially, literary texts (prose and poetry) [3].

It should be noted that one of the difficult tasks is the development of universal tools for the analysis of texts of literary and business styles. As indicated in [4], “when the words have been recognized in a business text, the most significant factor is familiarity with the text (its subject, structure and most frequent words); the keywords and theme elements are recognized relatively well; the end of the text is predictable and well recognized. For a literary text, a large “accent” falls on the initial (preamble) and middle (plot development) compositional fragments and in different ways relates to the components of communicative and semantic division: with the topic for the preamble, with dialogue (especially keywords or rema) for the middle fragment. Thus, speaking about text structures and analysis procedures, we must take into account various types of context, in particular, the functional style, compositional structure and rhetorical connectivity of the text” [4].

Currently, text processing is an actively developing field of IT. A review of works in this area is available, for example, in [5–8]. Note that in the last decade, the main direction of the development of algorithms for processing text documents has been the use of machine learning methods (see, for example, [9–11]). In general, the following approaches can be classified for automatic text analysis [12]:

1. Rule-based with patterns. This approach uses such tools as part-of-speech-taggers and parsers. Another option is to use N-grams to define the frequently used combinations that merged into words. In particular, when solving problems of text sentiment analysis, these N-grams are assigned positive or negative estimates;

2. Unsupervised learning. The main difference from supervised learning is the lack of manual markup for model training. In the case of a statistical model of the corpus of texts, the most weight in the text belongs to such terms that are more often found in this text and at the same time are found in a small number of texts of all sets;

3. Supervised learning. The training set is manually marked up by qualified experts or dataset engineers. Then, the marked-up set is used to train various classifiers, among which the Naive Bayes Classifier [13], Support Vector Classifier (SVM) [14], as well as algorithm ensemble, for example, boosting [15], when several machine learning methods can be combined into an ensemble, in which each subsequent method is trained on the errors of the previous one, and artificial neural networks (ANN) of various configurations [16, 17];

4. Hybrid method. This approach can combine machine learning methods, as well as use rule templates;

5. A method based on graph-theoretic mod-

els. With this approach, the division of the corpus into words is used, with each word having its own weight. Such a weight is used, for example, in problems of sentiment analysis: some words have more weight and more strongly affect the sentiment of the text;

6. Pre-trained models based on deep neural networks (transfer learning), when a pretrained model is retrained to solve specific problems, for example, the very popular BERT model [18].

In particular, the task of sentiment analysis has been repeatedly solved and is actively used in commercial developments. The latter include, for example, the system of linguistic text analysis of the modular type Eureka Engine, which allows us to extract new knowledge and facts from unstructured data of large volumes[1]. In addition to sentiment analysis, the system solves the problem of the definition of the subject of the text (i.e. classification) and named-entity recognition (NER). The module for automatic classification of texts TextClassifier is implemented on machine learning; there are also modules for automatic determination of named entities, normalization of words, and a morphoanalyzer. The internal structure of the system is not given. The system was used as a tool for sentiment analysis in the media regarding the same event [19]. We can also note work [20], which presents the results of a study of the method of sentiment analysis using the analysis of Twitter messages and reviews of the Kinopoisk portal as an example. The authors used machine learning algorithms as a toolkit: the SVM, the Naive Bayes classifier, and random forest methods. Additionally, the work is providing an overview of similar works in sentiment analysis problems.

A variety of algorithms for NLP suggests the possibility of their implementation in the form of an independent software product. Due to this, the structure of the created software system should be aimed at its interaction with both the end user and the other systems. This article formulates the requirements for the structure of the created software system and defines the role functions of users. After this, the structure we developed is described; this includes a data processing subsystem, a storage and a subsystem for constructing analytical reports.

The main features of the developed system, distinguishing it from comparable systems, are:

1. Automatic thematic modeling, which allows us to identify trends in real time without manually generating a list of keywords or queries. This allows us to automatically identify relevant and socially significant topics online, all of which is critical in making managerial decisions in various fields;

2. Expert marking at the subject level allows us to reduce the volume of objects required for marking (by comparison with use of deep learning networks);

3. From the previous paragraph it follows that prompt and relatively inexpensive markup is possible according to an arbitrary set of criteria, not limited only by semantics. Criteria can be selected individually for the specific requirements of the client (for example, assessment of innovativeness, social significance, opposition, social trust, inflation expectations, etc.).

1. Statement of the problem

The process of text analysis in natural language is described according to the following steps, which analyze the characteristics of the text:

✦ initialization — the formation of the text corpus and its preprocessing for subsequent analysis;

[1] Eureka Engine: http://eurekaengine.ru/

✦ structural analysis (only for poetic texts) – determination of the low-level characteristics of the text (phonetics and metrorhythmics of the poem);

✦ semantic analysis – the definition of semantic constructions taking into account synonymy and named entity linking (NEL). Analysis of scientific texts is usually limited to this level;

✦ pragmatic analysis – definition of genre and style features for literary texts; constructions that determine the destructive impact for news messages, etc.

✦ synthesis of the obtained results – determination of the effect of lower levels on higher, as well as aggregation of results in a convenient form for perception and search.

Let’s formulate requirements for the functionality of the system based on its purpose: scraping, storage, streaming analytics and the formation of analytical reports with visualization.

1. Reliable storage of texts corpus of large volumes, while the system must be configured to work with multi-style texts: scientific, journalistic and artistic;

2. Fast parallel access, filtering and aggregation of data for stream processing: preprocessing, building thematic models and classifiers, aggregation and uploading for real-time reports, etc.;

3. Flexibility of the system and the ability to store unstructured and weakly structured data to support storage and access arbitrary data structures for statistical analysis and various computational experiments based on modern text analysis methods.

The structure of the software system should allow us to solve large-scale problems consisting in storing corpus of volumes of several million texts and batch processing online of several thousand documents. Such, for example, is a real-time monitoring project of the Russian-language media of the Republic of Kazakhstan [21] designed to create the following types of reports:

1. The thematic structure of news publications in Kazakhstan Republic electronic media both at the level of major topics (economics, education, politics) and subtopics (pre-school education, a single state exam, higher education and science), and at the level of informational occasions (specific narrow topics that describe a specific event or group of closely related events);

2. Evaluation of individual publications, topics and the media on an arbitrary set of criteria. Such an assessment involves preliminary marking by an expert or a panel of experts;

3. Reports and alerts for identified anomalies. The anomalies are considered at two levels: at the level of dynamics (for example, a sharp increase in publications on a certain topic or a sharp increase in publications with a negative assessment according to the “semantic” criterion) and at the thematic level – the emergence of groups of publications with non-standard, “anomalous” topics that were not earlier met with (for example, the theme of cryptocurrencies, the theme of feminism in Kazakhstan Republic, etc.).

The first two types of reports can be obtained both dynamically and statically (for example, media assessment by certain criteria over the past year). The anomaly report involves an analysis of the dynamics with reference to publication time.

Conceptual design included the formation of the capabilities of the created software system. The created software system should have the following features:

1. Providing access to the texts corpus;

2. Automated processing of the texts corpus stored in the database;

3. Input the characteristics obtained into the repository (database);

4. Flexible planning for various data processing tasks;

5. Statistical processing of the characteristics obtained and their presentation in user-friendly form for the researcher;

6. Updating and improving the algorithms used to analyze the texts corpus.

In the current study the task was set to design the structure of a system for processing natural text corpuses. The scope of this system begins with the analysis of text corpus of a journalistic style. In the future, the scope of the system can be extended to literary texts, due to the modularity of the system and the flexibility of the technologies used.

The designed system consists of the following subsystems:

1. Data processing subsystem. A combination of hybrid methods is used (supervised learning and dictionaries);

2. Data store. To ensure quick user interaction, as well as reduce resource consumption, several types of storages are used;

3. Subsystem for building analytics based on the data obtained.

The information system should take into account the stages of text analysis. The structure of the system consists of the components listed in the description of the problem statement. At the preprocessing stage, the text is pre-processed for further analysis. The preprocessing methods used depend on the algorithm that works with this data (training and analysis). The following types of processing can be classified:

✦ giving as a result of "bag of words"; this type also includes the TF-IDF method;

✦ giving as a result of processing each semantic unit of the corpus (for example, news) its embedding, for example, distribution by tokens / words / phrases / sentences. In this case, it is possible to use recurrent neural networks (RNN);

✦ giving as a result of processing each semantic unit of the corpus one text embedding; for such preprocessing, standard classification methods may be used.

Structural analysis is used for literary style texts and can be performed by currently existing tools, for example, [22]. Semantic analysis can be performed both at the stage of text preprocessing (for example, lemmatization of words), and may not be performed at all — the chosen toolkit will depend on the methods of machine learning and may change over time. Pragmatic analysis in the system is carried out using a combination of machine learning algorithms and compiled frequency dictionaries. The synthesis of the results is ensured by aggregating the results in some storages and outputting these results in a form that will most accurately satisfy the needs of the user.

Based on the capabilities of the system described above, the following requirements for the developed system can be distinguished:

to ensure the operation of subsystems in the form of separate independent components, each of which can be quickly replaced if necessary;

✧ to organize parallelization of calculations, including on several machines;

✧ to implement automated processing of the texts corpus at the request of the user;

✧ monitor tasks in real time, including providing timely reporting of exceptions;

✧ to display data from the analysis of texts in the user interface;

✧ to update the algorithms used in the system to improve the quality of analysis and expand their scope.

2. The system's structure

All components of the system are organized in the Docker containers. All the containers have access to one virtual network, which provides the ability to exchange data using standard network protocols (TCP). Such an implementation ensures the operation of subsystems in the form of independent components, each of which can be replaced if necessary.

The interaction of the components, the subsystem for building analytics and the subsystem for data processing, is carried out using a storage system. The general scheme of the interaction of components is organized on the principle of ETL (extract, transform, load): the user receives a request for data in ElasticSearch (if data is rarely used) or in Redis (if data is often used). In addition, the processing subsystem uses the Airflow scheduler, which records in Redis information on the distribution of tasks by workers; they, in turn, report to Redis on the status of their tasks. During the design process, components can be used according to their intended purpose.

Visualization of the system structure is shown in the *Figure 1.*

Analysis of text corpus (at this stage – corpus of news messages in Kazakhstan's Russian-language Internet media in the amount of 1.5 million documents with constant replenishment) is carried out by loading "workers". New documents are uploaded to the data processing subsystem using a special parser: at this stage, the download is done manually at the user's request, in the future the receiving new news will be configured according to the schedule (jobs running). With a given frequency, reports will be generated that require a lot of compu-

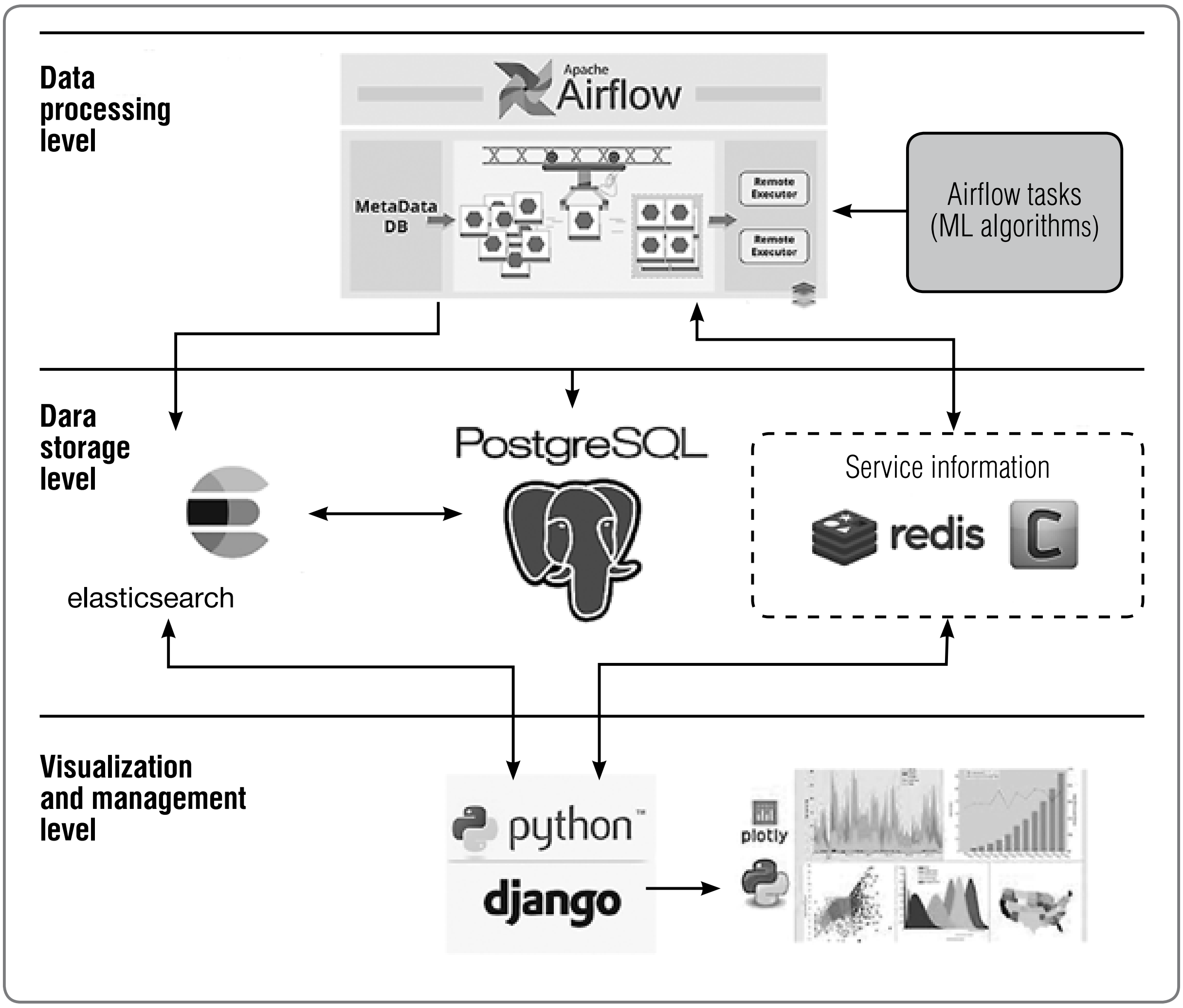


Fig. 1. Structure of the system developed

tational time; the results will be placed in the repository (this approach will reduce the waiting time for results from the data processing subsystem). Based on the data collected, additional model training will be performed (1–2 times per month), which will include recalculation of the set of key characteristics of the text corpus. In the event that additional training of the model does not lead to an increase in the accuracy index (for example, in the task of semantic analysis), the use of other ML-algorithms or their combination is provided.

The role system includes the following roles:

1. Custom user – has access to the basic functionality of the system: search, filtering, digital information panels (dashboards);

2. Advanced user – has access to custom reports, automatic alerts about "hot topics", the ability to filter by named entities (for example, person, organization, region) in articles. Such separation of users is due to subsequent use of the system by government bodies;

3. Developer – has access to the Airflow admin panel and to the repository where the Airflow DAG is stored. He can add and change his tasks, run and track their implementation;

4. Administrator – super-user, has a full set of rights to work with the system.

The role model is shown in *Figure 2*.

The following subsections describe the selected tools for each of the listed subsystems in more detail.

2.1. The data processing subsystem

During the analysis, Apache Airflow, an open source software platform, was chosen to these needs. The main components of this platform:

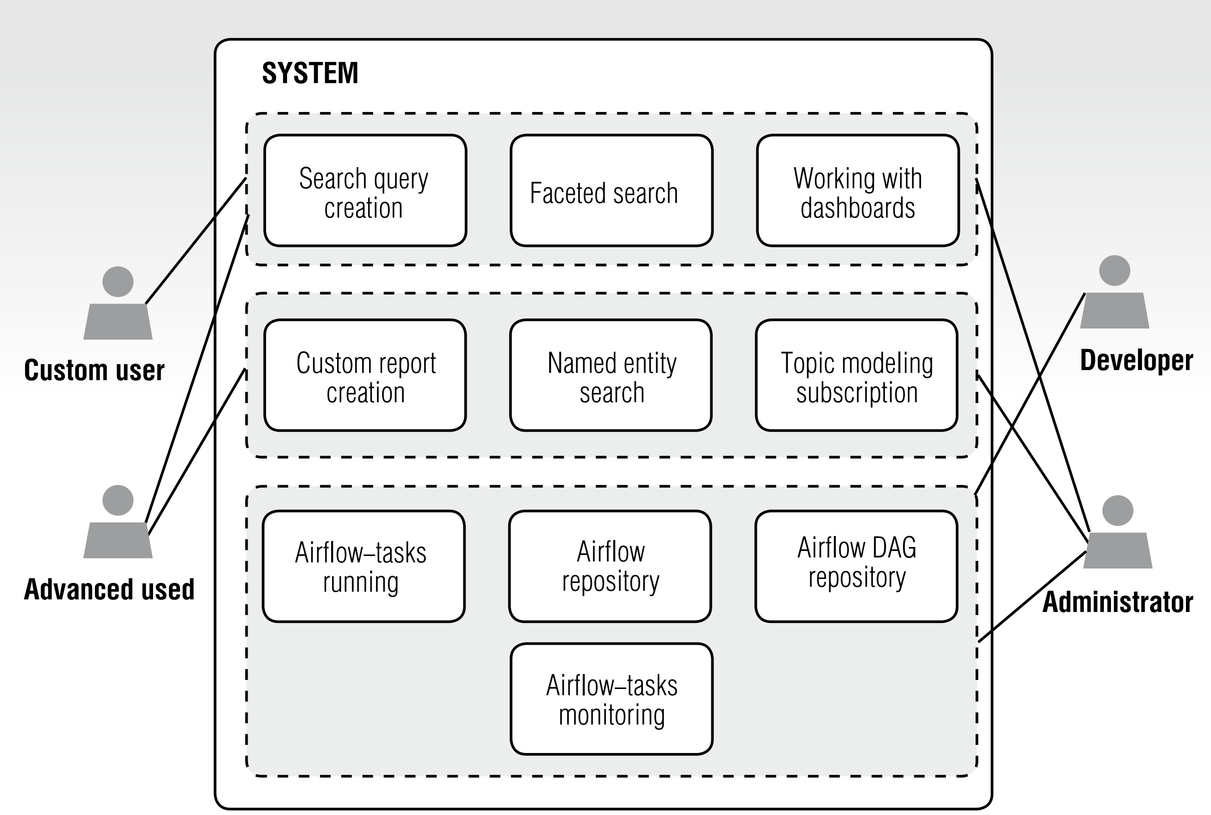


Fig. 2. The role model of the system

1. Airflow-worker — the main component that performs data processing. It can be scaled horizontally, including to individual servers or cloud virtual machines. In the current version of the architecture, the necessary dependencies are built into the Airflow-worker container image in advance. However, in principle, the process of dependency injection can occur in various ways, including by dynamically obtaining Docker containers from public or private repositories;

2. Airflow-scheduler — a component responsible for assigning tasks to Airflow-workers in the order defined by Airflow DAG. Airflow DAG is a non-cyclic directed graph that describes the order in which certain tasks are performed, and also contains information about the schedule, priorities, behavior in case of exceptions, etc.;

3. Airflow web server — a web interface that allows us to monitor and control the progress in tasks.

The machine learning algorithms are implemented in the system as separate Airflow tasks.

2.2. The data storage

There are three storage types in the system provided:

1. PostgreSQL — acts as a persistent storage for structured data. Its use is due to the wide capabilities of this relational database (among freeware) and interaction with a wide range of tools. The main data types stored in this database:

✦ news and metadata;

processed data at the level of different basic units of analysis (token / word / phrase / sentence / text), including vectorization, results of lemmatization, cleaning, etc.;

✦ the results of thematic modeling;

✦ the results of the classification of news on various grounds (semantic, politicization, social significance, etc.).

2. ElasticSearch — in-memory NoSQL storage designed for storing unstructured or weakly structured data, as well as quick search (including full-text) and filtering and streaming access. Compared with other NoSQL databases for storing documents with an arbitrary structure, such as MongoDB and CouchDB, ElasticSearch stands out with advanced tools for indexing text, which allow for full-text search in large volumes of documents in almost real time. Also, due to the possibility of constructing advanced indexes for data, it is possible to perform complex aggregations in the database itself, including distributed. ElasticSearch performs several functions:

✦ main storage for access, retrieval and filtering of data by the end user;

✦ main storage for ETL (extract, transform. Load) data processing processes, including the recording of any intermediate results in free form;

✦ storage for caching certain calculation results necessary for building dashboards and reports in the system;

ElasticSearch duplicates data stored in PostgreSQL as persistent storage, since ElasticSearch is an in-memory database without guarantees regarding data persistence and integrity.

3. Redis — a fast key-value storage used to cache individual pages and items, as well as to cache authorization sessions. Redis stores service data, as well as a cache of pages and items that are accessed frequently.

All three primary storages of the system can be easily scaled to several separate computers. Both horizontal scaling and replication are supported, with ElasticSearch and Redis showing near-linear increases in horizontal scaling performance.

A separate PostgreSQL cluster is used to store service data, such as task execution states. To run and track the progress of tasks, a bunch of Celery + Redis is used.

2.3. The subsystem for constructing analytical reports

The interface of the subsystem presents as an HTML + CSS + JS website with access via HTTP. The choice of the HTML + CSS + JS technology stack for the interface is justified by the fact that it is the web interfaces that are the most common and universally supported technology for building user interfaces, with the ability to access from any devices and operating systems from anywhere in the world, provided there is a web browser and Internet connection.

The web application is realized in the Django framework (Python), Gunicorn acts as the web server; the reverse proxy is Nginx. The web application has access to both the PostgreSQL persistent repository and ElasticSearch. Django has a built-in Cache Framework that allows us to cache pages and page elements in Redis. For example, if it is assumed that the page will be visited frequently, and it takes a long time for reading (for example, three seconds), then it is better to cache such a page in Redis, which will speed up access to the necessary data.

The Django framework was chosen for the following reasons:

1. The ability to quickly Agile-develop a web interface and data storage model. The development speed with Django is significantly higher than with competing products like Spring (Java), Yii (PHP) and Node.js (JavaScript);

2. Due to the project's involving the analysis of data and the construction of machine learning models, including for NLP, Python is the best choice, since most of the "state-of-the-art" models and ML/AI and NLP methods are developed by the community specifically in Python;

3. Django ORM works better with a PostgreSQL database.

The web application implements a series of pages for filtering, searching and accessing various dashboards and reports. At the first stage of the implementation of the system, dashboards are calculated in advance manually. With further development of the system, faceted search from Elastic Search will be used. The Plotly data visualization library will be used to create the graphs.

Examples of information that graphs can display are:

Dynamics by tonality (as well as manipulativeness, politicization, etc.), topics, number of views and comments, filtered by media, topics, authors, tags (including full-text search);

Distribution of topics, tonality values, etc. in statics, with filtering and search;

Identification of anomalies for analytical reports (the hottest topics, etc.).

Conclusion

This article formulates the requirements for the structure of a software system designed to process large (over one million units) corpus of text documents, including, in particular, the implementation of automated processing of corpus of texts, the ability to parallelize calculations and the preparation of analytical reports. The user role functions are defined. The structure of the software system was developed, including a data processing subsystem based on the Apache Airflow service, several types of storages that provide quick access to system components, and a subsystem for building analytical reports, which is generated in the Python Django application using the Plotly visualization library. The flexibility of the system allows us to select a different ensemble of machine learning algorithms, providing an increase in the quality and accuracy of the analysis of corpus of text documents.

Currently, the system is used to analyze news text corpus for the purpose of comparative analysis of news media corps in the Republic of Kazakhstan. ■

Acknowledgments

This work was funded by grant No BR05236839 of the Ministry of Education and Science of the Republic of Kazakhstan, by the Russian Fund of Basic Research, project No 19-31-27001 and within the framework of the state task theme No AAAA-A17-117120670141-7 (No 0316-2018-0009).

References

1. Barakhnin V.B., Kuchin Ya.I., Muhamedyev R.I. (2018). On the problem of identification of fake news and of the algorithms for monitoring them. Proceedings of the *III International Conference on Informatics and Applied Mathematics, Almaty, Kazakhstan, 26–29 September 2018*, pp.113–118 (in Russian).
2. Shokin Yu.I., Fedotov A.M., Barakhnin V.B. (2010) Technologies for construction of processing software systems dealing with semistructured documents aimed at information support of scientific activity. *Computational Technologies*, vol. 15, no 6, pp. 111–125 (in Russian).
3. Barakhnin V.B., Kozhemyakina O.Yu., Borzilova Yu.S. (2019) The development of the information system of the representation of the complex analysis results for the poetic texts. *Vestnik NSU. Series: Information Technologies*, vol. 17, no 1, pp. 5–17 (in Russian). DOI: 10.25205/1818-7900-2019-17-1-5-17.
4. Bolshakova E.I., Klishinskii E.S., Lande D.V., Noskov A.A., Peskova O.V., Yagunova E.V. (2011) *Automatic natural language text processing and computer linguistics.* Moscow: MIEM (in Russian).
5. Pang B., Lee L., Vaithyanathan S. (2002) Thumbs up? Sentiment classification using machine learning techniques. Proceedings of the *2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002), Philadelphia, PA, USA, 6–7 July 2002*, pp. 79–86. DOI: 10.3115/1118693.1118704.
6. Choi Y., Cardie Cl., Riloff E., Patwardhan S. (2005) Identifying sources of opinions with conditional random fields and extraction patterns. Proceedings of the *Conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT 2005). Vancouver, British Columbia, Canada, 6–8 October 2005*, pp. 355–362.
7. Manning C.D. (2011) Part-of-speech tagging from 97% to 100%: Is it time for some linguistics? Proceedings of the *12th International Conference "Computational Linguistics and Intelligent T ext Processing" (CICLing 2011), Tokyo, Japan, 20–26 February 2011*, pp. 171–189.
8. Mukhamedyev R., et al. (2020) Assessment of the dynamics of publication activity in the field of natural language processing and deep learning. Proceedings of the *4th International Conference on Digital Transformation and Global Society, St. Petersburg, Russia, 19–21 June 2019.* Springer, 2020 (in press).
9. Tarasov D.S. (2015) Deep recurrent neural networks for multiple language aspect-based sentiment analysis. *Computational Linguistics and Intellectual Technologies: Proceedings of Annual International Conference "Dialogue–2015"*, no 14 (21), vol. 2, pp. 65–74.
10. Garcia-Moya L., Anaya-Sanchez H., Berlanga-Llavori R. (2013) Retrieving product features and opinions from customer reviews. *IEEE Intelligent Systems*, vol. 28, no 3, pp. 19–27. DOI: 10.1109/MIS.2013.37.
11. Mavljutov R.R., Ostapuk N.A. (2013) Using basic syntactic relations for sentiment analysis. Proceedings of the *International Conference "Dialogue 2013", Bekasovo, Russia, 29 May – 2 June 2013*, pp. 101–110.
12. Prabowo R., Thelwall M. (2009) Sentiment analysis: A combined approach. *Journal of Informetrics*, vol. 3, no 2, pp. 143–157. DOI: 10.1016/j.joi.2009.01.003.
13. Dai W., Xue G.-R., Yang Q., Yu Y. (2007) Transferring naive Bayes classifiers for text classification. Proceedings of the *22nd National Conference on Artificial intelligence (AAAI 07). Vancouver, British Columbia, Canada, 26–27 July 2007*, vol. 1, pp. 540–545.
14. Cortes C., Vapnik V. (1995) Support-vector networks. *Machine Learning*, vol. 20, no 3, pp. 273–297. DOI: 10.1023/A:1022627411411.

15. Friedman J.H. (2001) Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, vol. 29, no 5, pp. 1189–1232.

16. Zhang G.P. (2000) Neural networks for classification: A survey. *IEEE Transactions on Systems, Man, and Cybernetics. Part C (Applications and Reviews)*, vol. 30, no 4, pp. 451–462.

17. Schmidhuber J. (2015) Deep learning in neural networks: An overview. *Neural Networks*, no 61, pp. 85–117. DOI: 10.1016/j.neunet.2014.09.003.

18. Devlin J., Chang M.-W., Lee K., Toutanova K. (2018) BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv:1810.04805.*

19. Vladimirova T.N., Vinogradova M.V., Vlasov A.I., Shatsky A.A. (2019) Assessment of news items objectivity in mass media of countries with intelligence systems: The Brexit case. *Media Watch*, vol. 10, no 3, pp. 471–483. DOI: 10.15655/mw/2019/v10i3/49680.

20. Romanov A.S., Vasilieva M.I., Kurtukova A.V., Meshcheryakov R.V. (2018) Sentiment analysis of text using machine learning techniques. Proceedings of the *2nd International Conference " R. Piotrowski's Readings in Language Engineering and Applied Linguistics (Saint-Petersburg, 2017)*, pp. 86–95 (in Russian).

21. Barakhnin V.B., Mukhamedyev R.I., Mussabaev R.R., Kozhemyakina O.Yu., Issayeva A., Kuchin Ya.I., Murzakhmetov S.B., Yakunin K.O. (2019) Methods to identify the destructive information. *Journal of Physics: Conference Series*, vol. 1405, no 1. DOI: 10.1088/1742-6596/1405/1/012004.

22. Barakhnin V.B., Kozhemyakina O.Y., Zabaykin A.V. (2014) The algorithms of complex analysis of Russian poetic texts for the purpose of automation of the process of creation of metric reference books and concordances. *CEUR Workshop Proceedings*, vol. 1536, pp. 138–143.

About the authors

Vladimir B. Barakhnin

Dr. Sci. (Tech.), Associate Professor;

Leader Researcher, Institute of Computational Technologies,
Siberian Branch of the Russian Academy of Sciences,
6, Academician M.A. Lavrentiev Avenue, Novosibirsk 630090, Russia;

Professor, Faculty of Information Technologies, Novosibirsk State University,
1, Pirogova Street, Novosibirsk 630090, Russia;

E-mail: bar@ict.nsc.ru

ORCID: 0000-0003-3299-0507

Olga Yu. Kozhemyakina

Cand. Sci. (Philol.);

Senior Researcher, Institute of Computational Technologies,
Siberian Branch of the Russian Academy of Sciences,
6, Academician M.A. Lavrentiev Avenue, Novosibirsk 630090, Russia;

E-mail: olgakozhemyakina@mail.ru

ORCID: 0000-0003-3619-1120

Ravil I. Mukhamediev

Dr. Sci. (Eng.);

Professor, Satbayev University, 22a, Satbayev Street, Almaty 050013, Kazakhstan;

Leader Researcher, Institute of Information and Computational Technologies,
125, Pushkin Street, Almaty 050010, Kazakhstan;

Professor, ISMA University, 1, Lomonosova Street, Riga LV-1019, Latvia;

E-mail: ravil.muhamedyev@gmail.com

ORCID: 0000-0002-3727-043X

Yulia S. Borzilova

Doctoral Student, Institute of Computational Technologies,
Siberian Branch of the Russian Academy of Sciences,
6, Academician M.A. Lavrentiev Avenue, Novosibirsk 630090, Russia;

E-mail: i.borzilova@alumni.nsu.ru

ORCID: 0000-0002-8265-9356

Kirill O. Yakunin

Doctoral Student, Satbayev University, 22a, Satbayev Street, Almaty 050013, Kazakhstan;

Developer Engineer, Institute of Information and Computational Technologies,
125, Pushkin Street, Almaty 050010, Kazakhstan;

E-mail: yakunin.k@mail.ru

ORCID: 0000-0002-7378-9212

DOI: 10.17323/1998-0663.2019.4.73.86

Application of the committee machine method to analysis of stock market technical indicators

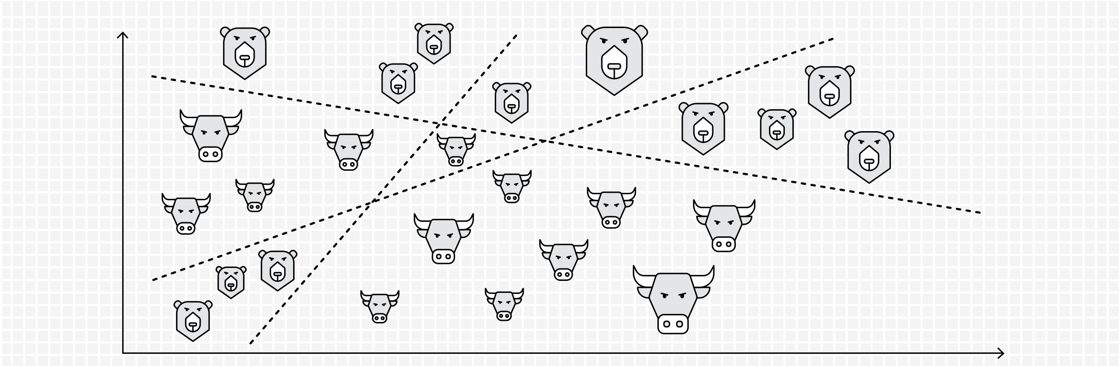
Nikolay P. Chernavin
E-mail: ch_k@mail.ru

Institute of Economics, the Ural Branch of Russian Academy of Sciences
Address: 29, Moskovskaya Street, Ekaterinburg 620014, Russia

Abstract

In this article we study problems of the committee machine method when applied to decision-making when there are many signals from different technical indicators of a stock exchange market. The committee machine method is a data classification method which can find non-linear data dependencies by construction of several linear classifiers. In the framework of this research, the basis for committee machine construction is a unified partially integer programming model, within which various logics of committee structures can be implemented. The subject of the research is the interrelation of indicators of technical indicators of a stock exchange market with pricing for financial instruments of stock exchange trading. Accordingly, the goal of the research is to show the efficiency of committee structures for solving the problems of forecasting the future value of financial instruments listed on stock exchange markets. To accomplish this goal, basic stock exchange data on Sberbank shares were collected from the Moscow Stock Exchange for the period from 2010 to 2019. On the basis of this, the technical indicators and interrelated parameters were calculated. They were used as data for the committee machine models with different numbers of committee members and voting logics. The result of the calculation was to obtain definitive rules, which when applied in speculative trading on the stock exchange market can generate stable profits. For comparison, we show the solutions of a similar problem by classical classification methods. The comparison shows that methods which work with the non-linear data dependencies provide results in terms of classification quality similar to committee machine results. This research may be interesting to the professional traders, investment analysts, specialists in data science and students with a mathematical and/or financial specialization.

Graphical abstract



Key words: committee machine method; stock exchange market; technical analysis; machine learning.

Citation: Chernavin N.P. (2019) Application of the committee machine method to analysis of stock market technical indicators. *Business Informatics*, vol. 13, no 4, pp. 73–86. DOI: 10.17323/1998-0663.2019.4.73.86

Introduction

At present, methods of technical analysis can rightly be called the most popular analytical tools in the financial markets industry. Technical analysis is a set of methods for predicting the dynamics of stock quotes through study of internal market factors, such as changes in prices, trading volumes, open interest, etc. The popularity of technical analysis methods is largely explained by two points. First, simple analytical methods do not require a strong professional knowledge base. Second, the possibility of obtaining economic benefits from the results of technical analysis is available to almost any person who has the minimum capital to open a stock exchange account.

A general picture of the technical analysis methods including their classification by data processing complexity is presented in *Figure 1* (compiled by the author using [1, 2]).

In accordance with *Figure 1*, technical analysis can be distinguished by formal-analytical and visual-graphic methods. The formal-analytical methods study stock exchange market data over the past period in the form of different indicators. On this basis, these methods predict growth or fall in stock prices. The visual-graphic methods are based on analysis through the study of graphical models built on stock exchange market data. This article explores the use of the committee method to evaluate the results of the technical analysis of stock prices.

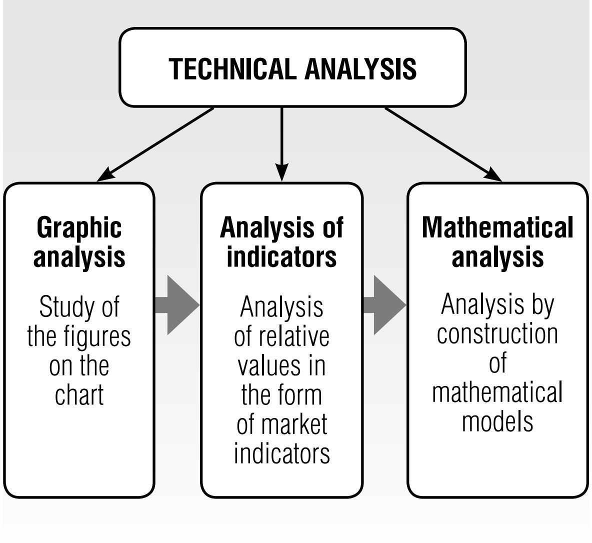

TECHNICAL ANALYSIS
Graphic analysis
Study of the figures on the chart
Analysis of indicators
Analysis of relative values in the form of market indicators
Mathematical analysis
Analysis by construction of mathematical models

Fig. 1. Classification of technical analysis methods based on data processing complexity

The committee machine method is one of the machine learning algorithms of binary classification. The term "committee machine" appeared for the first time in 1965 in an article by Ablow and Kaylor [3]. Further significant theoretical and practical research on the committee machine method was conducted by Osborne [4], Takiyama [5], as well as in proceedings of the N.N. Krasovskii Institute of Mathematics and Mechanics of the Ural Branch of Russian Academy of Sciences (IMM UB RAS), mainly by Mazurov and Khachai [6–16].

There is also interesting research into committee machine optimization by its confidence in the decision chosen that was conducted by Kuvshinov and Shiryaev [17–18]. Considerable reduction of the problem of committee machine construction to the task of linear partial-integer programming was studied by Chernavin and Nikonov [19–22].

1. Methodology of the research

Committee constructions are used to solve the problems of nonlinear discriminant analysis. Discriminant analysis is a group of machine learning methods that allow you to identify differences between groups and make it possible to classify objects according to the principle of maximum similarity. A necessary condition for applying discriminant analysis is the availability of a set of features (variables) [23], including a classifying (dependent) variable representing the class of the object, and discriminatory (independent) variables that reveal differences between objects of different classes.

The result of discriminant analysis is a discriminant function. The canonical discriminant function of the linear form has the following mathematical representation [23]:

$$d = \beta_0 + \sum_{i \in I} \left(x_{ij} \cdot \beta_i \right), \; j \in J, \qquad (1)$$

where d – classifying variable;

I – a set of features;

J – a set of all the considered objects in the space of the i-th number of features;

β_0 – free coefficient ensuring the fulfillment of the required conditions;

β_i – a discriminant function coefficient for the i-th feature;

x_{ij}– a discriminatory variable for the i-th feature of the j-th object.

Committee machine constructions use several discriminatory functions called the committee machine members. Depending on a value of the classification variable, it is said that a member of the committee votes "for" or "against" a particular decision. The final decision is made relying on the decisions of all the committee members and their processing using the committee machine logic. There are three main committee machine logics: unanimity (CU), majority (CM) and seniority (CS) logics.

To clarify the difference between different logics of the committee machines, we consider graphical examples for the class of objects: "circles" and "stars". These objects are characterized by two parameters along the axes X_1 and X_2. Accordingly, the necessary conditions for discriminant analysis are fulfilled by the presence of a classifying variable (binary division into the classes: for example, "circles" = 0 and "stars" = 1), as well as the presence of discriminatory variables, expressed by the values of each object along the X_1 and X_2 axes. The results of discriminatory analysis are displayed in the form of lines representing individual members of the committee machine (linear discriminant functions), and the arrows for each line which indicate the direction of voting of the corresponding committee member.

In a case of the CU, an object belongs to a certain class only if all committee machine members vote for this class, otherwise it will belong to another class. A graphical example of the CU of three members is presented in *Figure 2.*

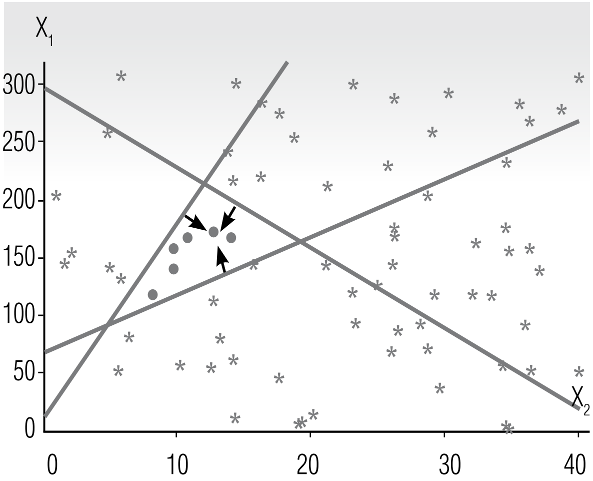


Fig. 2. An example of CU in a two–dimensional space

Figure 2 shows that the "circles" are in the area of the graphic for which all committee machine members voted unanimously. All observations in the remaining areas will contain "stars."

In accordance with the names, CM is a committee machine where an object belongs to a certain class if a majority of committee machine members vote for this class; otherwise it belongs to another class. A graphical example of CM of three members is presented in *Figure 3.*

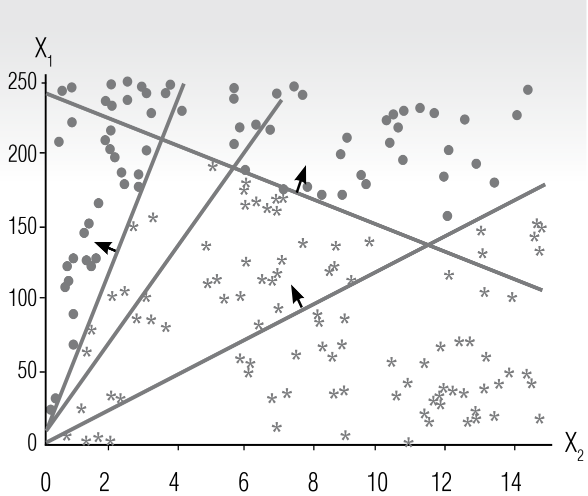


Fig. 3. An example of CM in the space of two features

Figure 3 shows that the "circles" are in the area of the graphic, for which two or more committee machine members voted. All observations in the remaining areas will contain "stars."

The last logic requires different committee members to have certain weights reflecting their power in the voting process. As a result, an object belongs to a certain class if there are enough weights from the votes of the committee machine members for this class; otherwise it belongs to another class. A graphical example of CS of three members is presented in *Figure 4* (in addition to the previous designations, the lines correspond to the weights of the committee members).

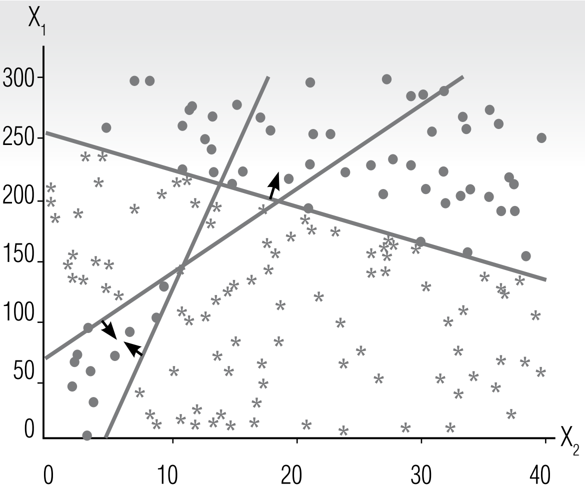


Fig. 4. An example of CS in the space of two features

Figure 4 shows that the "circles" are in the area of the graphic, which have a sum of vote weights more than three. All observations in the remaining areas will contain "stars."

The examples presented reflect cases where the committee machine allows us to classify objects into classes precisely. However, in real life, as a rule, not all objects can be accurately attributed to a certain class. Therefore, the committee machine should classify the objects in such a way that the quality of the division into classes would be best.

To illustrate this idea, let’s suppose that there is some initial division of objects into classes of “circles” and “crosses.” In *Figure 5*, a graphical example of such classification is presented.

Figure 5 shows the case of the CU committee machine of three members in the space of two features, where there are examples when objects are classified to the wrong class. For clarity, such objects are circled on the graph. The presence of the classification errors is a natural phenomenon which can be explained by many factors, such as incorrect initial markup of objects into classes, lack of initial information for more accurate classification, insufficient number of committee members, incorrect committee machine logic, etc.

Mathematically, a committee machine can be described as a linear partially integer programming model. Below an example of such a model is presented:

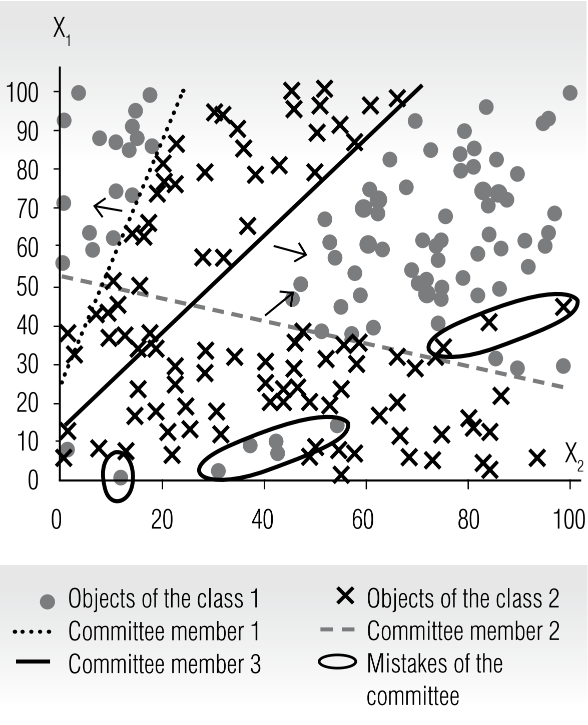


Fig. 5. An example of CM in the space of two features in a case where there are classification mistakes

$$
\begin{aligned}
&\sum_{i\in I}\left(x_{ij}\cdot\beta_i^t\right)+\beta_0^{\,t}-L\cdot z_j^t\le-\varepsilon,\ j\in J_1,\ t\in T,\\
&\sum_{i\in I}\left(x_{ij}\cdot\beta_i^t\right)+\beta_0^{\,t}+L\cdot z_j^t\ge\varepsilon,\ j\in J_2,\ t\in T,\\
&\sum_{t\in T}\left(z_j^t\cdot V^t\right)\le m+L\cdot d_j,\ j\in J_1,\\
&\sum_{t\in T}\left(z_j^t\cdot V^t\right)\le sv-m-1+L\cdot d_j,\ j\in J_2,\\
&\sum_{j\in J_1}\left(d_i\right)\le K_1\cdot\varphi,\\
&\sum_{j\in J_2}\left(d_i\right)\le K_2\cdot\varphi,\\
&\min\varphi.
\end{aligned}
\qquad (2)
$$

where J_1 – a set for classification in the space of the *i*-th number of features belonging to class 1;

J_2 – a set for classification in the space of the *i*-th number of features belonging to class 2;

I – a set of observation parameters;

T – a set of the committee members;

i, j, t – indices of the corresponding sets;

x_{ij} – *i*-th parameter of *j*-th observation (discrimination variables);

β_i^t – a coefficient for the *i*-th parameter of *t*-th committee machine (decision variables);

$\beta_0^{\,t}$ – an absolute term of the *t*-th committee machine (decision variable);

z_j^t – a Boolean variable to fixate vote direction of the *t*-th hyperplane for the *j*-th observation;

d_j – a Boolean variable to fixate errors of a committee machine classification for the *j*-th observation;

L – a number much greater by value than parameters of a model (introduction of such constant is necessary to make it possible for $L\cdot z_j^t$ and $L\cdot d_j$ to be used as the residuals);

ε – a small number close to 0 (introduction of such constant makes not strict inequalities equivalent to the strict inequalities (> or <), so this way we exclude as a decision a case when all coefficients are equal to zero);

V^t – a weight of the *t*-th committee machine member (constants which are calculated as a power of two);

φ – a Chebyshev approximation (minimax) variable. Using this value for minimization allows us to search for the optimal solution in such a way that the share of classification errors for each class of objects will be minimized at the same time;

sv – a sum of all committee machine weights $\left(\sum_{t\in T} V^t\right)$;

m – a minority of a committee machine (a variable in the range $0 \le m \le sv - 1$).

This model (2) can be used for each of the committee machine logics. If $V = 1$, then depending on *m* there may be constructed CU ($m = 0$ or $m = sv - 1$) or CM ($0 < m < sv - 1$). If $V = 2^t$, then depending on *m* there may be constructed CU ($m = 0$ or $m = sv - 1$) or CS ($0 < m < sv - 1$).

2. Description of assigned task

The main problem of the technical analysis methods is the presence of a multitude of false signals, the screening of which requires experience and intuition. Accordingly, we will try to conduct such screening with implementation of the committee machines. To accomplish this task, 10 market indicators were selected (*Table 1*).

The basis for the selection of indicators is a set of indicators described by Elder in his Triple Choice system [24]. At the same time, there is no preservation of the philosophy of the Triple Choice system associated with triple sifting of a deal before making a decision, since it is assumed that the committee method based on the composition of the indicators is able to reveal a decisive rule that has greater potential profitability.

Let's consider in more details the calculation of the indicators presented in *Table 1.*

To calculate indicators P_1 and P_2, it is necessary to calculate MACD histogram in the following way:

1. Calculate 12-day EMA of daily close prices;

2. Calculate 26-day EMA of daily close prices;

3. Calculate a "fast" line, also known as MACD line (26-day EMA of daily close prices calculated on 12-day EMA of daily close prices);

4. Calculate a "slow" line, also known as signal line (9-day EMA calculated on values of a "fast" line);

5. Calculate an MACD histogram as the difference between "fast" MACD line (step 3) and "slow" signal line (step 4).

Indicator P_1 shows how many days the MACD histogram grows or falls. Moreover, if the indicator falls, then the number of days is taken with the "minus" sign.

Indicator P_2 shows how many days the MACD histogram is more or less than zero. In this case, if the indicator is less than zero, then the number of days is taken with a "minus" sign.

The *SSO* in the indicators P_3 and P_4 is used to define the overbought and oversold market states. It is calculated as follows:

$$SSO = Mean_n\left(100 \cdot \frac{\sum_s (C_c - Min_r)}{\sum_s (Max_r - Min_r)}\right), \qquad (3)$$

where *r* – time period of SSO (in this research $r = 5$);

s – a time interval of the averaging (in this research $s = 3$);

Max_r – a maximum in a time interval *r*;

Min_r – a minimum in a time interval *r*;

C_c – a closing price on the moment of SSO calculation;

$Mean_n$ – function to calculate a mean value in a period of *n* days (in this research $n = 5$).

There are different opinions regarding the overbought and oversold market states. Usually, it is considered that the market is over-

Table 1.

Analyzed indicators

Indicator	Description	Value range
P_1	How many days in a row does a moving average convergence divergence histogram (MACD) decrease or increase	Integer
P_2	How many days in a row is the MACD[1] histogram more or less than zero	Integer
P_3	Signal from a slow stochastic oscillator (SSO[2])	from 0 to +1
P_4	How many days in a row does SSO give a strong signal	Integer
P_5	Signal from a relative strength index (RSI[3])	from 0 to +1
P_6	How many days in a row does RSI give the same signal	Integer
P_7	Trend calculated by dynamics of an exponential moving average (EMA) from the price maximums	–1, 0, +1
P_8	How many days in a row does the Chaikin indicator equal more or less than zero	Integer
P_9	How many days in a row does the Chaikin indicator[4] decrease or increase	Integer
P_{10}	Is the price maximum higher or is the minimum lower compared to the previous day. Otherwise it equals zero.	–1, 0, +1

sold when $SSO \leq 0.2$, and is overbought, if $SSO \geq 0.8$. Consequently, there is no meaning to use this indicator directly in a model, because it would not fully reflect the logic of oversold and overbought. Thus, as a model parameter this indicator needs to be transformed by the following formula:

$$P_3 = (2 \cdot (SSO - 0.5))^3 \qquad (4)$$

Indicator P_4 is used for a more detailed description of P_3. It indicates how many days SSO is in the overbought area (taken with a “minus” sign) or oversold (taken with a “plus” sign). This indicator is used for approximate digitization of an estimate of an SSO chart by the eyes of a trader, when not only current values are estimated, but also previous dynamics.

Indicator P_5 is based on the RSI indicator, which is calculated as follows:

$$RSI = 1 - \frac{1}{1 + RS}, \qquad (5)$$

where RS – a ratio of the average value of increasing closing prices for seven days to the average value of lowering closing prices for seven days.

In this research the market is considered oversold when $RSI \leq 0.3$, and is overbought, if $RSI \geq 0.7$. This indicator is transformed the same way as P_3 by the following formula:

$$P_5 = (2 \cdot (RSI - 0.5))^3 \qquad (6)$$

Indicator P_6 additionally analyzes RSI by showing how many days in a row the RSI sends

[1] For more details on the calculation and application, see [24]

[2] For more details on the calculation and application, see [24]

[3] For more details on the calculation and application, see [24]

[4] The Chaikin indicator shows what is happening with the trend based on the ratio of the difference between closing and opening prices to the difference between the maximum and minimum prices multiplied by the volume of trades. For more details on the calculation and application, see https://www.opentrainer.ru/articles/ostsillyator-chaykina-chaikin-oscillator/ (access date: 01 September 2019)

a signal of the same type, and this indicator can be either positive or negative depending on the type of signal (that is, if the signal is less than zero, then the number of days is taken with a "minus" sign).

The daily trend indicator was chosen as indicator P_7. Classically, the trend is determined by the slope of the line drawn through the local price highs. However, it is impossible to draw an ideal line through all local highs in the selected study interval; therefore, as a rule, the line is drawn empirically, so as to pass through the greatest number of price maxima with the minimum deviation from them. A trend has three main states: rising, falling and no trend (flat). In this study, the trend is determined by the change in the 22-day exponential moving average (EMA) from a maximum daily price. If the EMA is growing for the second day in a row, then this indicator equals one until it begins to fall. Accordingly, and vice versa, if the EMA falls for the second day in a row, then this indicator equals –1 until it starts to grow. If the EMA grew or fell within just 1 day, then the indicator is zero.

Indicators P_8 and P_9 reflect the signals given by the Chaikin indicator that is calculated as follows:

$$CHI = EMA_3(A/D) - EMA_5(A/D),$$
$$A/D = \frac{C_c - C_o}{Max - Min} \cdot V, \tag{7}$$

where A/D – accumulation/distribution indicator;

C_c – closing price in a day of calculation;

C_o – opening price in a day of calculation;

Max – maximum price in a day of calculation;

Min – minimum price in a day of calculation;

V – trade volume in a day of calculation;

EMA$_3$, EMA$_5$ – operations of EMA calculation with a time period of 3 and 5 days respectively.

Indicator P_8 shows how many days in a row the Chaikin indicator is more or less than zero. Moreover, if the indicator is less than zero, then the number of days is taken with a "minus" sign.

Indicator P_9 shows how many days in a row the Chaikin indicator is rising or falling. Moreover, if the indicator falls, then the number of days is taken with the "minus" sign.

The last indicator P_{10} shows whether the maximum price was higher than the previous day (equals 1) or the minimum lower than the previous day (equals –1). Otherwise, if prices have not gone beyond the price range of the previous day, then the parameter equals zero. The market also has a case when during the day the maximum was simultaneously higher than the previous day, and the minimum was lower than the previous day. Mostly they are linked with unpredictable events like news and are quite rare. Considering this fact, they were excluded from a list of the samples to reduce information noise from the intersection of the samples differing by the trading environment.

Examples of indicators for PJSC Sberbank shares are presented in *Table 2.*

3. Trading strategy

To build a committee machine model, one needs to form a trading strategy. The basic parts of a trading strategy are:

1. Start-up capital, that is, a fixed amount within which you can open a position[5];

2. Stop-loss (a preset value of maximum loss at which the current open position will be closed[6]).

Let's determine that a trading signal comes from P_3 and P_5. If both parameters are less than

[5] To open a position means to initially purchase or sell a certain volume of a financial instrument

[6] To close a trading position means to complete a reverse trade operation in relation to opening a position. For example, if a position was opened as part of a purchase transaction of a certain volume of a financial instrument, then the reverse transaction will be expressed by the subsequent sale of the acquired volume of the financial instrument either in full (complete closing of the position) or some part of it (partial closing of the position)

Table 2.

Examples of indicators for PJSC Sberbank shares

Date	Indicators									
	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9	P_{10}
05.02.2010	0.2	1	–0.001	0	–0.36	1	–1	–2	–2	–1
08.02.2010	0.3	2	–0.108	0	–0.342	2	–1	–3	–3	–1
09.02.2010	0	3	–0.291	1	–0.14	3	–1	–4	–4	–1
10.02.2010	–0.1	4	–0.375	2	–0.075	4	–1	–5	0	1
11.02.2010	–0.2	5	–0.166	3	–0.144	5	–1	–6	1	1

0.25, then this is a signal to buy (open a long position), and if they are more than 0.75, then a signal to sell (open a short position) is given. According to the strategy, the signal is formed at the end of the trading day, but we will assume that the position with the corresponding signal opens in the last minutes. In this case, we determine that the position is closed if a signal appears with a reverse direction, or losses are greater than a chosen stop-loss value.

There are various methods for choosing a stop loss value. In the framework of this study, we used the calculation formula with reference to an average price volatility[7] of seven days, which will limit losses for each observation. Accordingly, a stop-loss will be calculated using the following formula:

$$SL = \overline{V_7} \cdot s, \qquad (8)$$

where $\overline{V_7}$ – an average price volatility of seven days;

s – a coefficient that shows how much of the average daily volatility is allowed to risk in the framework of the trading strategy.

The described strategy can be schematically presented as presented in *Figure 6.*

The trading strategy shown in *Figure 6* is built by using two indicators of technical analysis – *SSO* and *RSI.* Trades in a trading strategy are formed only upon receipt of a same signal to buy or sell from both indicators. However, it must be understood that even after receiving strong signals from technical indicators, it is possible to experience losses from the trading activities.

4. A problem of machine learning

To reduce potential losses, a trading strategy must be presented as a discriminant analysis problem, in which the following classes of objects will be recognized:

1. Class 1 is a set of the positions that were closed with a profit and the profit was greater than the maximum loss[8] during the holding of the position[9];

2. Class 2 is a set of the positions that were closed with a loss, or a profit that was less than the maximum loss while the position was held.

After determining the strategy, it is necessary to choose a financial instrument for which the decision rule will be built. As part of the research,

[7] Daily volatility is a value that shows by how many percent the daily maximum is higher than the daily minimum

[8] When a position is opened on the stock exchange, it is constantly reevaluated at the current prices. Accordingly, the maximum loss on a position is the maximum amount of decrease in the value of a position relative to its value at the opening

[9] The period of position holding is a period from the time of an initial purchase or sale of a certain volume of a financial instrument to the time of a trade operation that is inverse to the first deal. More information on the concepts of the position opening and closing can be found at: https://www.metatrader5.com/ru/mobile-trading/android/help/trade/positions_manage/open_positions (access date: 01.09.2019)

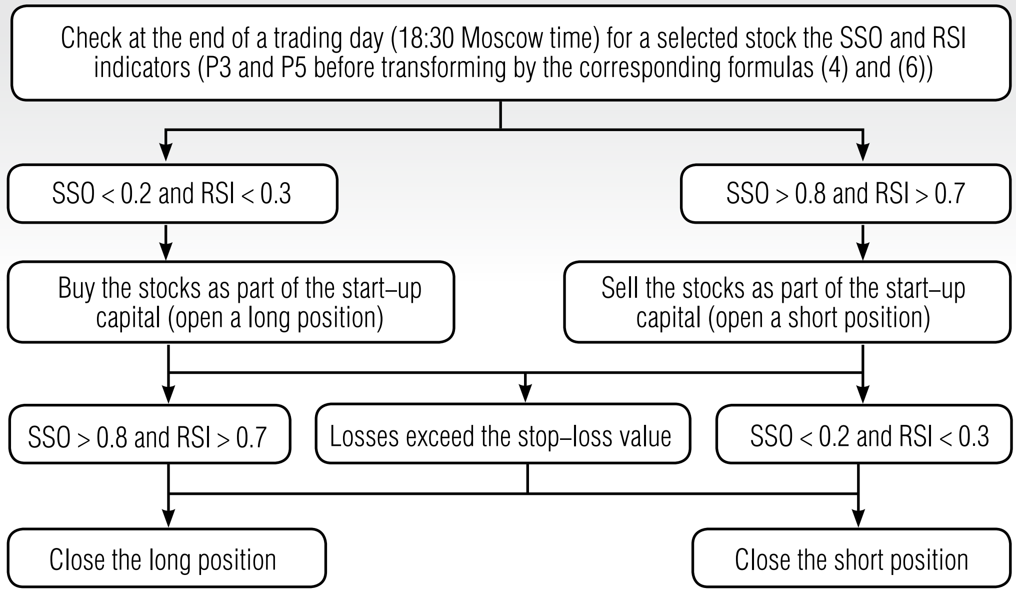

Check at the end of a trading day (18:30 Moscow time) for a selected stock the SSO and RSI indicators (P3 and P5 before transforming by the corresponding formulas (4) and (6))
SSO < 0.2 and RSI < 0.3
SSO > 0.8 and RSI > 0.7
Buy the stocks as part of the start-up capital (open a long position)
Sell the stocks as part of the start-up capital (open a short position)
SSO > 0.8 and RSI > 0.7
Losses exceed the stop-loss value
SSO < 0.2 and RSI < 0.3
Close the long position
Close the short position

Fig. 6. Trading strategy

Sberbank shares were chosen as such a tool given that they are the most liquid asset on the Moscow Exchange's share market. For the analysis, daily quotes on Sberbank shares were downloaded from the Finam[10] website for the period from 2010 to December 2018.

All observations were divided into training (a set with observations, on which a model trains) and test (a set with observations, on which a model does not train) sets, as shown in *Table 3.* This distinction is necessary to be able to check the consistency of the model on these two sets, which reduces the risk of retraining the model [25].

Table 3.

Description of the training and test sets

Set	Period	J_1	J_2	J
Training	2010–2017	332	116	448
Test	2018	37	15	52
Total	2010–2018	369	131	500

Table 3 provides a short description of the training and test sets for a trading strategy described in *Figure 6*[11]. According to it, the number of samples in J_1 is about 2.5–3 times greater than in J_2, and the test set equals 11.6% size of the training set. With the classification task and sets defined let's start our calculation of the mathematical model.

5. Classification results

For comparison, we constructed various committees from 3 to 7 committee machine members. Classification quality was evaluated using the F-measure metric, calculated by the following formula:

$$SL = \overline{V_7} \cdot s, \qquad (9)$$

where A – accuracy (the share of objects assigned by a committee machine to the desired class and at the same time really being objects of this class);

[10] Finam – Moscow exchange stocks – Sberbank: https://www.finam.ru/profile/moex-akcii/sberbank/export/ (access date: 01 September 2019)

[11] Samples were built for a trading strategy with a stop-loss selected by formula (8) with a coefficient $s = 2.3$

C – completeness (the share of objects of the desired class that a committee machine was able to recognize).

Studying the results among CU, the decisive rule for the case of five members with the result of 86.4% for J_1 and 68.9% for J_2 has shown the best separating ability on the training set. For the CS, the best case consists of six members with the result 89.1% for J_1 and 73.6% for J_2. Here are not presented results for the CM, since the mathematical model has not found an appropriate quality solution for the majority logic[12].

To evaluate the quality of the model, the most important thing is its results on the test set. This is linked to the fact that the results obtained on the test set have no risk of over-training and subject to the condition that the size of a test set is big enough it can objectively assess the predictive ability of a model. Accordingly, the F-score metric results of the model on the test set are presented in *Figure 7*.

Studying the results on the test set in *Figure 7* it can be seen that the previously reviewed CU of five members and the CS of six members also have the best result. Thus, we can state that the training and test sets really have strong links between each other, and that proves the fact that our problem can be studied by construction of committee machines. By further comparing these two committee machine decisions by the F-score metric, it can be seen that CU results of five members are better than for CS of six members.

So in this model we studied minimization of a number of errors in predicting the correct or incorrect signals from technical indicators. However, in the framework of this model, the value of the potential profits or losses, depending on the accepted decisions was not taken into account. Since the ultimate goal of modeling is to make a profit, introducing this evaluation criterion into the model can improve the final result. To do this, in the model for each observation of the training and test samples, we calculate the value of potential gains or losses by the following formula:

$$PL = \left(\frac{C_o}{C_c} - 1\right) \cdot D, \qquad (10)$$

where C_0 – the opening price of a position (initial purchase/sale price of a financial instrument, as a result of which a trading position was formed);

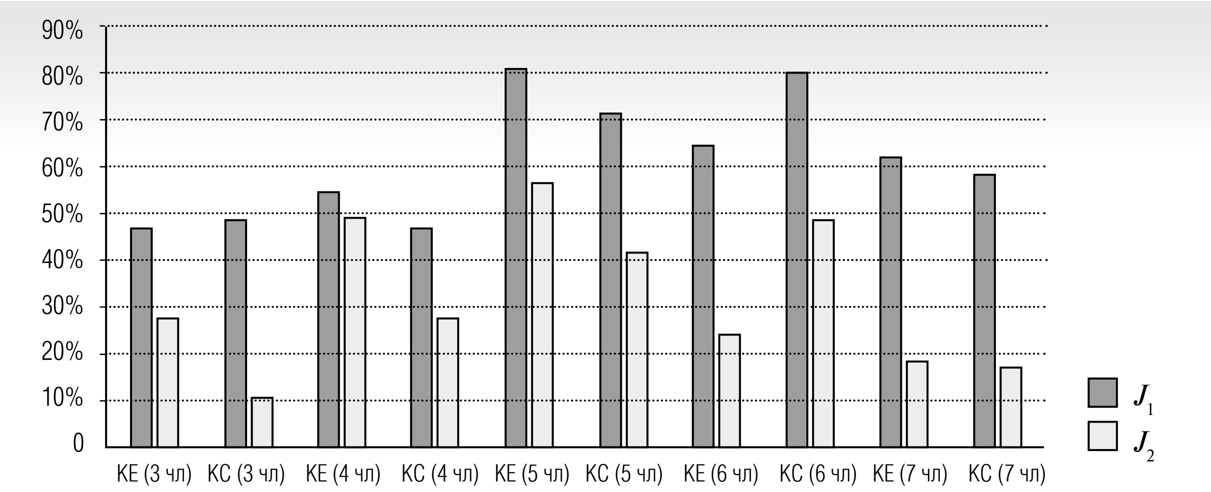


Fig. 7. F–score results on the test set

[12] While searching for a solution in the framework of the model (2) with $V = 1$, the solution with the lowest value of φ was found at $m = 0$, which corresponds to CU

C_c – the closing price of a position (price at which a previously opened position was closed on a stock exchange, as a result of which profits or losses were recorded on the position);

D – position direction (1 – if a financial instrument was bought; –1 – if a financial instrument was sold).

If the value obtained by the formula (10) is greater than zero, then a position gives a profit, otherwise there is a loss by position. Accordingly, the previous model (2) may be changed by addition of weights to each observation with addition of corresponding change in restrictions $\sum_{j\in J_1} d_j$ and $\sum_{j\in J_2} d_j$ in the model (2):

$$\begin{aligned} \sum_{j\in J_1} d_j w_j &\le \sum_{j\in J_1} w_j\varphi, \\ \sum_{j\in J_2} d_j w_j &\le \sum_{j\in J_2} w_j\varphi, \end{aligned} \tag{10}$$

where w_j – weight of the j-th observation.

Let's also additionally reduce the coefficient s in the stop loss calculation formula (8) to 1.5 in order to reduce the level of losses for each position formed by the trading strategy. With addition of this stop-loss parameter we changed the number of observations for each class on the training set, because 16 observations from the J_2 were closed by the stop-loss rule. In other words, they have been classified as the J_1. Accordingly, on the new training set we constructed various committees taking into account the conditions (3) and the best results were obtained by CM of seven members with the F-score metric results 92.1% for J_1 and 76% for J_2 on the training set, and 83.1% for J_1 and 51.9% for J_2 on the test set. A corresponding decision rule for CM of seven members is presented in the *Table 4.*

Let's compare the previous best decision rule with the decision rule from *Table 4* by the value of potential return. To simplify calculations, let's assume that each deal is formed with the same startint capital. Consequently, *Table 5* presents the results of comparison by years of CM of seven members and CU of five members taking into account the chosen stop-losses.

Based on the profitability results in *Table 5*, we see that for a CM of seven members, income from correctly recognized observations from J_2 and losses from wrongly recognized observations from J_1 have a much smaller variation of values than for a CU of five members. At the same time, losses on wrongly recognized observations from J_1 for a CM of seven members are two times less than for a CU of five members, while incomes from correctly recognized observations from J_2 are only 9.6% less. Moreover, for the considered CM of seven members there is no such annual period when the losses were greater than the income, whereas for the

Table 4.

Decision rule for the CM of seven members (*m* = 2)

Members	Coefficients										
	β^t_0	β^t_1	β^t_2	β^t_3	β^t_4	β^t_5	β^t_6	β^t_7	β^t_8	β^t_9	β^t_{10}
$t = 1$	–1.095	–0.041	0.088	0.006	–0.297	–0.021	0.471	1	0.296	–0.001	–0.298
$t = 2$	–0.352	0.64	–0.006	–0.531	–0.419	0.222	0.168	1	–0.355	0.158	0.224
$t = 3$	–123.18	–153.46	–4.664	6.719	89.83	–39.54	–33.635	–1	25.408	4.748	–75.751
$t = 4$	–1.859	–0.437	–0.193	0.112	0.004	–0.067	0.042	1	–0.062	0.157	0.045
$t = 5$	0.469	–1.666	–0.626	–0.369	–0.787	–0.1	–0.435	1	–1.118	–1.843	–0.142
$t = 6$	–1.573	0.164	0.03	0.058	–0.548	0.05	0.149	–1	–0.155	0.55	–0.723
$t = 7$	–0.906	0.211	0.022	–0.151	0.104	0.124	–0.141	–1	0.073	–0.011	0.263

Table 5.

Profit and loss evaluation of decision rules

Committee type	Class	% of profit or losses by year									
		2010	2011	2012	2013	2014	2015	2016	2017	2018	Total
CM of 7 members	J_1	–13.6	+0.6	–7.4	–32.0	–26.9	–24.6	–8.5	–5.8	–17.3	–136
	J_2	+87.1	+134.3	+37	+46.4	+78.6	+34.8	+67.9	+63.3	+62.8	+612
CU of 5 members	J_1	–53.3	–8.7	–4.5	–14.3	–66.7	+3.8	–43.7	–55.6	–28.4	–271
	J_2	+83.5	+137.8	+35	+65.2	+89.7	+28.3	+101.7	+52.7	+83.7	+677

case of the CU of five members in 2017 there was an excess of losses over the income. Thus, we can say that optimization with minimization of losses in a case of work with technical indicators provide better results than just minimization by the number of classification errors.

Conclusion

As the result of the study, we have shown what the committee machine method is and how various committee machine logics can be calculated within the framework of a single mathematical model of partially integer programming. Using a practical example, we consider the use of the committee method to increase the accuracy of forecasting prices on the stock exchange market while performing technical analysis. The use of committee machine constructions made it possible to derive objective decision rules that determine when to follow recommendations based on market indicators and when to refrain from active trading on the stock exchange market. Moreover, within the framework of committee machine constructions, as models of partially integer programming, complex optimization criteria can be formed which take into account the problem of maximizing profits and minimizing losses. A model with this formulation of the problem allows us to get more stable results with a potentially higher total profit, compared with models based on a simple minimization of the number of classification errors. Thus, committee machine constructions can be used on the financial markets both to form new trading strategies and to increase the profitability of existing strategies. ■

Acknowledgements

The article was prepared as part of the state assignment for the Institute of Economics, the Ural Branch of Russian Academy of Sciences for 2019.

References

1. Malyshenko K.A., Malyshenko V.A., Kvyatkovskaya E.O. (2017) Theoretical bases of the analysis of the stock market: System of indicators and classification of methods. *Scientific Journal of KubSAU*, vol. 129, no 5, pp. 1292–1303 (in Russian).
2. Neiman E. (2015) *The small encyclopedia of trader*. Moscow: Alpina Publisher (in Russian).
3. Ablow C.M., Kaylor D.J. (1965) Inconsistent homogeneous linear inequalities. *Bulletin of the American Mathematical Society*, vol. 71, no 5, p. 724.
4. Osborne M.L. (1977) The seniority logic: A logic for a committee machine. *IEEE Transactions on Computers*, vol. 26, no 12, pp. 1302–1306. DOI: 10.1109/TC.1977.1674798.
5. Takiyama R.A. (1978) General method for training the committee machine. *Pattern Recognition*, vol. 10, no 4, pp. 255–259. DOI: 10.1016/0031-3203(78)90034-1.
6. Mazurov Vl.D. (1990) *The method of committees in optimization and classification problems*. Moscow: Nauka (in Russian).

7. Mazurov Vl.D., Khachai M.Yu. (2004) Committees of systems of linear inequalities. *Automation and Remote Control*, vol. 65, no 2, pp. 193–203. DOI: 10.1023/B:AURC.0000014716.77510.61.
8. Mazurov Vl.D., Khachai M.Yu. (2013) Boosting and the polynomial approximability of the problem on a minimum affine separating committee. *Trudy Instituta Matematiki i Mekhaniki UrO RAN*, vol. 19, no 2, pp. 231–236 (in Russian).
9. Mazurov Vl.D., Khachai M.Yu., Poberii M.I. (2008) Combinatorial optimization problems related to the committee polyhedral separability of finite sets. *Trudy Instituta Matematiki i Mekhaniki UrO RAN*, vol. 14, no 2, pp. 89–102 (in Russian).
10. Mazurov Vl.D., Khachai M.Yu., Rybin A.I. (2002) Committee constructions for solving problems of selection, diagnostics, and prediction. *Trudy Instituta Matematiki i Mekhaniki UrO RAN*, vol. 8, no 1, pp. 66–102 (in Russian).
11. Mazurov Vl.D., Khachai M.Yu. (2003) Committee constructions as generalization of solutions to controversial problems of operations research. *Discrete Analysis and Operations Research*, ser. 2, vol. 10, no 2, pp. 56–66 (in Russian).
12. Mazurov V.D., Khachai M.Yu. (2007) Parallel computations and committee constructions. *Automation and Remote Control*, vol. 65, no 2, pp. 193–203. DOI: 10.1134/S0005117907050165.
13. Mazurov Vl.D., Smirnov A.I. (2012) Interpretation of contradictory images by means of systems of linear inequalities. *Trudy Instituta Matematiki i Mekhaniki UrO RAN*, vol. 18, no 3, pp. 144–154 (in Russian).
14. Khachai M.Yu. (2010) Computational complexity of recognition learning procedures in the class of piecewise-linear committee decision rules. *Automation and Remote Control*, vol. 71, pp. 528–539. DOI: 10.1134/S0005117910030136.
15. Khachai M.Yu. (1997) The existence of the majority committee. *Discrete Mathematics*, vol. 9, no 3, pp. 82–95 (in Russian).
16. Khachai M.Yu. (1997) Estimate of the number of members in the minimal committee of a system of linear inequalities. *Computational Mathematics and Mathematical Physics*, vol. 37, no 11, pp. 1399–1404 (in Russian).
17. Kuvshinov B.M., Shiryaev O.V. (2002) The method of committees in the problems of pattern recognition in the conditions of uncertainty of a priori information. *Bulletin of the South Ural State University, Series "Mathematics. Mechanics. Physics"*, vol. 2, no 3, pp. 34–43 (in Russian).
18. Kuvshinov B.M., Shiryaev O.V., Bogdanov D.V., Shaposhnik I.I., Shiryaev V.I. (2001) The system for classification of multiparametric objects for pattern recognition problems with inaccurate a priori information. *Information Technologie*s, no 11, pp. 37-43 (in Russian).
19. Nikonov O.I., Chernavin F.P. (2014) Construction of rating groups of individual borrowers using the method of committees. *Money and Finance*, no 11, pp. 52–54 (in Russian).
20. Nikonov O.I., Chernavin F.P., Medvedeva M.A. (2015) Classification problems: the method of committees. Proceedings of the *XII International Scientific and Practical Conference on the Problems of Economic Development in the Modern World "Sustainable Development of Russian Regions: Economic Policy in the Conditions of External and Internal Challenges", Ekaterinburg, 17–18 April 2015*, pp. 867–874 (in Russian).
21. Nikonov O.I., Medvedeva M.A., Chernavin F.P. (2015) Using the committee machine method to forecasting on the FOREX. Proceedings of the *IEEE 2015 Second International Conference on Mathematics and Computers in Sciences and in Industry (MCSI 2015), Sliema, Malta, 17–19 August 2015*, pp. 240–243.
22. Chernavin F.P. (2018) Application of the method of committees for classification problems. Proceedings of the *XII International Conference "Russian regions in focus of change", Ekaterinburg, 16–18 November 2017.* Ekaterinburg, UPI, part 1, pp. 437–447 (in Russian).
23. Kim J.O., Mueller C.W., Klecka W.R. (1989) *Factor, discriminant and cluster analysis.* Moscow: Finance and Statistics (in Russian).
24. Elder A. (2016) *Trading for a living: Psychology, trading tactics, money management.* Moscow: Alpina Publisher (in Russian).
25. Shiryaev V.I. (2007) *Neural network methods in the analysis of financial markets.* Moscow: KomKniga (in Russian).

About the author

Nikolay P. Chernavin

Junior Researcher, Institute of Economics, the Ural Branch of Russian Academy of Sciences,
29, Moskovskaya Street, Ekaterinburg 620014, Russia;

E-mail: ch_k@mail.ru

ORCID: 0000-0002-2093-9715