Challenges of data access in economic research based on Big Data technology

Chumak V.G., Ramzaev V.M., Khaimovich I.N.

International Market Institute Samara

Abstract. At present the world faces exponential increase of amount and variety of data. The data received as a result of research is used in joint projects of different countries by applying specialized up-to-date tools including Big Data. The present article indicates both the development of method application and data analysis facility, and data management in financial collective areas, it points out unprecedented rapid expansion of spectrum of tasks which require the solution based on data received, accumulation of solving similar tasks experience and making it possible to use it in many disciplines. The article gives applicability of economic data collections in Russia.

Let us consider the problems to access to financial data in Russia. There are two problems here: where to take them from and for which purpose to use. Intensive data use for solving financial tasks has just began in Russia, but in other countries, for instance in the USA it has been already used actively. The projects which are implemented in other countries are worth considering.

On November 11, 2014 in the United States National Institute of Standards and Technology has issued requirements for the financial data collection to use this data intensively in risk analysis and for financial markets development forecasting [1-13]. These requirements, made by NIST standard for Big Data, describe the use of intensive data in the financial sector, description of current solutions, as well as define the objectives and problems of the use of Big Data for financial data.

According to the requirements for financial data which are contained in Big Data, the following information is included:

— bank concerning commercial, retail, credit cards, consumer crediting, corporate banking services, banking operation, trade financing and global payments;
— data in the area of valuable securities and investments, such as retail brokerage services, services to individuals, asset management, institutional brokerage, investment banking, trust banking, safe custody and clearing services;
— data in the area of insurance including personal property accident insurance and collective property accident insurance.

Also these requirements contain one of the approaches to develop an efficient strategy for the financial services industry. When working with large collections of data in Big Data the organization must perform the following steps to define initial status for working with big data on the Internet within cloud systems of the public and private financial institutions offering financial services inside the United States; Great Britain, the EU and China.

In accordance with this strategy, every financial data should include description of people (resources), processes (time / cost / return on investment), technology (various operating systems, platforms), and statutory regulation (depending on the different and several controlling bodies).

Further these data must be identified, analyzed, evaluated, reviewed, tested and revised to in the processes of:

1. project initiation;
2. risk assessment;
3. analysis of the impact on business;
4. development and testing the strategies ensuring the business continuity;
5. emergency response;
6. development and implementation the plans ensuring continuity of financial projects;
7. carrying out of trainings.

Currently big data with hadoop methodology in cloud technology on the Internet is used as a part of hybrid systems as a tool for carrying out of the risk and fraud analysis, apart from assistance in process structuring ("know your customer"). These are three areas, where intensive data from Big Data has given good account.

At the same time, the traditional client/server/data bank/RDBM - systems are used for handling, processing, storage and archival subdivision of financial data. Recently, the financial community has approved the initiative for the presentation of the financial statements in xbrl (Extensible Business Reporting Language) as of 13th of May, 2013.

Currently, areas of concern associated with the use of intensive data storage on the Internet under the control of Big Data, include the aggregation and storage of data (confidential and other) from a number of sources that could create administrative and managerial problems concerning the following issues:

• access control;
• management/administration;

- data ownership.

However, based on the current analysis, these problems are well known and are being solved at the moment, for example, in R&D (research and development).

Russia does not practice the data exchanging (information received from the Sberbank) to predict changes in financial markets and to analyze the risks in the banking sector, so it is possible to use the American standard, provided that the necessary modifications must be made to adapt to the financial markets of Russia.

Based on these requirements the US and EU carry out their projects in the area of intensive financial data usage by the following organizations:

1. The IFRS, Securities and Markets Working Group, www.xbrl-eu.org is a group of organizations which deals with the issues of security and financial data usage in market sphere.
2. PCAOB http://www.pcaob.org is a nonprofit corporation created by Congress to oversee the audits of joint stock companies in order to protect the interests of investors and further the public interest in the preparation of informative, accurate and independent audit reports.
3. http://www.ey.com/GL/en/Industries/Financial-Services/Insurance - the organization monitors the major regulatory changes in the financial documents, analyzes the behavior of emerging markets, estimate the existing economic uncertainty, predicts consumer expectations.
4. http://www.treasury.gov/resource-center/fin-mkts/Pages/default.aspx is the website of the Treasury Department which is responsible for a wide range of activities such as advising the President on economic and financial issues, encouraging sustainable economic growth, and fostering improved governance in financial institutions. The Department of the Treasury operates and maintains systems that are critical to the nation's financial infrastructure, such as the production of coin and currency, the disbursement of payments to the American public, revenue collection, and the borrowing of funds necessary to run the federal government.
5. CFTC http://www.cftc.org is an organization which deals with insurance and financial issues.
6. SEC http://www.sec.gov is US Securities and Exchange Commission which protects investors, maintains fair, orderly, and efficient markets.
7. FDIC http://www.fdic.gov is the Federal Deposit Insurance Corporation.
8. COSO http://www.coso.org is the Committee of Sponsoring Organizations of the Treadway Commission (COSO) which is dedicated to providing thought leadership through the development of frameworks and guidance on enterprise risk management, internal control and fraud deterrence.
9. isc2 International Information Systems Security Certification Consortium, Inc.: http://www.isc2.org is a consortium which carries out wide ranging studies in the field of business management and financial flows taking into account financial risks analysis for the enterprises.
10. ISACA Information Systems Audit and Control Association: http://www.isca.org is an association which makes financial analytics.

These organizations carry out projects in the US and the EU on the stress testing practices of financial markets, i.e. they assess the stability of financial institutions toward adverse changes on the market and also they assess the systemic risk within the EU system according to special standardized procedures, road map, and generate key proposals of exercises. Also they create aggregate statistical data on the banking sector in EU, credit risk, operational risk, market risk, and supervisory actions and measures starting from 2007 to 2013 in the form of documents posted on the website https://www.eba.europa.eu/supervisory-convergence/supervisory-disclosure/aggregate-statistical-data.

For instance, there is some data related to credit risk in EU banks, which is brought according to stress tests on a real time basis in the form of final tables (see fig. 1).

Table Data related to Credit Risk

Data related to Credit Risk*			2010	2011	2012	2013
Credit Institutions: Own Funds Requirements	Own Funds Requirements Credit Risk % of Total Own Funds Requirements		90,02 %	89,07 %	90,81%	90,45%
Credit Institutions: Distribution by Approach	% Number	Standardised Approach (SA)	91,97 %	92,23 %	92,35%	91,42%
		Foundation IRB Approach (FIRB)	7,63 %	7,37 %	7,38%	8,42%
		Advanced IRB Approach (AIRB)	0,39 %	0,40 %	0,27%	0,15%
	Own Funds Requirements % of Own Funds Requirements Credit Risk**	Standardised Approach (SA)	68,77 %	69,40 %	71,52%	69,92%
		Foundation IRB Approach (FIRB)	18,76 %	18,27 %	15,07%	16,29%
		Advanced IRB Approach (AIRB)	12,47 %	12,34 %	13,41%	13,79%
Credit Institutions: Distribution by IRB Exposure Class	Exposures % of Risk Weighted Assets	Central Governments & Central Banks	2,85 %	1,30 %	2,09%	1,13%
		Institutions	17,53 %	21,22 %	21,02%	16,62%
		Corporate	37,64 %	39,70 %	37,37%	42,78%
		Retail	19,90 %	14,15 %	14,55%	17,69%
		Equity	6,43 %	7,43 %	7,53%	8,17%
		Securitisation positions	2,13 %	1,63 %	1,07%	1,65%
		Other Non Credit Obligation Assets	0,12 %	0,06 %	0,01%	0,17%

[data as per 31.12]

* Data provided by the OeNB.
** If an institution uses more than one approach, it will be counted accordingly.

Fig. 1. – An example of credit risks assessment in EU as of 2013, carried out according to intensive data collection

Also these organizations conduct the registry of EU credit institutions. It does not contain any data on Russian financial market and credit institutions, but this data can be useful for investment projects of joint financing, as for example in rocket and space branch in the design and One of the largest financial projects in the USA with data-intensive in Big Data is the project of prediction of systemic risks of crediting in banking system [14-16]. The initial data for this project in the USA are contained in the documents on joint credits, posted on the website http://www.sec.gov/.

In Russia and in the regions the prediction of systemic risks of crediting (MIDAS)in today's complicated economic conditions is extremely vital. It is possible to define the goals of such project:

1) To define the most important financial institutions in case of loss of liquidity when the economic system evolves into an uncontrollable state;

2) To define the companies which most likely depend on credit resources (including external credit);

3) To define the effects of changes in the parameters and indicators of the financial system (the key rate, the change in exchange rates, changes in the value of bonds, etc.);

4) To conduct stress tests to define stability of entire system on the whole.

The initial data for these projects may be: requirements of SEC regulator (the analogue is former Federal Commission for Securities Market, now it is Central Bank

of Russian Federation), which oblige the issuers of securities to disclose their reporting data. This reporting data is made with application of GAAP standard, which is quite similar to IFRS and it allows to tell about the definite structure of data, which easily fits into the system of intensive processing of data. In Russia the requirement of Central Bank of Russian Federation to disclose the reporting data is applied to the issuers of securities, crediting institutions etc.

Using intensive financial data may take place, for example, according to the following algorithm: 1) forming a graph of joint crediting; 2) analysis of critical hubs; 3) visualization of corporate relations with other companies and key personalities.

Based on initial data the graph of joint crediting is formed. The graph nodes are the organizations. Two nodes are linked together by the edge if the relevant organizations present a joint credit (may be with other organizations). Edge has a weight depending on the amount of joint credits (the sum of money) the couple of organizations have. For each graph node normalized centrality of proper vector (the centrality is a measure of the graph node importance) is calculated. By such centrality measure, the contribution to the node importance is provided by the important nodes, which are connected. Nodes with the highest level of centrality with high probability are extreme hubs for the network of joint crediting. Extreme hubs then are the subject to a more detailed analysis:

- Relations with other companies (affiliated companies, competitors, credited parties etc.);
- management (taken positions and dates for each person, control over other companies);
- aggregated financial data (borrowing, investment etc.).

As the result America succeeded in creation relational graph of certain banks with other companies and banks, as well as in distinguishing key ventures in banking system. The examples of relations with other companies and key people for Citigroup are shown in figures 2 and 3.
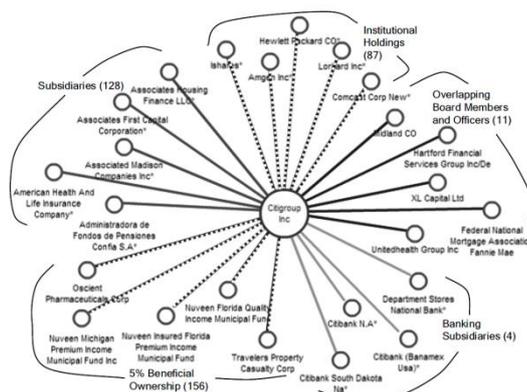

Fig. 2. – The example of connection between the Citigroup bank with other banks
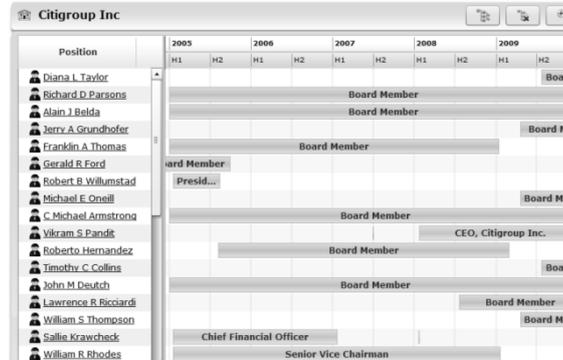
Fig. 3. – The defining of the key people for Citigroup bank within the MIDAS project

As a result, it can be concluded that the intensive use of financial data with the use of Big Data in Russia will allow us to solve problems of joint crediting in the banking sector and to predict the risks in time and develop coordinated management decisions before the crisis. Correct and timely decisions will avoid synergistic effects on the financial market and make the market more predictable and stable.

You can also use the BIG DATA technology in public and municipal administration of the territory on the basis of competitiveness. Competitiveness is an essential characteristic of social and economic systems, including territories. For several years a research on competitiveness of areas, such as region, urban districts, including small and single-industry towns, municipal districts and rural settlements was carried out.

Our approach is based on an understanding of competitiveness as the ability to compete in the competition for scarce resources [17,18].

The basis of the methodology that we developed is the economical and mathematical model of the additive type to assess the territory competitiveness state:

$$\begin{cases} KS = (\xi_1 GF + \xi_2 PRF + \xi_3 EF + \xi_4 PPF + \xi_5 APF + \xi_6 SF + \\ + \xi_7 FEF + \xi_8 IfF + \xi_9 UVF + \xi_{10} IF + \xi_{11} InF + \xi_{12} DF) \\ 0 \le \xi_i \le 1, \ i = \overline{1, 12} \\ \sum\limits_{i=1}^{12} \xi_i = 1 \\ 0 \le GF \le 1; \ 0 \le PRF \le 3; \ -2 \le EF \le 1; \ -3 \le PPF \le 12; \ 0 \le APF \le 6; \ -3 \le SF \le 29; \\ \le 0 FEF \le 11; \ -2 \le IfF \le 13; \ 0 \le UVF \le 1; \ 0 \le IF \le 2; \ 0 \le InF \le 3; \ 0 \le DF \le 5 \end{cases}$$

where KS – competitiveness; GF – geographic factor; PRF – natural-recourse factor; EF – ecological factor; PPF – industrial production factor; APF – agricultural production factor; SF – social factor; FEF – financial and economic factor; IfF – infrastructural factor; UVF – factor of coordination level with higher public authorities; IF – innovative factor; InF – investment factor; DF – spiritual factor; ξ – coefficient of factor importance (is defined from the experts opinion).

The study highlighted 12 factors of competitiveness specific to the current level of socio-economic development of territories. Each factor has its own importance, which determines its weight, contribution to the total value of competitiveness. Importance of factors is different for different types of territories, reflecting the differentiation in the current state of development.

As for the management decision making, realizability and visibility of information are important, we propose a multi-dimensional visualizing results of the analysis and evaluation of competitiveness. By selecting the dimensions of the space it is possible to illustrate the level and the contribution of different factors of competitiveness to management purposes.

Moreover, it is obvious that a sustainable competitive economic development cannot be endless. For example, according to one theory it is limited by energy and natural resources of the planet, and we are already coming to the point of maximum use. There are other theories, according to which the development curve is approaching to the saturation range.

Despite this, competition or rivalry is still the most important factor which is stimulating the development and qualitative growth of socio-economic systems. Accordingly, the application of the above mentioned state models for the management, in practice based on extensive direct addition of factors components, has its limitations. In the area close to saturation, such models are characterized by large inaccuracy or are inadequate.

In consideration of the foregoing, we propose a more precise methodology for the control of competitive development, which is based on correlations of competitiveness factors that allow defining control vectors using key target parameters. Part of this approach is similar to the method of evaluation of a synergistic effect. However, the latter is quite complicated in terms of numerical values.

During research 6 groups of correlated factors were revealed.

From a practical standpoint, the factor grouping allows the most effectively control the competitiveness, as the maximum increase of the level of competitive development can only be obtained through co-regulatory factors within groups.

Analyzing the correlations it was found that the only factor that correlated with all the others is an investment. Therefore, the target function of management in terms of increase of competitiveness is supplemented by model of control action, which is defined as a limited resource investment:

$$
\begin{cases}
\Delta_{KS} = 0,058(GF + \Delta GF(L)) \times 0,072(PRF + \Delta PRF(L)) \times 0,064(EF + \Delta EF(L)) \times \\
\times 0,11(PPF + \Delta PPF(L)) \times 0,075(APF + \Delta APF(L)) \times 0,115(SF + \Delta SF(L)) \times \\
\times 0,113(FEF + \Delta FEF(L)) \times 0,076(IfF + \Delta IfF(L)) \times 0,057(UVF + \Delta UVF(L)) \times \\
\times 0,101(IF + \Delta IF(L)) \times 0,104(InF + \Delta InF(L)) \times 0,055(DF + \Delta DF(L)) \to \max \\
\Delta...F = ...F(L_m(\bar{\phi})) - (...F) \\
L_m(\bar{\phi}) = \sum_{n=1}^{N} \sum_{m=1}^{M} \dfrac{PV_n^{\inf}}{(1+r^{\inf})^n} \cdot \dfrac{1}{IR} \cdot \phi_{im} \\
m = \overline{1,\ 12} \\
i = \overline{1,\ 12} \\
0 \le GF \le 1,\ 0 \le PRF \le 3,\ -2 \le EF \le 1,\ -3 \le PPF \le 12,\ 0 \le APF \le 6,\ -3 \le SF \le 29, \\
0 \le FEF \le 1,\ -2 \le IfF \le 13,\ 0 \le UVF \le 1,\ 0 \le IF \le 2,\ 0 \le InF \le 3,\ 0 \le DF \le 5
\end{cases}
$$

where ΔF is a change of discounted effects of each factor of competitiveness in municipal entity.

The models which take into account the correlation interaction, are effectively used for the selection of investment projects of territorial development, including the competitive distribution, because they are focused not only on account of direct financial results, but also correlated indirect effects of competitiveness increase and its individual factors. For example, environmental, social and others.

Moreover the development of a modern economy is inherently unstable - a sharp acceleration is followed by the deceleration and vice versa.

For example, the GDP decline of the European economy was 6% in 2008-2009, 1.5% in 2011, 0.2% in 2012. In 2013, the EU GDP is forecasted to grow by 0.6%, in 2014 - by 1.2%. The budget deficit of the euro zone has fallen from 4.2% of GDP in 2011 and to 3.7% in 2012. Further decline in the deficit to 2.8% of GDP in 2013 is expected to happen. At the same time the government debt increased to 91% of GDP in 2012 from 87% in 2011. The unemployment rate is from 4 to 27% in different EU countries.

The Russian economy as less stable showed more prominent changes. GDP growth for the period of 2001-2008 was 6.6%. In 2009-2011 there was a drop to 0.2%, and again growth in 2012 to 3.4%. The size of the public debt amounted to minus 9.5% of GDP in 2011 and 3% in 2012. The budget deficit of 4% in 2010, gave way to a surplus of 0.8% in 2011. However, they could not retain positive dynamics and 2012 was with a deficit of 0.02%.

Moreover, these processes occur in a highly intense competitive environment. Therefore, it is obvious that in such a situation those territories, clusters, companies win which are highly sensitive systems - react to changes dynamically and increase their competitive state. Under these conditions, traditional static linear models cannot provide a solution to control tasks, as:

— They are situational and are applicable in short time intervals, it does not allow carrying out long-term and strategic management;
— They ignore the rate of control response;
— Modern complex multidimensional communication and processes are non-linear.

The most effective management of competitiveness in these environments is possible on the basis of dynamic models that have just begun to be used in modern economics.

Thus, we formulate the concept of competitiveness as a dynamic characteristics, defines the application response to any changes in the external socio-economic environment.

Due to the limited volume, we will give the results of dynamic modeling on the example of the social and economic systems of industrial clusters.

Under the proposed new approach, the identified factors of competitiveness, taking into account their mutual correlations, are structured into three dominants: production, labor and capital. In the process of modeling it is necessary to allocate the cluster formation of the overall economic system of the region, i.e., define their boundaries.

Analyzing territorial cluster formation from a position of control, 2 key types are allocated:

1) Functional or managed cluster is that:

— generates as a result of deliberate external influence in strategically important sectors of the state and in the implementation of strategic plans;
— uses the financial, economic and political support of the state;
— as a rule, has a nuclear structure.

2) Self-organizing and entrepreneurial cluster is that:

— is not a product of government agencies;
— generates spontaneously, on the initiative of the business and on the basis of economic relations;
— is not under control and does not have the institutional partners to ensure its survival;
— as a rule, has a matrix structure.

Since the cluster is an open dynamic system, its boundaries are unclear, which led to the use of fuzzy logic and fuzzy sets. Thus the degree of membership is determined by the elements of the system on the basis of necessary and sufficient conditions for the existence of the cluster. The use of fuzzy logic allows to find the area of the intersection of clusters, which are special areas of innovation activity and may provide a breakthrough in the development of cluster systems.

Dynamic models of functional competitiveness and business clusters have a certain look.

Next, the starting point for the management of competitiveness of regional industry clusters of Samara region is determined. For this objective function state model is supplemented by complex of parameters CL that we have developed. The complex CL-parameters includes:
1) Type the cluster according to the criterion of control;
2) Type of cluster development dynamics;
3) The type of the cluster structure;

4) Enterprises producing key products - the leaders of the cluster.

As a result the system characteristics of industrial clusters in the region were formed, on the basis of which it is possible to assess the necessity and the degree of control action and apply the appropriate management models

$$\frac{\partial u_i}{\partial t} = c_i u_i + \sum_{j \neq i}^{n} d_{ij} u_j - \sum_{j \neq i}^{n} b_{ij} u_i u_j + D_i \Delta u_i \quad i = \overline{1, n} \ ,$$

where members of the coefficients *dij* describe the dependence of production in the *i* element from production in the other elements of the cluster; members with the coefficients *bij* take into account the competition among manufacturers.

Dynamic modeling of territorial social and economic systems allowed directly:

— to determine the stationary states of the system, being the target results of management;
— to evaluate state variables of the system when one or another of its parameters is changed, i.e., to trace the impact of management;
— to assess the degree of approximation of the actual system state to the predefined target values and choose the most effective motion path in these conditions.

Despite the apparent complexity of the models, the use of application software at the user level allows to simply interpret the results and formulate management decisions.

To bring the territorial system, for example, the industrial cluster to the area of sustainable development of competitiveness, it is necessary to make adjustment of parameters of the cluster system. It was found that some of the parameters are rather inert. These include: the duration of the production cycle, the pace of staff rotation, tax liabilities and etc. Others have higher dynamics. These are extensive labour capacity, number of employed in production, the unit value of manufactured goods, and others. The most effective control parameters are nonlinear, namely intensive labor productivity growth is achieved through innovation and introduction of new technologies, as well as the employee displacement as a result of intense productivity growth.

The use of dynamic models allow to determine the best configuration of controlled parameters, to assess desired degree of the impact, target results and the speed of their achievements, all of which gives competitive advantage in managing the social and economic system.

To develop the models which predict the competitiveness of the territories it is vital to use large amounts of streaming data on a real time basis. The purpose of this research is to develop models and methods for making management decisions on the basis of forecasting the competitiveness of the territories. The objectives of this study are: to identify factors of competitiveness, to develop a model of competitiveness of the territories using expert estimates, to form information on experts with help of BIG DATA technology. The results of the study include management decision making models of competitiveness of the territories using expert assessments and BIG DATA technology. Practical results include improving the quality and timeliness of decision making on territory management on the basis of forecasting models of the area.

References

1. DRAFT NIST Big Data Interoperability Framework. Use Cases and General Requirements. National Institute of Standards and Technology, 2014; 4.
2. Big Data is a Big Deal. The White House, Office of Science and Technology Policy. Source: <http://www.whitehouse.gov/blog/2012/03/29/big-data-big-deal>.
3. **Shantenu Jha, Judy Qiu, Andre Luckow, Pradeep Mantha, and Geoffrey C. Fox**. A Tale of Two Data-Intensive Approaches: Applications, Architectures and Infrastructure. 3rd International IEEE Congress on Big Data Application and Experience Track, 2014.
4. **Judy Qiu, Shantenu Jha, Andre Luckow, and Geoffrey C.Fox**. Towards HPC-ABDS: An Initial High-Performance Big Data Stack, 2014.
5. **Geoffrey Fox, Judy Qiu, and Shantenu Jha.** High Performance High Functionality Big Data Software Stack. Big Data and Extreme-scale Computing (BDEC), 2014.
6. Source: <http://www.exascale.org/bdec/sites/www.exascale.org.bdec/files/whitepapers/ fox.pdf.>.
7. **Geoffrey C.Fox, Shantenu Jha, Judy Qiu, and Andre Luckow.** Towards an Understanding of Facets and Exemplars of Big Data Applications, 2014. Source: <http://grids.ucs.indiana.edu/ptliupages/publications/OgrePaperv9.pdf>.
8. **Geoffrey Fox and Wo Chang.** Big Data Use Cases and Requirements, 2014. Source: <http://grids.ucs.indiana.edu/ptliupages/publications/NISTUseCase.pdf>.
9. **Geoffrey Fox.** INFO 590 Indiana University Online Class. Big Data Open Source Software and Project, 2014.
10. Source: <wttp://www.whitehouse.gov/mgi>.
11. Source: <http://www.whitehouse.gov/open>.
12. NIST Integrated Knowledge EditorialNet (NIKE). Source: <http://xpdb.nist.gov /nike/term.pl>.
13. Source: <https://rd-alliance.org/group/metadata-standards-directory-working-group.html>.
14. **Balakrishnan S, Chu V, Hernandez MA, Ho H, Krishnamurthy R, Liu S, Pieper J, Pierce JS, Popa L, Robson C, Shi L, Stanoi IR, Ting EL, Vaithyanathan S, Yang H.** Midas: Integrating PublicFinancial Data. SIGMOD, 2010; 1187-1190.
15. **Arasu A, Re C, Suciu D.** Large-Scale Deduplication with Using Dedupalog. ICDE, 2009; 952-963.
16. **Chiticariu L, Li Y, Raghavan S, Reiss F.** Enterprise Information Extraction: Recent Developments and Open Challenges. SIGMOD, 2010; 1257-1258.
17. **Ramzaev VM, Khaimovich IN, Chumak PV.** Management of investment projects during energy modernization of enterprises in the region. Economic sciences, 2013; 4(101): 109-113.
18. **Ramzaev VM, Khaimovich IN, Chumak PV.** Models and methods to manage energy efficiency in the organizations taking into account the limitation of investment resources. Modern problems of science and education4. Source: <http://www.science-education.ru/110-9960>.