УДК 004.05(075.8)
Луцький національний технічний університет[1]
Університет Бєльсько-Бяли, Польща[2]
І.Є. Андрущак[1], В.А. Кошелюк[1], О.М. Сіваковська[1], М.І. Потейчук[1], В.П. Марценюк[2]

PECULIARITIES OF TA INSTRUMENTAL BARRIERS OF STRAINS TECHNOLOGY DATA MINING.

І.Є. Андрущак, В.А. Кошелюк, О.М. Сіваковська, М.І. Потейчук, В.П. Марценюк. Особливості та інструментальні засоби застосування технології DATA MINING. У цій статті докладно розглядаються методи, інструментальні засоби та застосування Data Mining. Викладаються основні концепції сховищ даних і місця Data Mining в їх архітектурі. Обговорюється процес аналізу даних за допомогою технології Data Mining. Докладно розглядаються етапи цього процесу. Аналізується ринок аналітичного програмного забезпечення, описуються продукти від провідних виробників Data Mining, обговорюються їх можливості.
Ключові слова: Data Mining, комбіновані методи, методи обмеженого перебору, Web-mining.

И.Е. Андрущак, В.А. Кошелюк, М.И. Потейчук, А.Н. Сиваковська, В.П. Марценюк. Особенности и инструментальные средства применения технологии DATA MINING.В данной статье подробно рассматриваются методы, инструментальные средства и применение Data Mining. Излагаются основные концепции хранилищ данных и места Data Mining в их архитектуре. Обсуждается процесс анализа данных с помощью технологии Data Mining. Подробно рассматриваются этапы этого процесса. Анализируется рынок аналитического программного обеспечения, описываются продукты от ведущих производителей Data Mining, обсуждаются их возможности.
Ключевые слова: Data Mining, комбинированные методы, методы ограниченного перебора, Web-mining.

I. Andrushchak, V. Koshelyuk, M. Poteychuk, O. Sivakovskaya, V. Martsenyuk. Peculiarities of ta instrumental barriers of strains technology DATA MINING. This article discusses in detail the methods, tools and application of Data Mining. Outlines the basic concepts of data warehouses and the place of data mining in their architecture. The process of data analysis using Data Mining technology is discussed. Details are considered stages of this process. Analyzed the market for analytical software, describes products from leading manufacturers of Data Mining, discusses their capabilities.
Keywords: Data Mining, combined methods, limited search methods, Web-mining.

Formulation of the problem. Data mining is a powerful new technology with great potential to help companies focus on the most important information in the data they have collected about the behavior of their customers and potential customers. It discovers information within the data that queries and reports can't effectively reveal. Generally, data mining is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases. Data mining consists of more than collection and managing data; it also includes analysis and prediction. People are often do mistakes while analyzing or, possibly, when trying to establish relationships between multiple features. This makes it difficult for them to find solutions to certain problems. Machine learning can often be successfully applied to these problems, improving the efficiency of systems and the designs of machines. There are several applications for Machine Learning (ML), the most significant of which is data mining.

Setting up tasks. External providers of IT outsourcing services work exclusively in the field of information technology and, due to narrow professional specialization, provide high-quality services, the cost of which is lower than the cost of using their own IT services. The IT outsourcer has a lot of experience in solving various problems that he faced with his clients. That is, there is a base of problem situations and methods for their possible solutions. In addition, the outsourcing company takes on the implementation of processes that divert people and resources from performing the basic functions. Thanks to IT outsourcing, the company can significantly reduce the cost of owning its own information system. Thanks to IT outsourcing, it is possible to optimize the distribution of all company assets. At the same time, the contract concluded with the IT outsourcing company is a reliable guarantee that the computer system will function properly, and all the problems will be eliminated promptly. In addition, all new hardware and software components will be implemented competently and qualitatively. Due to the fact that the world around us is unstable, and the reaction to various business signals should be lightning fast - IT outsourcing has become quite popular. This type of outsourcing is very popular today among quite young companies that are actively developing, and who do not have the desire to expand their own staff of IT specialists [1].

Basic material presentation. Mining information on the World Wide Web is a huge application area. Search engine companies examine the hyperlinks in web pages to come up with a measure of "prestige" for each web page and website. Dictionaries define prestige as "high standing achieved through success or influence." A metric called PageRank, introduced by the founders of Google and used in various guises by other search engine developers too, attempts to measure the standing of a web page. The more pages that link to your website, the higher its prestige. And prestige is greater if the pages that link in have high prestige themselves. The definition sounds circular, but it can be made to work. Search engines use PageRank (among other things) to sort web pages into order before displaying the result of your search.

Another way in which search engines tackle the problem of how to rank web pages is to use machine learning based on a training set of example queries - documents that contain the terms in the query and human judgments about how relevant the documents are to that query. Then a learning algorithm analyzes this training data and comes up with a way to predict the relevance judgments for any document and query. For each document a set of feature values is calculated that depend on the query term, whether it occurs in the title tag, whether it occurs in the document's URL, how often it occurs in the document itself, and how often it appears in the anchor text of hyperlinks that point to this document. For multiterm queries, features include how often two different terms appear close together in the document, and so on. There are many possible features: typical algorithms for learning ranks use hundreds or thousands of them. Search engines mine the content of the Web. They also mine the content of your queries - the terms you search for - to select advertisement that you might be interested in. They have a strong incentive to do this accurately, because they only get paid by advertisers when users click on their links. Search engine companies mine your very clicks, because knowledge of which results you click on can be used to improve the search next time Online booksellers mine the purchasing database to come up with recommendations such as "users who bought this book also bought these ones"; again they have a strong incentive to present you with compelling, personalized choices. Movie sites recommend movies based on your previous choices and other people's choices: they win if they make recommendations that keep customers coming back to their website.

Anomaly detection is the process of finding the patterns in a dataset whose behavior is not normal on expected. These unexpected behaviors are also termed as anomalies or outliers. The anomalies cannot always be categorized as an attack but it can be a surprising behavior which is previously not known. It may or may not be harmful. The anomaly detection provides very significant and critical information in various applications, for example Credit card thefts or identity thefts. When data has to be analyzed in order to find relationship or to predict known or unknown data mining techniques are used. These include clustering, classification and machine based learning techniques.

Anomaly detection can be used to solve problems like the following:

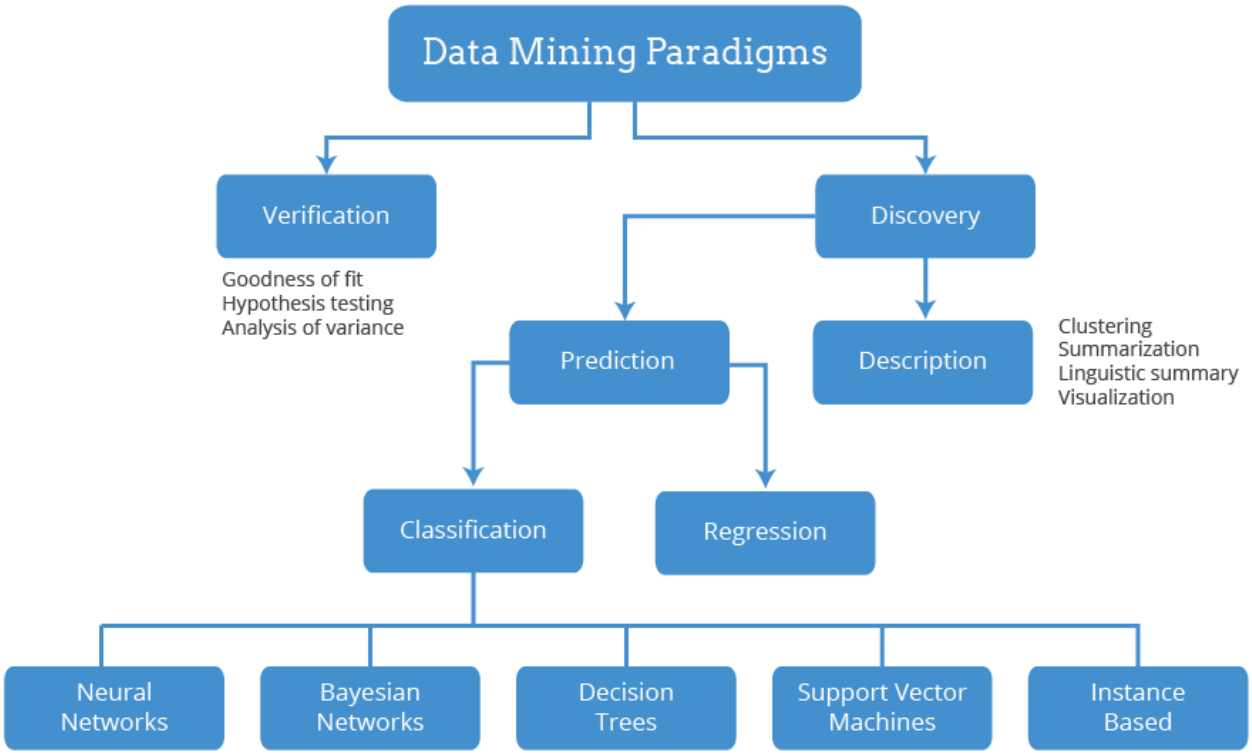
- a law enforcement agency compiles data about illegal activities, but nothing about legitimate activities. How can suspicious activity be flagged? The law enforcement data is all of one class. There are no counter-examples. – Insurance Risk Modeling — An insurance agency processes millions of insurance claims, knowing that a very small number are fraudulent. How can the fraudulent claims be identified? The claims data contains very few counter-examples. They are outliers. Claims are rare but very costly. – Targeted Marketing - Given demographic data about a set of customers, identify customer purchasing behaviour that is significantly different from the norm. Response is typically rare but can be profitable. – Health care fraud, expense report fraud, and tax compliance. – Web mining (Less than 3 % of all people visiting Amazon.com make a purchase). – Churn Analysis. Churn is typically rare but quite costly.

– network intrusion detection. Number of intrusions on the network is typically a very small fraction of the total network traffic. – Credit card fraud detection. Millions of regular transactions are stored, while only a very small percentage corresponds to fraud. – Medical diagnostics. When classifying the pixels in mammogram images, cancerous pixels represent only a very small fraction of the entire image (Pic 1).

Association rule learning is a method for discovering interesting relations between variables in large databases. It is intended to identify strong rules discovered in databases using some measures of interestingness. Based on the concept of strong rules, Rakesh Agrawal introduced association rules for discovering regularities between products in large-scale transaction data recorded by point-of-sale (POS) systems in supermarkets. For example, the rule {onions, potatoes} {burger} found in the sales data of a supermarket would indicate that if a customer buys onions and potatoes together, they are likely to also buy hamburger meat. Such information can be used as the basis for decisions about marketing activities such as, e.g., promotional pricing or product placements. In addition to the above example from market basket analysis association rules are employed today in many application areas including Web usage mining,

intrusion detection, Continuous production, and bioinformatics. In contrast with sequence mining, association rule learning typically does not consider the order of items either within a transaction or across transactions [2].

Clustering - the task of discovering groups and structures in the data that are in some way or another «similar», without using known structures in the data. Clustering is the grouping of a particular set of objects based on their characteristics, aggregating them according to their similarities. Regarding to data mining, this metodology partitions the data implementing a specific join algorithm, most suitable for the desired information analysis.



Pic.1 Data mining paradigms

This clustering analysis allows an object not to be part of a cluster, or strictly belong to it, calling this type of grouping hard partitioning. In the other hand, soft partitioning states that every object belongs to a cluster in a determined degree. More specific divisions can be possible to create like objects belonging to multiple clusters, to force an object to participate in only one cluster or even construct hierarchical trees on group relationships [3].

Classification models predict categorical class labels; and prediction models predict continuous valued functions. For example, we can build a classification model to categorize bank loan applications as either safe or risky, or a prediction model to predict the expenditures in dollars of potential customers on computer equipment given their income and occupation. Popular classification techniques include decision trees and neural networks. Regression - a data mining (machine learning) technique used to fit an equation to a dataset. The simplest form of regression, linear regression, uses the formula of a straight line ($y = mx + b$) and determines the appropriate values for m and b to predict the value of y based upon a given value of x. Advanced techniques, such as multiple regression, allow the use of more than one input variable and allow for the fitting of more complex models, such as a quadratic equation.

Summarization - providing a more compact representation of the data set, including visualization and report generation. Data visualization is a general term that describes any effort to help people understand the significance of data by placing it in a visual context. Patterns, trends and correlations that might go undetected in text-based data can be exposed and recognized easier with data visualization software. Today's data visualization tools go beyond the standard charts and graphs used in Excel spreadsheets, displaying data in more sophisticated ways such as infographics, dials and gauges, geographic maps, sparklines, heat maps,

and detailed bar, pie and fever charts. The images may include interactive capabilities, enabling users to manipulate them or drill into the data for querying and analysis. Indicators designed to alert users when data has been updated or predefined conditions occur can also be included. Most business intelligence software vendors embed data visualization tools into their products, either developing the visualization technology themselves or sourcing it from companies that specialize in visualization.

There are many tools to solve data mining problems. In this paper, we will consider a few of them:

1. RapidMiner. Written in the Java Programming language, this tool offers advanced analytics through template-based frameworks. Offered as a service, rather than a piece of local software, this tool holds top position on the list of data mining tools. In addition to data mining, RapidMiner also provides functionality like data preprocessing and visualization, predictive analytics and statistical modeling, evaluation, and deployment. What makes it even more powerful is that it provides learning schemes, models and algorithms from WEKA and R scripts.

2. WEKA. The original non-Java version of WEKA primarily was developed for analyzing data from the agricultural domain. With the Java-based version, the tool is very sophisticated and used in many different applications including visualization and algorithms for data analysis and predictive modeling. It's free under the GNU General Public License, which is a big plus compared to RapidMiner, because users can customize it however they please. WEKA supports several standard data mining tasks, including data preprocessing, clustering, classification, regression, visualization and feature selection. WEKA would be more powerful with the addition of sequence modeling, which currently is not included.

3. R-Programming. It's a free software programming language and software environment for statistical computing and graphics. The R language is widely used among data miners for developing statistical software and data analysis. Ease of use and extensibility has raised R's popularity substantially in recent years. Besides data mining, it provides statistical and graphical techniques, including linear and nonlinear modeling, classical statistical tests, time-series analysis, classification, clustering, and others.

4. Orange. Python-based, powerful and open source tool for both novices and experts. It also has components for machine learning, add-ons for bioinformatics and text mining. It's packed with features for data analytics.

5. KNIME. Data preprocessing has three main components: extraction, transformation and loading. KNIME does all three. It gives you a graphical user interface to allow for the assembly of nodes for data processing. It is an open source data analytics, reporting and integration platform. KNIME also integrates various components for machine learning and data mining through its modular data pipelining concept and has caught the eye of business intelligence and financial data analysis. Written in Java and based on Eclipse, KNIME is easy to extend and to add plugins. Additional functionalities can be added on the go. Plenty of data integration modules are already included in the core version [4].

6. Scikit-learn Scikit-learn is a free software machine learning library for the Python programming language. It features various classification, regression and clustering algorithms including support vector machines, random forests, gradient boosting, k-means and DBSCAN, and is designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy.

Another advantage of IT outsourcing is the provision of uninterrupted operation of a whole staff of specialists. That is, regardless of the time of the day, calendar holidays or weather conditions, the customer is guaranteed the provision of a qualified specialist to solve the problems encountered. At the same time, an important advantage is that, according to the contract, the outsourcing company constantly diagnoses the operation of the company's IT infrastructure and eliminates the problem. The specialist already knows all the features of the system, which significantly reduces the time spent troubleshooting and helps to prevent them from appearing in the company. further.

Despite the fact that IT outsourcing can take a variety of forms, there are three main ones:

- resource outsourcing. In our country, this type of IT outsourcing is more common today. With this option, the client uses and manages external IT resources. However, he carries all the risks that are associated with the result of his activities.

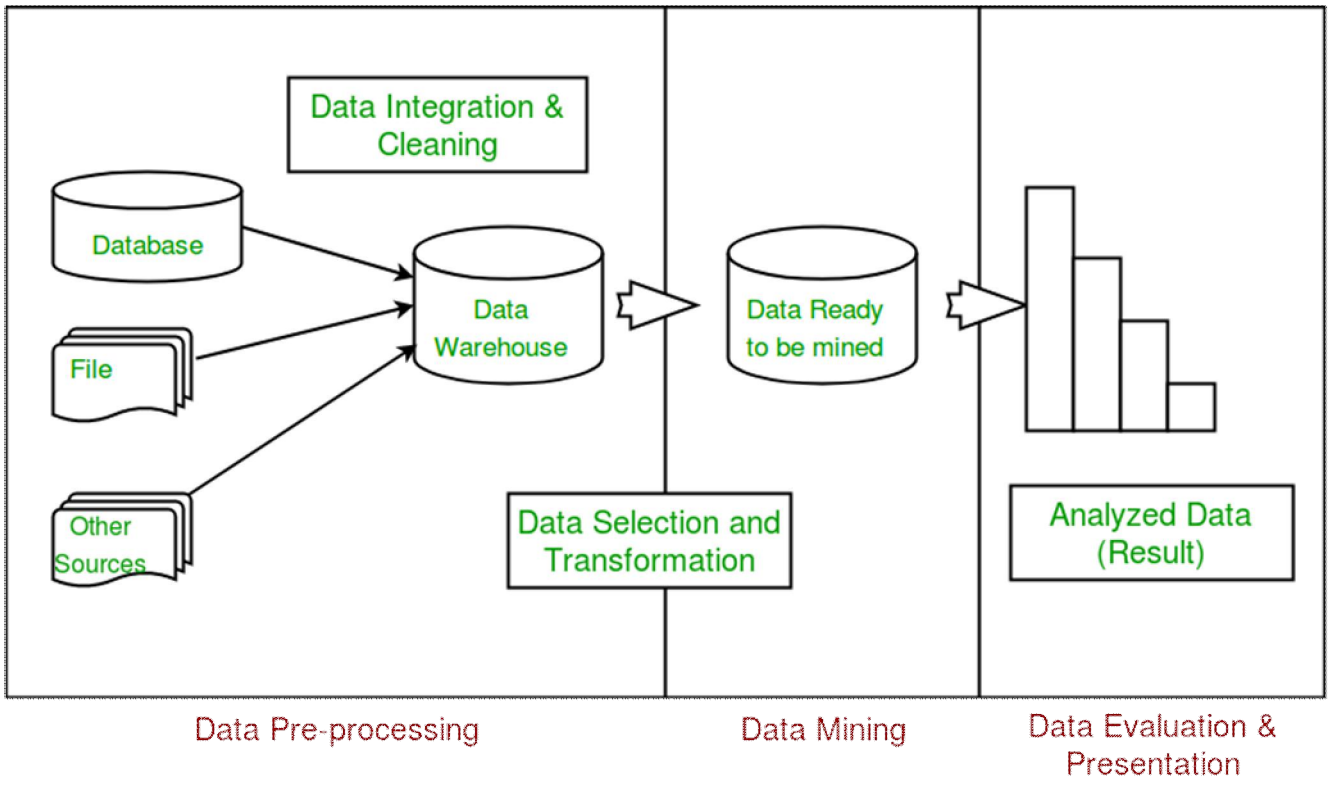
- functional outsourcing. With this option, the outsourcing company transfers the functions. The responsibility of the service provider, as well as the criteria for its operation in this case are stipulated in the agreement. Such criteria may include: the speed of response to various incidents, the frequency of prevention, the recovery time after the failures, etc.

- strategic outsourcing. In this case, a complete transfer to a complex IT outsourcing of the entire IT infrastructure of the enterprise is carried out.

For most enterprises, the main argument for applying to IT outsourcing is budget saving. The hiring of a full-time IT specialist requires a separate workplace, the fulfillment of financial obligations provided for by law and the provision of a social package, which includes both compulsory paid leave and sick leave. This is quite costly for many enterprises. Since problems with the operation of the IT infrastructure can occur at different intervals. Problems can arise or extremely rarely, or require constant support of a specialist, while he may be absent due to leave or illness. While the outsourcing company undertakes to provide competent professionals regardless of external circumstances. Another item of expenditure may be the training of an IT specialist. Computer technologies do not stand still and are in constant development. It is almost impossible to keep track of all innovations to one specialist, while IT outsourcing policy is based on constant improvement. Outsourcing companies are interested in the continuous improvement of the qualifications of their employees in different areas of IT technologies [5,6,7].

Involvement of outside support is most often needed for:

- the development of the infrastructure of the enterprise without distracting the staff from ordinary projects.
- reducing costs for maintenance of office equipment and other tasks in 2-3 times.
- improving the efficiency of performing a number of tasks in the field of support of information technologies in the enterprise.
- increase the level of responsibility of employees for the current state of servers, infrastructure, office equipment.



Pic.2 Data mining stage

Also IT outsourcing helps to reduce the cost of maintaining the infrastructure by 30-50%. For example, it is not profitable for a company to hire a full-time system administrator, because it will often hurt, work long, or vice versa - do it too quickly, and therefore most of the working time will not do anything.

We will admit, in office there are only 5 computers and 20 units of office equipment. To handle at every failure to the masters or to the service centers is also unprofitable - private masters take expensive, service centers repair a long time. The best way is to hire an incoming system administrator from a company that provides IT outsourcing: it will monitor the operation of office equipment remotely, and periodically come for physical service. The firm will pay him less than a full-time employee, while forgetting about expensive repairs.

Applying more and more in the work of information technology companies leads to the fact that firms try to use the labor of the most qualified personnel. At the same time, technology is increasingly dependent on the human factor, and therefore does not require constant intervention in its work. Servicing computers and repairing equipment are increasingly outsourced to almost every company in every country.

Many executives believe that it is undesirable to give out to IT-outsourcing important projects with high expected returns. Some believe that it is better to entrust them to full-time specialists. It is not recommended to outsource the maintenance of rare CRM and profile software developed by another team. To transfer ordinary tasks to outsourcing did not bring additional problems, carefully choose a legal entity. Pay attention to reviews - look for them on the Internet in profile forums, thematic sites, special resources. Pay special attention to official letters of thanks sent from partners - the more of them, the better [8-9].

In general, the benefits of data mining come from the ability to uncover hidden patterns and relationships in data that can be used to make predictions that impact businesses. Specific data mining benefits vary depending on the goal and the industry. Sales and marketing departments can mine customer data to improve lead conversion rates or to create one-to-one marketing campaigns. Data mining information on historical sales patterns and customer behaviors can be used to build prediction models for future sales, new products and services. Companies in the financial industry use data mining tools to build risk models and detect fraud. The manufacturing industry uses data mining tools to improve product safety, identify quality issues, manage the supply chain and improve operations.

Statistics show that at the moment IT outsourcing is at the stage of active development. More and more companies and organizations are turning to IT outsourcing companies. As the transfer of functions of IT infrastructure maintenance to highly focused specialists significantly increases the efficiency of the enterprise as a whole. This makes it possible to concentrate precisely on the directions in which the company specializes. Nevertheless, the market of IT services in our country is gaining momentum every year. More and more entrepreneurs and even large firms prefer external contractors. Today, the dependence of business on IT is extremely high. Modern technologies and solutions allow not only to maintain and accelerate existing business processes, but to change the very model of the company's activity on the market, to open new lines of business. The best option for an actively developing business is cooperation with experienced IT outsourcers [10].

Conclusion

This paper has presented different data mining tasks and tools to solve them. As the amount of data is expanding in all areas, it is easier to find a lot of useful knowledge by using data mining methods. As well as, above-mentioned tools will help us to implement data mining techniques in various areas.

1. Alders, R. IT Outsourcing: A Practical Guide / R. Alders; trans. with English. - M.: Alpina Business Books, 2003. - 300 p.
2. Agapov V., Yakovlev S., Pratusevich V. Review and assessment of the prospects for the development of the world and Russian information technology markets [Electronic resource] // URL: http://www.moex.com/n8686/?nt=106.
3. Heywood, J. Bryan Outsourcing: In Search of Competitive Advantages / J. Bryan Heywood; trans. with English. - M.: Publishing house "Williams", 2004. - 176 p.
4. Tyutina M.V. Analysis and prospects for the development of the information technology market [Text] // Innovative economics: materials IV Intern. sci. Conf. (Kazan, October 2017). - Kazan: Beech, 2017. - P. 9-13.
5. Wayle, P. IT governance: the experience of leading companies. How information technology helps to achieve Sunrise results / Peter Weil, Jinn W. Ross; trans. with English. - M.: Alpina Business Books, 2005. - 293 p.
6. Vorobiev K. Yu. Classification of outsourcing from the positions of the managerial approach // Vestnik Kostroma State University. N. A. Nekrasov. Scientific and methodical journal. Volume 19, №4. 2013 - P.53-56.
7. Information technologies of 2017 [Electronic resource] // BIT. Business & Information Technology. - 2017.-No. 01 (64).
8. Extract from the report World Electronic Industries 2012–2017 carried out by DECISION (March 2014), p. 4.
9. Six of the Best Open Source Data Mining Tools // The New Stack. URL: http://thenewstack.io/six-of-the-best-open-source-data-mining-tools/.
10. Artykov ME, Kurbanova O. U. A review on data mining tasks and tools // Young Scientist. - 2016 - №9.5. - P. 17-20. - URL https://moluch.ru/archive/113/29760.