УДК 004.8; 681.3.06-519.852.6

І.Є. Андрущак[1], Ю.Я. Матвіїв[1], А.А. Ящук[1], В.П. Марценюк[2]
[1]Луцький національний технічний університет, Україна
[2]Академія технічно-гуманістична в Бяльско-Бялі, Польща

ОСОБЛИВОСТІ ТЕХНОЛОГІЇ GRID ТА ІНТЕЛЕКТУАЛЬНОЇ ОБРОБКИ ДАНИХ DATA MINING

У статті розглядаються відмінності між Data Mining і класичними статистичними методами аналізу та OLAP-систем, а також обговорюються типи закономірностей, що виявлені в Data Mining. Описано обсяг застосування програми Data Mining та приклад системи ADAM що працює в середовищі Grid.

Ключові слова: *асоціація, Data Mining, класифікація, послідовність, кластеризація, прогнозування.*

И.Е. Андрущак[1], Ю.Я. Матвиив[1], А.А. Ящук[1], В.П. Марценюк[2]
[1)]Луцкий национальный технический университет, Украина
[2)]Академия техническо-гуманистическая в Бяльско-Бялей, Польша

ОСОБЕННОСТИ ТЕХНОЛОГИИ GRID И ИНТЕЛЛЕКТУАЛЬНОЙ ОБРАБОТКИ ДАННЫХ DATA MINING

В статье рассматриваются различия между Data Mining и классическими статистическими методами анализа и OLAP-систем, а также обсуждаются типы закономерностей, обнаружены в Data Mining. Описаны объем применения программы Data Mining и пример системы ADAM работающего в среде Grid.

Ключевые слова: *ассоциация, Data Mining, классификация, последовательность, кластеризация, прогнозирование.*

I.Ye. Andrushchak[1], Yu.Ya. Matviiv[1], A.A. Yashchuk[1], V.P. Marcenuyk[2]
[1)]Lutsk National Technical University, Ukraine
[2)]Akademia Techniczno-Humanistyczna w Bielsku-Bialej, Poland

FEATURES OF GRID TECHNOLOGY AND INTELLIGENT TREATMENT OF DATA MINING

The article discusses the differences between Data Mining and the classical statistical methods of analysis and OLAP-systems, and discusses the types of regularities found by Data Mining. The scope of the Data Mining application and an example of the ADAM system operating in the Grid environment are described.

Key words: *association, Data Mining, classification, sequence, clustering, forecasting.*

Formulation of the problem. Recently, the World Data Center branch and the national Grid infrastructure (academic and educational segments) began to operate in Ukraine, so that domestic scientists and specialists can now count on increased data from various industries that are processed in a united network of clusters in the country. The development of methods of recording and storing data has led to a rapid growth of volumes of collected and analyzed information. The data volumes are so significant that people simply can not analyze them on their own, although the need for such an analysis is quite obvious, because in these "raw data" there is a knowledge that can be used in decision-making.

In order to perform an automatic analysis of data, Data Mining is used (knowledge extraction). This is a new technology of intelligent data analysis in order to detect hidden patterns in the form of significant features, correlations, trends and templates. Modern data extraction systems use artificial intelligence-based methods of representation and interpretation, which allows them to find dissimilar in terabyte repositories of not obvious but very valuable information. In fact, we are talking about the fact that in the process of data mining, the system does not retreat from the hypotheses put forward, but offers them on the basis of analysis itself.

Setting up tasks. Traditional mathematical statistics, which for a long time claimed to be the main tool for data analysis, did not meet the emerging problems. The main reason - the concept of averaging by sampling, which leads to operations on fictitious quantities. Methods of mathematical statistics proved to be useful mainly for testing pre-formulated hypotheses and for "rough intelligence analysis," which forms the basis of operational analytical processing of OLAP data.

Data Mining is a multidisciplinary field that arose and develops based on the achievements of applied statistics, image recognition, methods of artificial intelligence, database theory. Hence the abundance of methods and algorithms implemented in various existing Data Mining systems. Many of these systems integrate at once several approaches. However, as a rule, each system has a key component on which the main rate is made.

Data Mining is a collection of many different methods of acquiring knowledge. The choice of method often depends on the type of data available and on what information is to be obtained.

The basis of modern Data Mining technology is the concept of templates (patterns), which reflect fragments of multidimensional relationships in the data. These patterns are regularities inherent in sub-types of data, which can be compactly expressed in a form that is understandable to a person. The search for templates is carried out using methods that are not bounded by the framework [1].

Analysis of recent researches and publications. The basis of modern technology Data Mining is the concept of patterns that represent the fragments of multidimensional data relationships. These patterns are regularities inherent in sub-types of data, which can be compactly expressed in a form that is understandable to a person. The search for stem cells is carried out using methods not limited by the a priori assumptions about the structure of the sample and the type of distribution of the values of the analyzed parameters.

It should be immediately determined that the scope of use of Data Mining is unlimited - it is everywhere where there is any data. There are two areas in which Data Mining can be used: as a mass product and as a tool for unique research. Now Data Mining technology is used in virtually all areas of human activity, where retrospective data is accumulated. Let's consider four main areas of application of Data Mining technology in more detail: science, business, retail and Web-direction [1,5,7].

Data mining methods can be used both for working with large data, and for processing relatively small amounts of data (obtained, for example, from the results of individual experiments, or when analyzing data on the company's activities). As a criterion for a sufficient amount of data, both the research area and the applied analysis algorithm are considered.

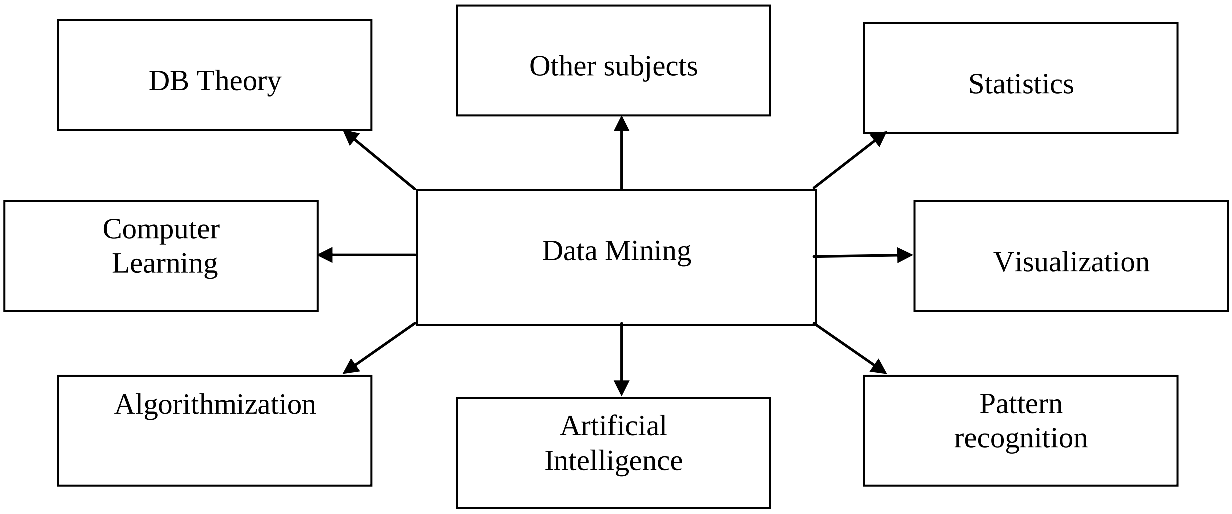
The development of database technologies first led to the creation of a specialized language - the language of queries to databases. For relational databases, this is SQL language, which provided ample opportunities for creating, changing and retrieving stored data. Then, it became necessary to obtain analytical information (for example, information on the activity of the enterprise for a certain period), and it turned out that traditional relational databases, well adapted, for example, to keep the company operational, are poorly suited for analysis. This led, in turn, to the creation of so-called. "Data warehouses", the very structure of which is the best way to conduct comprehensive mathematical analysis.

The traditional methods of data analysis (statistical methods) and on-line analytical processing of OLAP (OnLine Analytical Processing) are mainly aimed at verifying the verified driven data mining and on the "rough" intelligence analysis that forms the basis of OLAP at that time as one of the main postulates of Data Mining is the search for non-obvious laws. Data Mining tools can find these patterns independently and also independently build hypotheses about interconnections. Since it is precisely the formulation of the dependency hypothesis that is the most difficult task, the advantage of Data Mining in comparison with other methods of analysis is obvious.

Most statistical methods for identifying relationships in data use the concept of averaging by sampling, which results in operations on non-existent values, while Data Mining operates with real values.

OLAP is more suitable for understanding retrospective data, Data Mining is based on retrospective data to answer questions about the future.

Basic material presentation. The algorithms used in Data Mining require a large number of computations. Previously, this was a deterrent to the widespread use of Data Mining, but today's growth in performance of modern processors has eliminated the severity of this problem. Now for a reasonable time, you can conduct a qualitative analysis of hundreds of thousands and millions of records. Data Mining is an interdisciplinary industry that originated and developed on the basis of such sciences as applied statistics, image recognition, artificial intelligence, database theory, etc. (Dr. 1 [1]).



Dr. 1. **Data Mining as an interdisciplinary field**

Data Mining's potential gives a "green light" to expanding the scope of this technology. As for the prospects for Data Mining, the following development directions are possible [2]:

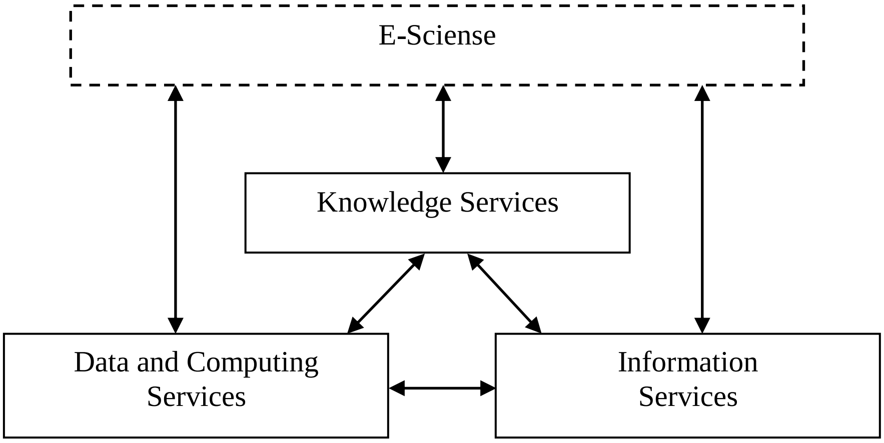
- selection of types of subject fields with their heuristics, formalization of which will facilitate the decision of the relevant Data Mining tasks related to these industries;

- creation of formal languages and logical means by which formalizing considerations and automation of which will become a tool for solving Data Mining tasks in specific subject areas;

- creation of methods of Data Mining, capable not only to "extract" from the data of the laws, but also to form some theories, which are based on empirical data;

- overcoming the significant lagging of the capabilities of the toolkit Data Mining from the theoretical advances in this area.

Looking at the future of Data Mining in the short term, it is obvious that the development of this technology is mostly directed at the industry related to Grid-systems for E-Science [3]. E-Science features characterize the computing infrastructure, which consists of three conceptual levels (Dr. 2).



Dr 2. **Three-level architecture of Grid-services**

1. **Data / computing services.** This level contains information about the location of the computing resources allocated to the calculation, and on the means of transferring data between different computing resources. It can process large volumes of data, providing fast networks, and provide diverse resources as a single metacomputer [4].

2. **Information services.** Indicates how information is transmitted stored, who has access to it. Here the information acts as data with value. For example, the identification of an integer as the temperature of the reaction process, the recognition that the string is the name of a person.

3. **Knowledge services.** Provides a way in which knowledge is acquired, used, found, published, to help users achieve their specific goals. Here, knowledge is provided as information used to achieve a

goal, solve a problem or make a decision.

The concepts considered are an integral part of the so-called information pyramid, which is based on data, the next level - information, then goes the solution, completes the pyramid level of knowledge.

Grid systems that are already built or built will contain some elements of all three levels. The degree of importance of using these levels will be determined by the user. Thus, in some cases, the processing of huge amounts of data will be a dominant task, while in other cases, the maintenance of knowledge - the main problem. So far, most of the Grid research work has focused on the level of data / computing and on the information level. At the same time, there are still many unresolved issues relating to the management of large-scale distributed computing and the effective access and dissemination of information from heterogeneous sources. It is believed that the full potential of Grid computing can only be acquired through long-term exploitation of the functionality and capabilities provided by the level of knowledge. Therefore, this level is required for automated direct simple access to operations and interactions. It's time for scientists and engineers to oppose Data Mining as an instrument for research (genetics, chemistry, medicine, nanotechnology). Developers of the national Grid infrastructure of Ukraine link the future of Data Mining with their use as Grid of intelligent applications embedded in virtual or corporate data warehouses, as well as in the network of World Data Centers [5-6].

The Data Mining Potential gives "green light" to extend the scope of technology. As for the prospects of Data Mining, the following development areas are possible:

- allocation of types of subject areas with corresponding heuristics, formalization of which will facilitate the corresponding Data Mining tasks related to these areas;
- creation of formal languages and logical means by which formalizing considerations and automation of which will become a tool for solving Data Mining tasks in specific subject areas;
- creation of Data Mining methods, capable not only of obtaining data from regularities, but also to form certain theories based on empirical data.
- to overcome the essential lagging of the capabilities of Data Mining toolkits on theoretical achievements in this field [7].

Looking at the future of Data Mining in the short term, it's obvious that the development of this technology is most directed to areas related to business. In the short term, Data Mining products can become as common and necessary as e-mail, and for example, be used by users to find the lowest prices for a particular product or the cheapest ticket.

In the long run, the future of Data Mining is really exciting - it may be the search for intelligent agents as new types of treatment for various diseases, as well as a new understanding of the nature of the universe.

However, Data Mining also has a potential danger - as more and more information becomes available on the Internet, including information of a private nature, and more and more knowledge can be obtained from it.

Before using data mining algorithms, it is necessary to prepare a set of analyzed data. Since the IAP can detect only the patterns that are present in these data, the initial data on one side must have a sufficient volume so that these patterns are present in them, and on the other - be compact enough that the analysis takes an acceptable time. Most often, data warehouses or data marts are used as input data. Preparation is required for the analysis of multidimensional data prior to clustering or data mining.

Then the data is cleared. Cleaning removes samples with noise and missing data. The cleared data are reduced to feature sets (or vectors, if the algorithm can only work with vectors of fixed dimension), one set of characteristics for observation. A set of attributes is formed in accordance with the hypotheses about which characteristics of raw data have a high predictive power in relation to the required processing power for processing. For example, a black and white face image of 100 × 100 pixels contains 10,000 bits of raw data. They can be transformed into a feature vector by detecting in the image the eyes and mouth. As a result, the data volume decreases from 10 thousand bits to the list of location codes, significantly reducing the amount of data analyzed, and hence the analysis time [8-9].

A number of algorithms are able to process missed data that have predictive power (for example, the customer does not have purchases of a certain type). For example, when using the method of associative rules, English is not processed by attribute vectors, but by sets of variable dimension.

The choice of the objective function will depend on what is the purpose of the analysis; The choice of a "correct" function is fundamental to the successful intellectual analysis of data. Observations fall into two categories: the training set and the test set. The training set is used to "learn" the data mining algorithm, and the test set is used to test the found patterns [10].

Conclusion. An important position in Data Mining is the non-triviality of wanted templates. This means that the found patterns should reflect non-obvious, unexpected (unexpected) regularities in the data, constituting so-called hidden knowledge (hidden knowledge). It has come to the understanding that raw data (raw data) contains a deep layer of knowledge, with a competent excavation of which true nuggets can be found.

The scope of Data Mining is unlimited - it's wherever there is any data. But in the first place Data Mining methods were intrigued by commercial enterprises today. The experience of many of these companies shows that the effect of using Data Mining can reach 1000%. For example, the annual savings of the UK supermarket network at the expense of Data Mining's implementation is 700,000. Data Mining is of great value for executives and analysts in their day-to-day business.It's time for scientists and engineers to seize Data Mining as an instrument for research (genetics, chemistry, medicine, nanotechnology, etc.). Developers of the national Grid infrastructure of Ukraine link the future of Data Mining with its use as Grid-intelligence added to the virtual or corporate data warehouses as well as to the World Data Centers network. But an interdisciplinary task requires the integration of the efforts of Ukrainian specialists (maybe within the framework of the relevant state program) that work in universities and academic institutions and are well-known in mathematical methods and have the experience of creating many unique information processing algorithms to create modern Data Mining with broad opportunities.

References

1. Barseghyan F. Methods and models of data analysis of OLAP and DataMining. / Barseghyan F., Kupriyanov M., Stepaneenko V., Kholod I. - SPb BHV-Petersburg, 2008.

2. Chubukova I.A. Data Mining: a Manual. - M .: Internet University of Information Technologies: BINOM: Knowledge Lab, 2006. - 382 p. (http://www.intuit.ru/department/database/datamining/)

3. Data Mining and Image Processing Toolkit. - http: //datamining.itsc.uah. edu / adam /.

4. Dyuk V. Data Mining: the course (+ CD) /. Dyuk V., Samoilenko A. - St. Petersburg: Izd. Peter, 2001. - 368 pp.

5. Knowledge Discovery Through Data Mining: What Is Knowledge Discovery? - Tandem Computers Inc., 1996.

6.Krechetov N. Products for the analysis of data. - // Market of Software, N14-15_97, c. 32-39

7. Kiselev M. Means of obtaining knowledge in business and finance / Kiselev M., Solomatin E. - / / Open Systems, No. 4, 1997, p. 41-44.

8. Pertrenko A.I. Grid and intelligent data mining. / A.I. Pertrenko // System Research & Information Technologies, 2008, No. 4 97-110.

9. Methods and models for data analysis OLAP and Data Mining / F. Barseghyan, M. Kupriyanov, V. Stepanenko, I. Kholod. - SPb .: BHV. - 2008 - 267 pp.

10. Data Mining, Web Mining, Text Mining, and Knowledge Discovery. - http://www.kdnuggets.com.

Стаття надійшла до редакції 20.03.2018