“Mircea cel Batran” Naval Academy Scientific Bulletin, Volume XIV – 2011 – Issue 2
Published by “Mircea cel Batran” Naval Academy Press, Constanta, Romania

CONCEPTS OF DATA MINING AND KNOWLEDGE DISCOVERY IN DATA WAREHOUSES

Loredana-Maria PAUNESCU[1]
[1] Asistant drd., Department of Modeling, Economic Analysis and Statistics, Petroleum&Gas University, Ploiesti, Romania

Abstract: *Data mining tools perform data analysis and reveal major data models to determine the directions of development of various fields (economic, scientific, medical, educational) that are carried by economic bodies to achieve their goals.*
Keywords: *data mining, data warehouse, decisions, models, information.*

1. INTRODUCTION

Rapid and continuous growth of data volumes imposed the emergence and use of data mining tools to extract useful information and knowledge from huge amounts of data collected and stored in large data warehouse.

In essence, data mining tools perform data analysis and reveal major data models to determine the directions of development of various fields (economic, scientific, medical, educational) that are carried by economic bodies to achieve their goals.

1.1. Data mining Concept (data mining)

This concept defines the process of knowledge discovery models and / or useful information from large amounts of data collected and storage depots in different types of data (data warehouse) in order to use them to base management decisions on all component levels from an economic system.

In reality, data mining is an improper expression which associates, in a very suggestive way, the discovery of patterns of knowledge and / or useful information from a large data quantity with the process of extraction of minerals from rocks, in order to make it easier to be understood .

The data mining process is an essential part of the knowledge discovery process because it finds patterns in data as "hidden" data to be assessed, in accordance with the user’s requirements. In terms of its functionality, data mining is the process of knowledge discovery that is of interest from a large amount of data stored in databases or data warehouses (Data Warehouses).

The rapid and continuous growth of data volumes imposed the emergence and use of data mining tools to extract useful information and knowledge from huge amounts of data collected and stoate in large and very large data warehouses.

In their essence, data mining tools makes data analysis and highlight important data models to determine the directions of evolution of various fields (economic, scientific, medical, educational) that develops economic bodies to achieve their goals.

The data mining process interacts with the user through its dedicated knowledge base, the data models found, representing the user with important, in fact, new knowledge is stored in the base of knowledge to be presented.

Therefore, the concepts that characterize data mining are: data, information and knowledge.

1. Data represent the basic material of information. A data has meaning only in a certain context and will be transformed into an information.
2. Information is facts, perceptions or messages that increase the degree of knowledge of a human being in relation with the environment, its features are: clarity, novelty, utility.
3. Knowledge is the summation of all information acquired while in a specific area or on a particular object and can be exploited in decision-making processes.

Flow associated with data conversion is shown in Figure 1.

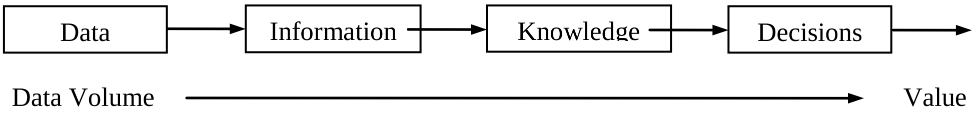

Fig.1. Transforming data into useful knowledge decision making process

The basic requirement for any computer system is to provide the accurate and timely information to all stakeholders of a company, the work towards the systems is accepted by most specialists through the use of computing on two fields: as background in business decisions and to assist current activity carried out within a scaffold [1].

The first level is characterized by the existence of alternatives analysis software, specialized in decision-making, aiming to find the most favorable decisions for the scaffolding at the time of their adoption.

The second plan, aims to develop economic management applications at work or group of activities within the company.

To obtain performance of the assisted business processes, it is necessary that these systems control the environment by receiving data, processing and return their results quickly enough to be able to influence the operation of the processes in real time.

Real time processing system requires immediate entry of data messages transmitted from the source terminals. A large number of stations, and remote system connected via high-speed communication devices can work simultaneously: some updated files, other others involved in queries, etc.. Real-time information systems based on advanced technology solutions: powerful processors, direct link between the computer and electronic communication system, adapted to different terminals; improvements of equipment and communication channels, storage space allocation and preliminary processing network, the development of production techniques ensuring the evaluated software's necessary programming.

1.2. Organization and data processing in data warehouses

Accumulation of large amounts of data necessary for the operation and making correct decisions is not a problem technically. Now the question is extracted from the multitude of existing information most relevant to this issue in a short period of time and a lower cost. Thus was born the idea of structuring information in data warehouses (data warehouse).

The emergence of data warehouse technology is the result of several factors: economic environment, changing business vision, the accumulation of large amounts of data, rapid progress of technology and data processing organization [2].

- the economic environment is increasingly competitive, global and complex and requires information developed to support strategic decisions;
- changing business vision resizing organizations, reengineering business processes and reduce market cycles have forced organizations to accelerate business and to recognize and quickly adopt the change. Strategic planning and decision making have become critical to the extent that organizations aimed at increasing profitability and achieving competitive advantage and need for this constant and continuous information;
- large amounts of data - the success of databases, especially relational led to accumulation of large amounts of data.
- wave of new technologies - in the processing of analytical data significant new technologies have emerged. OLAP products have matured. The software industry is based on data organization in order to support decision making.

The new Web technologies with OLAP products provide a data retrieval much easier.

Data warehouse is one of the latest technologies and data processing organization quickly adopted by companies who understand that the information is a valuable resource and implement a data warehouse that provides solutions to adapt faster to business constantly changing and to obtain competitive advantages.

Data warehouses (data warehouse) are the product of the economic environment and advanced technologies. On the one hand, the economic environment is increasingly competitive, global and complex and requires information developed to support strategic decisions, and on the other hand, developments in information technology provide effective solutions for managing large volumes of data integrated terabytes s order, providing levels of synthesis / adequate detail.

Data warehouse architecture and tools provides executive leadership by organizing systematic understanding and use of data in making strategic decisions. A large number of organizations believe that the data warehouse systems have valuable tools in the economic environment which is competitive and in continuous evolution.

Deposits of information is an acute requirement of modern organizations and also implemented a technological reality ever more frequently.

A definition of data warehouse OLAP made by Council is as follows:

"A data repository (data warehouse) is a centralized storage of detailed data from all relevant sources within an organization and allows for querying dynamic and detailed analysis of all information" [12].

Data structures in a data warehouse is optimized for rapid retrieval and analysis. The data are historical and are updated at regular intervals, depending on the reporting requirements.

The definition of William Inmon, known as the father of this concept is very concise: "a data warehouse is a database-oriented topics, integrated, nonvolatile historical and support for managerial decision-making process."

Large volume of data repositories requires the use of special tools and data mining technologies such as multi-dimensional dynamic analysis, statistical methods of forecasting and mathematical methods applied to a large volume of data.

Statistical analysis of data and knowledge extraction in data warehouses such as the received data mining, or mining the data, a term whose use is obvious. From large volume of data is extracted only the relevant data for decision support, the other being ignored or used for other purposes.

Also, you can see the importance of flexibility required to implement such systems, data warehouses. Here, flexibility is an organization-wide connectivity, so that different database servers can connect simultaneously to the slready existing deposit.

It is also very important to choose an architecture that easily adapt to changes in performance, capacity and connectivity.

Configuration processes, optimization and system administration, including rescue procedures - restoring and maintaining the system functionality that time can become very difficult operations that should be repeated every addition of new servers in the system.

To avoid these problems, you can choose a middle way and you can choose to carry out an in store to contain only relevant data required for analysis. Such deposits are called on in the marts and can be made to work on more modest configurations and resources than data warehouses in a shorter period of time.

Such a data mart is a data warehouse requirements specific to a subset or a specific department within the organization.

While a data warehouse contains data that can be used to answer any questions on certain aspects of a company, a data mart that contains pertinent data os a certain branch of the company.

Connecting together the different data marts of the the company's sites, thus forming a specific infrastructure, departments may share their data and can create a data warehouse built easier and more elastic.

The role of a data warehouse is to provide a coherent picture of data related to work organization and the context in which it acts [10].

The use of this collection may consist of extraction of reports (on demand or with a certain periodicity), extraction of data to be used for office applications (spreadsheet programs, word processors, presentation programs, etc..), especially to be used by specialized applications analysis.

These two categories are based on online analysis technologies (OLAP - On Line Analytical Processing) applications and technologies focused on multidimensional analysis to extract knowledge from data (data mining) focused on the discovery of significant patterns in data collections.

2. Knowledge discovery in data warehouses

Further development of data warehouses imposed the field emergence of new techniques for collecting, storing, processing, retrieval and transmission of data.

Data mining integrates features three main areas of interest, such as: statistics, artificial intelligence and database systems. The basic function in data mining is the extraction of knowledge from data models to the user, combining a variety of statistical algorithms, fuzzy logic, etc..

Data mining is found as a fundamental step in the process of Knowledge Discovery in Databases - KDD, as shown in Figure 2.

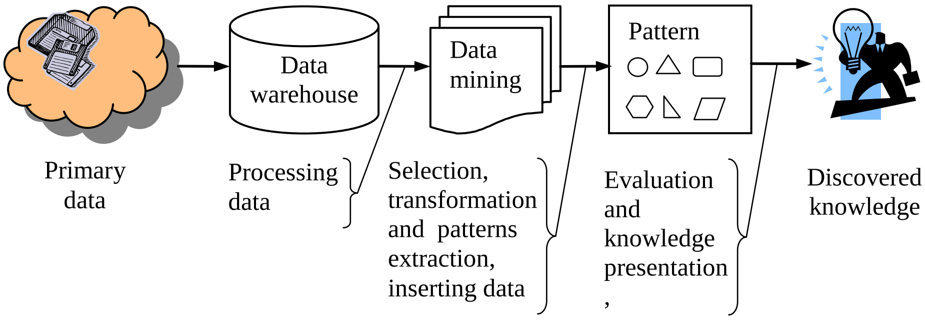
Data warehouse
Data mining
Pattern
Primary data
Processing data
Selection, transformation and patterns extraction, inserting data
Evaluation and knowledge presentation ,
Discovered knowledge

Fig.2. Data mining, a stage of knowledge discovery process

3. CONCLUSIONS

The fundamental steps describe the process of knowledge discovery in data warehouse are: data cleaning, data integration, data selection, data transformation, data mining, pattern evaluation and knowledge presentation.

The problems can be solved with data mining methods require two types of data mining, namely:

- Data mining focuses on finding descriptive patterns describing the data, interpreted by people and which in turn can produce information based on existing data set.
- Data mining which involve the use of predictive variables or fields in the data set to predict unknown or future values of other variables of interest, generating a predictive model.

As descriptive methods are mentioned: clutering, the rules of association, etc.. And the predictive methods: classification, regression, anomaly detection, etc.

REFERENCES :

[1] Adamson, C., *Mastering Data Warehouse Aggregates. Solutions for Star Schema Performance*, Wiley Publishing, Inc., Indinanapolis, 2006

[2] Airinei, D., *Depozite de date*, Editura Polirom, Iaşi, 2003

[3] *Aligulizev, R.,Clustering of document collection – A weighting approach, Expert Systems with Applications, vol. 36, nr. 4, 2009, pg. 7904-7916*

[4] Androne,M., *Analiza datelor stocate în depozite mari de date*, www.spiruharet.ro/sesiuni-comunicari/word/5.5.pdf

[5] Agrawal, R., Srikant, R., *Fast Algorithms for Mining Association Rules in Large Databases,* Research Report RJ 9839, IBM Almaden Research Center, San Jose, California, 1994

[6] *Berry, M.J.A., Linoff, G.S., Data mining techniques: for marketing, sales and customer relationship management, Wiley Publishing, Indianapolis, Indiana, 2004*

[7] Berson, A., Smith, S., Thearling, K. - *Building Data Mining Applications for CRM*, McGraw-Hill Companies, 1999

[8] Berry, M.J.A., Linoff, G., *Data Mining Techniques for Marketing, Sales and Customer Support,* Willey, New York City, 1997

[9] Berry, M.J.A., Linoff, G., *Mastering Data Mining: The Art and Science of Customer Relationship Management,* Canada, Wiley, 2000

[10] Bloodgood, J., Salisbury, D., *Understanding the influence of organizational change strategies on information technology and knowledge management strategies. Decision Support Systems*

[11] Burns, R.S., *Advanced Control Engineering*, Butterworth-Heinemann, 2001

[12] Chakrabarti, S., *Data Mining: know it all,* Morgan Kaufmann know it all series, 2009

[13] *Davidescu, N., Proiectarea Sistemelor Informatice prin limbajul Unified Modeling Language (PSI 2), Editura ALL Beck, Bucureşti, 2003*

[14] Duebel, C., *Data Mining* , Application Techniques to Industrial Processes to Improve Business Performance, www.knowledgeprocesssoftware.com

[15] Edelstein, H.A., *Introduction to Data Mining and Knowledge Discovery*, Third Edition, Two Crows Corporation, 1999

[16] Fayyad, U.M., *From Data Mining to Knowledge Discovery: An Overview, in Advances in Knowledge Discovery and Data Mining,* eds. Fayyad, U.M., Piatetsky-shapiro, G, Smyth, P. Si Uthurusamy, R., Menio Park, CA: AAAI Press, 1996.