EXTRACTING KNOWLEDGE FROM DATA - DATA MINING

DIANA ELENA CODREANU [1]
DENISA ELENA PARPANDEL[2]
IONELA POPA [3]

Abstract
Managers of economic organizations have at their disposal a large volume of information and practically facing an avalanche of information, but they can not operate studying reports containing detailed data volumes without a correlation because of the good an organization may be decided in fractions of time. Thus, to take the best and effective decisions in real time, managers need to have the correct information is presented quickly, in a synthetic way, but relevant to allow for predictions and analysis.
This paper wants to highlight the solutions to extract knowledge from data, namely data mining. With this technology not only has to verify some hypotheses, but aims at discovering new knowledge, so that economic organization to cope with fierce competition in the market.

Keywords: *data mining, date warehouse, database, analysis, decisions, information*

Introduction

Because the present policy makers in an organization is facing an avalanche of data, data systematic data yet unexplored, but also because computing, informatics, information technology has been developed in to an alert, the skyrocketed and the need to use new methods for information discovery, knowledge that is "hidden" in the data, information that can not be "discovered" by traditional means or by using just the human factor.

Data mining (data mining) sometimes called knowledge discovery in databases (Knowledge Discovery in its Bases - KDD) is a new technology, a young technology, technology "growing" but that seems to be on the verge of reaching a "key technology.

As with other technologies there are many understandings of information technology on the definition of this technology but also in terms of the roots, interdisciplinary nature of this technology. This paper wants to emphasize to point out these two issues, to highlight what "borrow" the process of data mining in other disciplines that cross over, but intense relationship between Data Mining and Business Intelligence. This paper aims to discuss this topic because there are now a number of companies that have field activity data mining applications and this is due to increasingly larger data mining services for example, the economic and financial market (we can give examples of areas like Business Intelligence, Business Performance Management - BPM, Customer Relationship Management - CRM and others).

Data mining has been defined as the automatic analysis of large and complex data sets as its main objective the development of new significant patterns or trends without using this technique could remain "unidentified." Because of recent research in data mining have developed more efficient methods FST for finding these new patterns, as well as huge volumes of data into knowledge based on effective methods of classification to clustering, analysis of frequent patterns, sequential or structural.

Basically data mining is one of the most recent technologies for data analysis with OLAP, data warehouse concept, text mining or web mining.

[1] Assistant, Ph. D. candidate, „Constantin Brancoveanu” University, Piteşti (e-mail: codreanudia@yahoo.com).
[2] Assistant, Ph. D., „Constantin Brancoveanu” University, Piteşti (e-mail: parpandeldenisa@yahoo.com).
[3] Lecturer, Ph. D., „Constantin Brancoveanu” University, Piteşti (e-mail: popaionela80@yahoo.com).

The interdisciplinary character of Data Mining Technology

The concept of data mining has emerged in the 80s and in that time and represented the process by which only applied to algorithms for extracting knowledge from large data collections. Year 1994 is another year of reference in terms of data mining technique, so in 1994 the company launched the SPSS data mining program called "Clementine," which proved later to be widely used and widely spread at that time. Another reference date is the data mining industry in 1996 with funds allocated by the Commission shall draw European design methodology CRISP-DM (Cross Industry Standard Process for Data Mining).

Over the years, the concept of data mining has evolved, so has been interpreted in a broader process that is taking place extracting knowledge from databases based on some information requirements and to validate the information obtained. This is the approach that has been accepted more and more lately.

Because now we are facing an avalanche of information in electronic text, data mining has seen a new specialization, namely to help *text mining* automatic extraction of knowledge from text.

Date mining became known in the '90s, when speaking of data mining *or mining the data "in* many environments, whether it is academic whether the business.

In 1997 Pregibon - Research Scientist Google Inc.., Says *that "Data mining is a mixture of Statistics, IA (Intelligence Artificial) and database research"* (D. Pregibon, Data Mining, Statistical Computing and Graphics Newsletter, 7, p.8, 1997).

We say that data mining is an interdisciplinary, as in disciplines such as statistics, database technology and artificial intelligence, has borrowed techniques and terminology work. It is very difficult to define each of these disciplines, but it is equally difficult to delimit the boundaries between data mining and each [one] of them. (Figure)

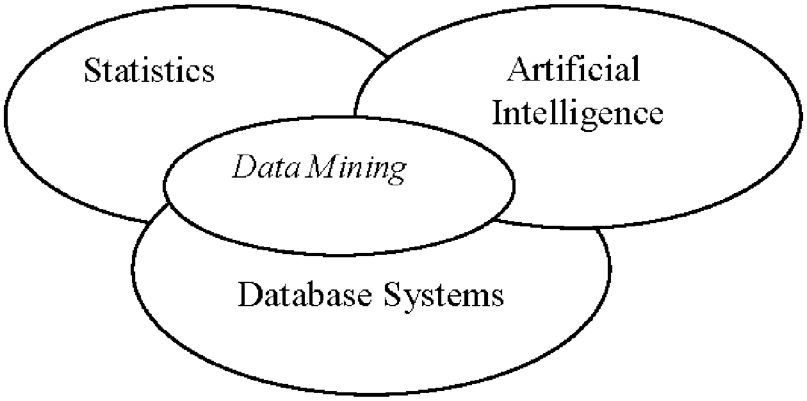

Figure: Data mining interdisciplinary nature of art

Statistics - Data Mining (DM) used in defining statistical techniques that are found in the literature as the *exploratory data analysis (EDA - Exploratory Data Analysis).* EDA is used to establish the systematic relationship between different variables in a situation where we have sufficient information on their nature. Among the best known and EDA techniques can be used to name [two:]

- *Computational methods:* descriptive statistical (repairs, CALS statistical parameters (mean, median, the lead standard), correlations), multivariate statistical methods (clustering, factor analysis, discriminate function analysis, classifications and regression trees, linear and nonlinear regression models;

- Because *data visualization* allows you to view information in visual form, is one of the best and used data mining methods. Among the most commonly used techniques can include: histograms, graphs rectangular, scatter graphs, contour graphs, charts, graphics array.

Artificial Intelligence *(AI - Intelligence Artificial)* Data Mining is its contribution in providing information processing techniques based on human reasoning.

Database Systems *(DBS - Database Systems),* is that the third discipline that provides material that must be operated using the methods mentioned above.

Like *statistics,* data *mining* is a business solution, it is just a technology[4]. In contrast to statistical analysis, *data mining* analyzes all relevant data from the database and extract models (patterns) hidden.

The term "data mining" is used in particular by organizations concerned with information processing companies, financial analysts, in marketing, fraud detection, but lately it has become increasingly used in scientific research scientific, extract information from large volumes of data.

Data mining is a complex process, taking roots in the learning process, a process that can achieve the following levels:

- Facts - definition of reality;

- Concepts - a concept represents a set of objects having certain common symbols SPECIFICATIONS;

- Procedures - a procedure consists of a series of actions performed to achieve an objective;

- Principles - Principles are underlying truths to other truths.

Practical data mining analysis are focused on the discovery of new knowledge from data.

Information obtained by using data mining techniques to be valid. The accuracy and completeness are the two characteristics of the underlying validity of the data. This information should not only be valid and the data mining process itself.

Looking for Data mining definition

In the literature there are many definitions for data mining / knowledge discovery in data base because it is a relatively new technology.

Gray and Watson believe that "data mining allows analysts and store managers to find the answers to company data, which they have not even put"[5]. Data mining has been described as *"the science of extracting useful information from large volumes of data and database"*[6].

Data mining in relation to economic resources planning is the statistical analysis and logical data volumes of transactions, looking for patterns that can help decision-making process.

The data mining process of extracting means knowledge bases or data warehouses, knowledge previously unknown, valid and operational at the same time[7].

Data mining seeks not only verify the hypotheses, but aims at discovering new knowledge, information totally unknown until then. Thus, the results are very valuable.

Extracting knowledge from data performed correctly, the following benefits:

- Helps reduce costs for decision-making processes;

- assist managers in making business decisions faster and better quality;

- Services to business customers an improved knowledge signified;

[4] Tudor Irina, Cărbuneanu Mădălina, *Utilizarea tehnicilor de data mining în comunitățile de învățământ virtual,* Conferința Națională de Învățământ Virtual, ediția a V-a, 2007;

[5] H. Mannila, P. Smyth , *Principls of Data Mining,* Mit Press, Cambrige, ISBN 0-262-08290-X, 2001;

[6] Gray P. H.J. Watson (1996) The new DSS: Data Warehouses, OLAP, MDD and KDD;

[7] Ellen Mank, Bret Wagner, *Concepts in Enterprise Resource Planning,* Second Edition, Thomson Cours, Tehnology, Boston, MA, ISBN – 0-619-21663-8, 2006;

We can say that the whole company management knows an improved significantly.

The main features presented by data mining systems are as follows[8]:

- Findings generated by the system have a very high probability; can be considered "safe";

- compared with information held by the user or system, the resulting information is not trivial;

- discovered user models are represented in a comprehensible form.

Data mining process can be used, in principle, any type of database. It is used in particular in data warehouses because to be accurate and reliable results, the data mining process is applied to data of high quality that is to be cleansed, integrated, selected and processed. Such data are found to meet these conditions within the data warehouse.

Data mining and Business Intelligence

Smooth running of an organization is provided by the decisions be taken in time, decisions that are correct as long as they are based on data and information on which to base those decisions. Thus, it is necessary that managers take into account all available information sources, both within the firm, the information system and beyond.

The main purpose of business intelligence is to obtain information from available data, data that may come from scattered sources or stored in a warehouse date. A correct decision based on accurate information.

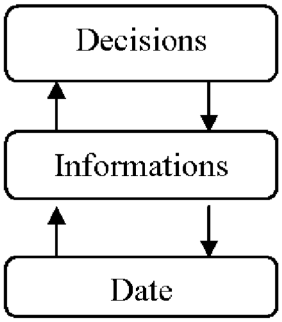

Figure data, information and decisions

Stages of processing data into information and information into decisions then form part of a cyclical process. Substantiation decisions require information and information resulting from the data. Apart from a context, no data are valuable information. To make the information, data should be in particular business contexts.

Using data mining analysis will identify new knowledge without being required human intervention. Basically, data mining analysis is aimed at discovering new knowledge.

Using applications of Business Intelligence (BI) offers users a number of significant advantages especially for those who have to make decisions about the good of the organization[9]

- Because the design and to develop business intelligence applications based on economic progress and requirements processes in organizations, the company will benefit because these applications;

[8] Bodea, C., Inteligenţa artificială şi sisteme expert, Bucureşti, Editura Inforec, 1998;

[9] Kimball R. Şi colectiv, *The Microsoft Data WarehouseToolkit with SQL Server 2005 and Microsoft Business Intelligence Toolset,* John Wiley & Sons, 2006;

- Business Intelligence applications using large volumes of processed data and information from the process are made available to a large number of users;

- Mechanisms for calculating the indicators of business performance analysis and are easily integrated business intelligence applications;

- In the design and development of Business Intelligence applications can be trained, usually decision makers within the firm.

And processed data from various sources are loaded into the data warehouse and then using OLAP and Data Mining technology data are converted into information, information that is presented to the beneficiaries in the form of report. A Model of Business Intelligence in the firm's decision making is presented in the following figure[10] :

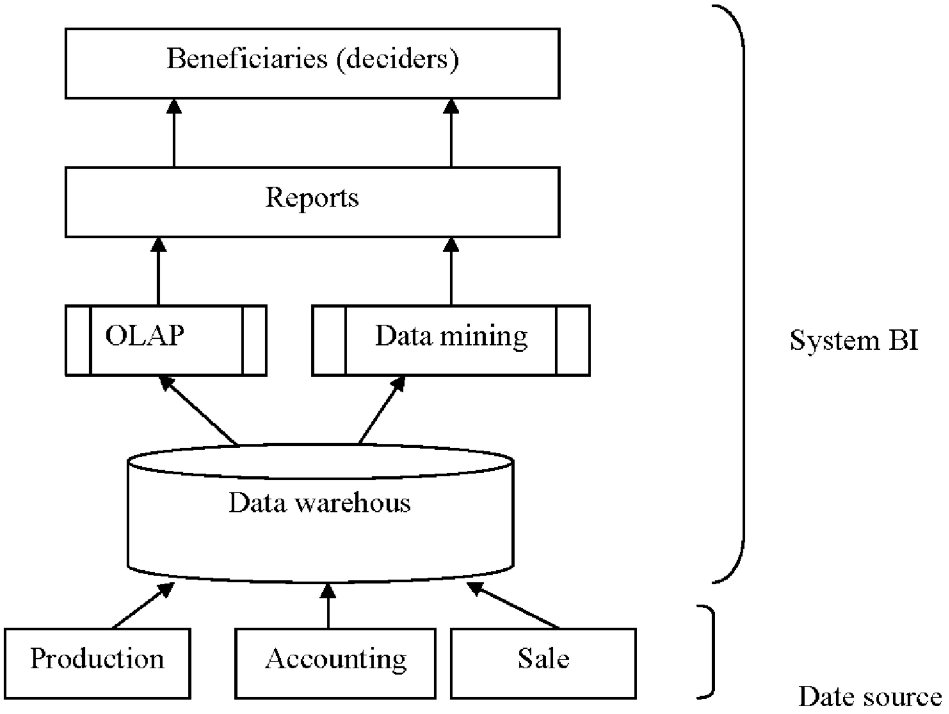

Figure: Business Intelligence in decision-making firm.

With data mining tools are analyzed all links between data characteristics. Also, data mining tools to solve problems of data inconsistency. Although systems containing integrated data source, cleaned and valid, they may contain data about the real world that is not true. These data are called noise and may as a result of human error in the input of data and information systems.

Also, the data may contain patterns (patterns) or trends that would be true only for a small subset of data. This can be examined from two points of view, which, in statistical terms but unattractive business point of view this is perhaps the most interesting customers serum due to their behavior. These types of sets are called local effects.

Using data mining techniques, data are analyzed from several angles simultaneously, which can help eliminate attributes that are considered singular and thus does not seem relevant.

[10] Giovinazzo, W. 2002 , Internet – Enabled Business Intelligence, Prentice Hall.;

Interdependencies between attributes are those that allow the extraction of data to information that is relevant.

Data mining is the process through which one or more machine learning techniques can extract hidden knowledge from data. The process of data mining is to identify trends and patterns in data sets.

The steps of extracting knowledge from data

Typically, when referring to the process of extracting knowledge from data refer only to the application stage, the different methods and data mining techniques, but do not pay attention to the steps needed for the process.

Main stages of extracting knowledge from data are:

1. Defining the problem - at this stage of acquisition of knowledge about the initial state, but also about setting goals and purpose of the application;

2. Building the database - making a single database where data are cleaned and stored to provide the necessary support for building (making) an appropriate model, also reduce the size of the data store, eliminating noise and finding how to replace the data missing;

3. Use of data - is the phase in which the data understanding, viewing them, but also the development of pivot tables to provide a clear view of the database consistency

4. Preparation is the stage when there is data to identify those variables that will constitute the components of the predictive model, you might like to consider all variables to ensure the best predictions. It is still quite difficult to achieve this because a large number of variables increases the time needed to build the model prediction and the decreased ability of the model results;

5. Building the model - is a process interactive as to Jung CAE better option, certain steps in the development process will be repeated, and it will also modify the data. The model can be considered final when there is completion of the training and testing.

6. Assessment model - considering the accuracy of results. Model is given by increasing the efficiency of certain indicators.

7. Interpretation of model results and improvement. This is the stage when the decision to the user based on the results it can decide to replay one or more phases of the project.

All the knowledge extraction process is centralized around an object on the problems encountered in the economic field, problems involving the identification, discovery of knowledge to help solve them.

We can say that the whole process of knowledge extraction is an iterative process, because during the course of this process, shows steps are executed repeatedly, sometimes by taking back some of them.

Although methods and techniques for extracting knowledge from data are applied in automatic mode, the data mining process requires considerable human effort, especially in the stages of analysis, but also in what the validation results.

Conclusions

Data mining is the process of discovering some models of knowledge or information of the user data stored in data warehouses usually type database or data warehouse, as well as advanced data analysis techniques used to discover these patterns.

Data mining has occurred because of developments in information technology, the avalanche of data produced by human society from its their activities and because of the need to transform data into information and knowledge for wide applications in analysis and production control, market analysis, detection fraud but also for medicine.

Data mining involves an integration of techniques from many fields. We can say that it is an interdisciplinary field that intersects the main subjects is technology database, data warehouse technology, statistics, business intelligence, digital technology.

We say that data mining is an advanced analytical processing of data is superior to limited analytical process of database systems and data warehouse analysis because data specific techniques more advanced.

Information that is discovered data mining process are used to substantiate decisions is made by the directors of a company.

References

- Bodea C., *Inteligența artificială şi sisteme expert*, Editura Inforec, 1998.
- Baragoin, C., et al., *Mining Your Own Business in Retail Using DB2 Intelligent Miner for Data*, Internațional Tehnical Support Organization, International Busins Machines Corporation, Red Book, San Jose, california, 2001.
- Giovinazzo, W. , *Internet – Enabled Business Intelligence*, Prentice Hall, 2002.
- Gorunescu F., *Data mining. Concepte, modele şi tehnici,* Editura Albastră, Cluj-Napoca, 2006.
- Gray P. H.J. Watson (1996) *The new DSS: Data Warehouses, OLAP, MDD and KDD,* 1996.
- Kimball R. Et al., *The Microsoft Data WarehouseToolkit with SQL Server 2005 and Microsoft Business Intelligence Toolset*, John Wiley & Sons, 2006.
- Mank Ellen and Wagner Bret, *Concepts in Enterprise Resource Planning*, Second Edition, Thomson Cours, Tehnology, Boston, MA, ISBN – 0-619-21663-8, 2006.
- Mannila H. and Smyth, *Principls of Data Mining,* Mit Press, Cambrige, ISBN 0-262-08290-X, 2001.
- Tudor Irina and Cărbuneanu Mădălina, *Utilizarea tehnicilor de data mining în comunitățile de învățământ virtual,* Conferința Națională de Învățământ Virtual, ediția a V-a, 2007.
- www.Wikpedia ro/datamining.