DATA MINING TECHNOLOGIES

Titrade Cristina-Maria[1]

Abstract

Knowledge discovery and data mining software (Knowledge Discovery and Data Mining - KDD) as an interdisciplinary field emersion have been in rapid growth to merge databases, statistics, industries closely related to the desire to extract valuable information and knowledge in a volume as possible.There is a difference in understanding of "knowledge discovery" and "data mining." Discovery information (Knowledge Discovery) in the database is a process to identify patterns / templates of valid data, innovative, useful and, in the last measure, understandable.

Keywords: data mining, knowledge discovery, data warehouse, data mining tools, data mining applications.

Intoduction

Data mining is a step in the process of discovery of information consisting of a set of data mining algorithms that, in accepted limits, discover significant patterns in data structures, which indicate the general market trends. Data mining uncovers patterns in data using predictive techniques. These models play an important role in making decisions because they highlight areas where business processes need improvement. Using data mining solutions, organizations can increase their profitability interact with their customers, detect fraud, management can improve high-risk activities, etc. Patterns discovered using data mining solutions help organizations make better decisions and in less time.
Most analysts separate data mining software into two groups:

- data mining tools - allow the user to a number of techniques that can be applied to any business problem
- data mining applications - incorporates techniques within a purpose-built applications to address a problem specifice.de business. Whether you realize it or not, our daily life is influenced by an application of data mining. For example, almost any financial transaction is processed by an application of data mining to detect fraud if there is bonding. What organizations are increasingly using data mining tools and applications together to develop predictive analysis.

Data mining tools are used to ensure flexibility and accuracy in analysis. They increase the effectiveness of data mining applications.

Components of Data Mining and KDD (Knowledge Discovery and Data Mining)

The main function of the DM is, therefore, is to extract knowledge from data models. For this, the DM uses a variety of algorithms from statistics, pattern recognition, classification, fuzzy logic, machine learning, genetic algorithms, neural networks, data visualization, etc.. The variety of algorithms can be grouped in the main components of

[1] Ph.D. Candidate at the Academy of Economic Studies, Bucharest and assistant teacher at the Romanian-American university, Bucharest

DM. The number of these components varies from one author to another. Thus, some believe that DM has three components, others four, etc..
The main components of DM are:

- model - which, like any computer model, is represented by a function in one dimensional or multidimensional space (a set of functions), depending on the parameters. It can be represented either as a linear function of parameters, either as a function of probability or as a fuzzy function, etc.. Getting the model is achieved by different algorithms, such as the classification and clustering;
- preference criteria - which may be of different nature, some sort based on other interpolation or the best approximation;
- selection algorithms - which lead to the selection of three important elements that appear in the DM, namely: the model, which models are selected from the data, which are selected from the database and are parameters, and the criterion or criteria preferences which selects the basis of criteria;
- determining violations - which generally consists of algorithms for determining the deflection and stability, a specific category of such algorithms are those statistics, determining deviations from the ideal model.

In terms of process there are three classes of data mining activities: discovery, predictive modeling and analysis of exceptions:

- Discovery is the process of database search to find models without a predetermined idea or hypothesis of what can be models. In other words, the program takes the lead in finding what the models interesting without the user needing to think about questions relevant advance. In large databases, there are so many models that the user would never think at all practical to questions should be asked. The key issue in this case lies in the richness models that can be found and expressed, and the quality of information delivered - elements that determines the strength and usefulness of the technique of discovery.
- Modeling predictive models found in the database are used to make predictions. Predictive modeling allows users to process records that have fields unknown value, and the system infer unknown values based on previous models in the database.
- Analysis is the process by which exceptions apply models extracted to find anomalies or unusual data elements. To discover anomalies, first learn what is normal, then detect those articles that deviate from the norm in a given interval. For example, once we observed that 90% of buyers were under 50 years, we ask about 10% of those buyers who are over 50 years and are based data. It is noted that the discovery can help us find 'common knowledge' while the analysis of exceptions find the unusual cases.

Knowledge Discovery Process

Of course, each commercial product uses several algorithms and each of them can find some or all of the above components in different proportions.

The authors distinguish between DM and KDD regarded as a complex interactive and iterative process, which includes DM. Thus, KDD is considered within the knowledge extraction is performed on the following steps:
- The first step is the understanding of the scope and problem formulation. This step is a prerequisite for extracting useful knowledge for choosing the most suitable method of

data mining for the third stage, in accordance with the destination application and the nature of the data.
- The second step is the collection and reprocessing of data, including selection of data sources, remove the outer layers, the treatment of missing data, data transformation and reduction. The most time consuming step of the whole KDD process.
- Step three is represented by data mining, the extraction process models or patterns hidden in data. An example is: a global representation of a systematic component structures that summarize the underlying data or data that describes how they arise. In contrast, a pattern is a local structure, probably associated with some variables and some conditions (cases). Most important are data mining predictive modelling methods for classification and regression, clustering, dependence modelling with graphical models and density estimation, etc..
- Fourth step is the interpretation (post-processing) found knowledge, especially in terms of description and interpretation of prediction - the two main purposes of the discovery system in practice. Experience shows that the models or patterns of data are not directly used as KDD process is repeated through the knowledge inevitably discovered. A standard way of evaluation is to divide the data into two sets, working on a set of data and testing on the second. We repeat the process a number of times, sharing data every time otherwise. Average score we use to estimate the performance rules.
- The final step is to put into practice the knowledge discovered. In some cases, you can use this discovery without it the hub of an integrated system, in other cases, the user uses to exploit this discovery by means of specialized software. Putting into practice the results is the ultimate goal of KDD's.

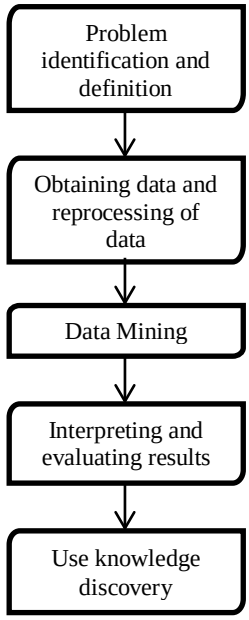

Figure no1 - KDD process for extracting knowledge

Relationship Data Warehouse, OLTP, OLAP and Data Mining

A relational database is designed with a specific purpose. Since the purpose of a data warehouse (data marts) differs from that of an OLTP, the design features of a relational database that supports a data warehouse differs from those of an OLTP database.

Data warehouse database	OLTP Databases
Designed to analyse a business size categories and attributes.	Designed for business operations in real time.
Optimized for large loads and query large, complex, unexpected to access multiple records from a table.	Optimized for a normal set of transactions, usually adding or deleting one record at a time on the table.
Loaded with data consistent, valid, does not require validation in real time.	Optimized for validating input during the transaction, using data validation tables.
Supports few current users compared to OLTP.	Supports thousands of current users.

Conclusions

Data Mining a Data Warehouse tool?
Data mining is a technology that uses sophisticated and complex algorithms to analyse data and to reveal interesting and necessary analysis by decision makers. While OLAP organizes data into a model suitable for exploitation by analysts, data mining performs analysis on data and provides results of the decision makers. Thus, OLAP-oriented model allows the analysis and data mining oriented data analysis easier.

Data mining has traditionally operated only on records from the database or data warehouse type text files extracted from the data warehouse database. In SQL Server 2000 Analysis Services provides data mining technology that allows data analysis in OLAP cubes, as well as data from relational data warehouse databases.

In addition, the results of data mining can be incorporated into OLAP cubes to give new features oriented analysis model provides a point of view the dimensional OLAP model. For example, data mining can be used to analyse the attributes of sales in exchange for buyers and create a new cube dimension to assist the analyst in discovering the information embedded in the data cube.

Bibliography

-Pieter Adriaans, Dolf Zantinge - Data Mining, 1996
-Soumen Chakrabarti, Earl Cox – Data Mining: know it all, 2008
-David Louis Olson, Dursun Delen – Advanced Data Mining Techniques, 2008
-Mehmed Kantardzic – Data Mining: concepts, models, methods, and algorithms, 2003