Original Research Paper

Volume-8 | Issue-9 | September-2018 | PRINT ISSN No 2249-555X

Economic

DATA MINING AND BUSINESS INTELLIGENCE

Stoica Liviu Constantin

PhD, Academy Of Economic Studies, Bucharest

ABSTRACT In this article titled Data Mining and Business Intelligence, I studied data mining and business intelligence technology and techniques. In the first part I studied the discovery of knowledge and data mining, ie the software and the main components used by it and the stages of the KDD process. In the second part we presented Data Techniques, namely: K-NN, classification, regression, decision trees. In the third part I presented Business Intelligence, that is, I presented: applications, concept, software and technologies.

KEYWORDS : business intelligence, data mining, tools, techniques, technology

1. DATA MINING TECHNOLOGY

Knowledge Discovery and Data Mining (KDD) have functioned as an interdisciplinary field that is rapidly developing and provides statistics, databases and areas of activity that are closely linked to the desire to obtain information and knowledge in a large volume of data.

There is a difference between "knowledge discovery and data mining", ie Knowledge Discovery in databases is the process used to identify valid, innovative, useful, and understandable models or templates.

Data Mining is the information discovery process and consists of a set of algorithms indicating discovery patterns that indicate market trend.

Data mining discovers models inside data using predictive techniques, and models are used to make decisions and highlight business processes.

Most of the analysts shared the software used in the data mining, as follows:

- Mining data tools that provide the user with techniques that can be applied to business issues and ensure accuracy and flexibility in analysis.
- data mining applications.

The primary function of Data Mining is to extract knowledge from data. For this, Data Mining uses algorithms in statistics, machine learning, neural networks, genetic algorithms, etc.

The main components of Data Mining are:

- the model that is represented by a function in the uni or multi-dimensional space;
- preference criteria that may be of a different nature, ie some are based on ordering, clustering, and interpolation;
- Selection algorithms lead to the selection of the most important elements that appear in Data Mining, ie: the model, data, and cross-referencing criterion;
- Determining deviations that consist of algorithms that determine stability and deviation.

From the studies conducted so far, it has been found that there are differences in DM and KDD and KDD has been found to be an interactive and complex process. KDD was designed in 1989 to designate a DM-based research area on form recognition, automatic learning, etc., and the first international conference on KDD and DM took place in 1995.

The discovery of knowledge in the databases is considered to be a process of identifying valid data patterns that can be understood as Fayyad showed in 1996 in "Advance in Knowledge Discovery and Data Mining" [1]. From its point of view, there are several stages in the discovery of knowledge, namely: data selection, data preprocessing, data transformation, data extraction, interpretation of the results.

These are exemplified in the following figure as follows:

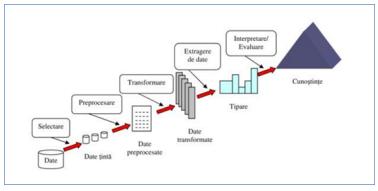

Figure .... KDD process steps (1)

In KDD knowledge extraction is done as follows:

- The first step is the stage of understanding the scope and problem formulation
- The second step is the process of collecting and reprocessing data, selecting the data source, etc. This is the most important stage in the KDD process
- The third stage is the data mining stage, the knowledge extraction stage or the patterns that are hidden in the data
- The fourth stage is the stage of knowledge interpretation, ie the interpretation in terms of prediction and description
- The fifth step is the implementation phase of the discovered knowledge, that is, the ultimate goal of the KDD.

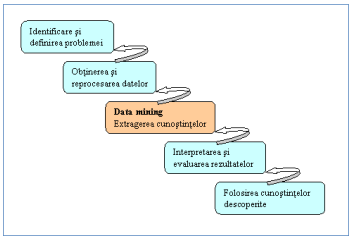

Figure .....

Relational databases are designed for a specific purpose. It is known that the purpose of the data warehouse differs from that of the OTP, the design features of the relational databases that support the data warehouse differ from those of the OLTP type. Data mining is a technology that uses algorithms used for data analysis, and OLAP organizes data for exploitation by analysts.

2. DATA MINING TECHNIQUES

Today, businesses and business produce large volumes of data that are used in day-to-day operations. In the analysis process, the information is extracted from the databases and can be exploited to build new models to identify relationships between database records, to produce economic forecasting models, to classify records, etc.

Data mining techniques can be grouped into three categories according to the problems they can shape, namely:

a) Classification and regression comprises the widest category of applications and is based on models to predict belonging to a set of classes;

b) Analysis of associations and successions also known as "shopping

cart analysis", a technique that generates descriptive models describing the correlation rules that are found between the attributes of the dataset;

c) Cluster analysis is a descriptive technique that is used to group similar entities in a set of data. Grouping techniques are based on demographic algorithms, neural networks, K-NN, etc. The difference between regression and classification is that in the case of classification, the predicted output represents the belonging to a particular class, and in the other case, the outpu estimated the value of an attribute.

I know that regression is used when the output is defined over a wide range, and regression problems can easily be transformed into classification issues and vice versa.

1) The statistical methods are:

a. Regression. Regression models are used to predict the value of a response with one or more predictive variables. Regressions are of several forms: linear regression, multiple regression, polynomial regression, nonparametric regression and robust regression;

b. Generalized linear patterns allow to establish relations between the variables that u respond critically and the series of predictive variables. These models include Poisson's logical regression and regression; [2]

c. The variability analysis is a technique that analyzes experimental data for two or more populations. Generally, ANOVA (Analysis of Variance) analysis involves a comparison of the K population to see if there are at least two of the different characteristics;

d. Mixed pattern models are used to analyze grouped data, data that can be classified by one or more grouping variables.

e. Factor analysis is used to determine the variables that are combined to generate a specific factor;

f. Discriminant analysis is a technique used to predict a categorical response variable;

g. Time series are techniques used in statistics to analyze time series data, such as: AutoRegressive Integrated Moving Average (ARIMA), autoregression methods, time series models; [3]

h. Survival analysis is a statistical technique and was designed to predict the probability that a person can survive for a while.

2) The k-NN (k-Nearst Neighbor) algorithm is a predictive data exploration technique that is used in particular for classification issues. This technique implies that the entire training set includes data and classifications for each item. After applying the algorithm, K represents the number of cases or similar items in the group. [4]

This algorithm has two parameters, namely: the number of closest k cases and the metric for measuring similarity.

3) Clustering or clustering is a statistical method used to group multidimensional data into algorithmically defined clusters. [5]

4) Techniques of the new generation:

- Rules. A combination rule is defined as I = {i1, i2, ..., im} a number of symbols called and elements, or D a set of transactions, D = {t1, ..., tm}, T is included in I, T is the transaction. Association rules are used to find common sets of databases containing consumer transactions.
- Bayes technique is a classification technique and is less implemented in data exploration applications. This technique owes the name of British Minister Thomas Bayes (1702-1761) and allows the analysis of the relationship between the dependent variable and the independent variable using conditional probability calculus.
- Neural Networks. This technique has concepts in the field of artificial intelligence, namely: artificial neuron is the basic unit for information processing within the neuron and the artificial neural network which is the synonym of artificial neurons that are connected by means of connections
- Decision trees are a data exploration technique that has both predictive and descriptive potential [6]. The great advantage of this technique is that most of the algorithms that make up the decision trees can be applied without any restriction as to the nature of the data. Decision trees include: Classification and Regression Trees (CART), Chi Intelligent Interaction Detection (CHAID). These algorithms provide sets of kingdoms that can be applied to a set of unclassified data to estimate the records that have a particular output. It is known that the CHAID algorithm segments a set of data that creates binary subsets, and the CART algorithm requires less data preparation than the CHAID algorithm (11). Also, from the studies we have done so far, I can say that the CART algorithm is a prediction and exploration algorithm and is robust about missing data 2 test that is in theand the CHAID algorithm is based on the contingency buckets in order to be able determines the categorical predictor that is furthest from independence of estimated values [7].

3. BUSINESS INTELLIGENCE

Companies to make decisions need information. In general, most are spread in IT, and the transformation of data into information can be analyzed for decision making which is rather cumbersome.

Business Intelligence, Data Warehouse, and decision support applications help businesses respond in real-time to complex questions.

There are many data mining techniques, and the choice of the appropriate tool leads to finding answers relevant to the company.

The process is dynamic, and the responses change as the business strategy changes.

Data warehouse and business intelligence lead to meeting business requirements, ie customer knowledge and knowledge of the organization. According to the 2000 Data Warehousing Institute, Data Warehouse and Business Intelligence technologies allow companies to analyze and integrate customer information.

An integrated system offers the following bendeficies:

- simpler systems;
- quick access to current data;
- Infrastructure management savings;
- more efficient management;
- Increase profit.

Business Intelligence applications provide companies with better data capture capabilities that result directly, and the implementation of Business Intelligence solutions can affect a company's profits directly.

The concept of Business Intelligence is the process of implementing strategies to acquire and analyze all information from different sources in a legal and ethical way in order to substantiate the decision within a group, a management entity, people and technologies.

BI software can become a valuable tool for business intelligence within a company, but only until the analysis stage is able to select, classify, and compare data and information.

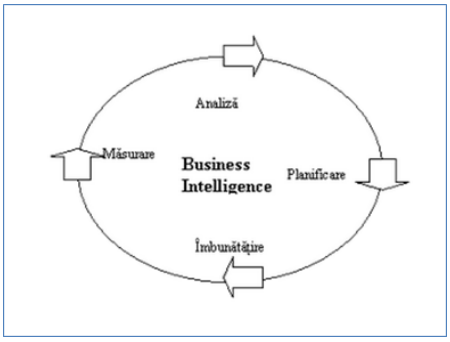

Figure ....

One study showed that about 40% of US companies implemented BI solutions, and the rest want to launch adoption of BI technologies in the next year.

Today, business is conducive to the spread and use of BI applications. There are, however, companies that have high spending on technology

purchases, and BI is a more or less obvious initiative in terms of profitability. I mean, these apps help in making the right decision.

Business Intelligence applications allow you to categorize and detail at the same time specific to the exact information process. I can say that BI is a young platform that until recently had many applications that were used only in several departments within the company. This has quickly changed, companies have developed or have purchased BI applications covering almost all of the company's functionalities, ie: financial analysis and sales analysis, human resources, key client developments, etc.

Thus, the company can expand its BI benefits to business participants: customers, employees, suppliers, shareholders or partners.

In order to be used by as many companies as possible and in a more efficient way, the technologies that make up a BI platform must be organized.

ABI platform may contain the following technologies:

1. Database. A BI platform must contain relational databases and multidimensional databases;
2. OLAP (On-Line Analytical Processing). OLAP is a core component of a BI platform, being the most widely used method of analysis;
3. Data Mining. Data Mining is the activity of extracting and analyzing data in order to discover the elements that are hidden or that are harder to find out of the databases. Data Mining can help determine relationships and correlations between data or data groups.
4. Interfaces must be friendly and link databases with OLAP and Data Mining.

CONCLUSION

As far as the storage of large volumes of data is concerned, I am thinking of the need to extract data based on data. Thus, the use of Data Mining is required to obtain different statistical data or forecasts across a wide range of domains. Data Mining is a relatively new field, old data exploration methods (clustering, regression, etc.) have been replaced by newer performing methods, for example decision trees.

REFERENCES

[1] Fayyad U.M., Piatetski-Shapiro G., Smyth P. and Uthurusamy R., Advances in Knowledge Discovery and Data Mining, AAAI/MIT Press, 1997
[2] Mosteller, F. and Tukey, J. W. (1977) Data Analysis and Regression. Reading, MA: Addison-Wesley
[3] Han, J. & Kamber, M. Data mining concepts and techniques (2nded.), edited by Morgan Kaufmann, V. Harinarayan, A. Rajaraman, J.D.Ulman. San Francisco, 2006
[4] Hart , P. E. , Cover , T. M. Nearest neighbor pattern classification, IEEE Transactions onInformation Theory, IT-13, 1967
[5] Barbara, D. An introduction to cluster analysis for data mining
[6] RokachL.,Maimon O. Data mining with decision trees-Theory and Application
[7] Nepomnjashiy, A., Data Mining Algorithms: Microsoft SQL Server 2000 vs. "Yukon" SQL Server, DatabaseJournal.com, 2004, http://www.databasejournal.com/