Data Mining for Increasing Economic Analysis Efficiency

Prep. Cristina Ofelia Sofran
Lect. dr. Alin Munteanu
Universitatea "Tibiscus", Timişoara

REZUMAT: Acumularea datelor în baze de date de dimensiuni neobişnuit de mari îngreunează prelucrarea acestora şi identificarea datelor utile, dar şi alcătuirea de statistici. Soluția este procedeul numit „data mining", ce permite extragerea de cunoştinţe utile ascunse în cadrul unor cantităţi mari de date brute.
Cuvinte cheie: date, informație, analiză, data mining

1 Information in Economical Context

Information is the most valuable asset of a corporation and that is the reason it should be wisely handled. It is well known that a lot of businesses fail because of a law data management or because of the manager's lack of skills to manage information.

Information derives from data. Any company, person or situation encounters a large amount of data, raw data, which requires certain skills and intelligence in order to select exactly the data that provides the needed information. The metamorphosis of data is presented in the Figure no. 1 bellow:

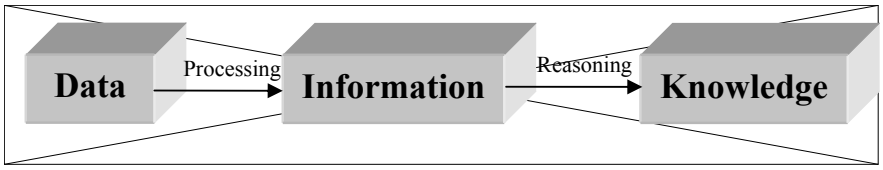

Figure no.1. Data – information – knowledge relation

Most of the twentieth century was dominated by the so called "industrial age" when among the three known production factors – labour, land, capital – the most important one used to be capital. At the end of the twentieth century, a new asset was discovered, and it became the main production factor of the contemporary "informational age": this asset is information. In the economical field, information's importance is growing each day and it is a continuous race in the economical life for achievement of information at the very moment it is required. In the current competition conditions on any market, it matters a great deal who will be the first to have certain information.

2 Reasons for Using Data Mining

Data mining - sometimes called data or knowledge discovery - is the process of extracting knowledge hidden from large volumes of raw data.

Data mining software allows as an analytical tool to analyze data from many different dimensions or angles, categorize it and summarize the identified relationships. Data mining is actually a process that is supposed to find correlations or patterns among dozen of fields in large relational databases.

The main reason for the necessity of data mining usage, meaning an intelligent data analysis, is the large volume of existing and newly appearing data that require processing. This volume of data is enlargening each day by storing large amount of information by each organization. Human analysts find themselves in an impossible situation when dealing with such overwhelming amounts of data.

Another reason for the need of automatic analysis instead of human analysis is that there are two major problems that surface when human analysts process data: first, the inadequacy of the human brain when searching for complex multifactor dependencies in data and second, the lack of objectiveness of human processed analysis.

Data mining process will substitute the work that should've been done by professional statisticians that were highly trained but also highly paid. This way, using data mining, an analyst that is not a professional in statistics or programming will easily manage to extract knowledge from data.

For a better understanding of how useful data mining is here are some tasks solved by data mining:

- Predicting: a task of learning a pattern from examples and using the developed model to predict future values of the target variable.

- Classification: a task of finding a function that maps records into one of several discrete classes.
- Detection of relations: a task of searching for the most influential independent variables for a selected target variable.
- Explicit modeling: a task of finding explicit formulae describing dependencies between various variables.
- Clustering: a task of identifying groups of records that are similar between themselves but different from the rest of the data. Often, the variables providing the best clustering should be identified as well.
- Market Basket Analysis: processing transactional data in order to find those groups of products that are sold together well. One also searches for directed association rules identifying the best product to be offered with a current selection of purchased products.
- Deviation Detection: a task of determining the most significant changes in some key measures of data from previous or expected values.

3 Data Mining Software

Most analysts separate data mining software into two groups: data mining tools and data mining applications. Data mining tools provide a number of techniques that can be applied to any business problem. Data mining applications, on the other hand, embed techniques inside an application customized to address a specific business problem. Regardless of whether we are aware of it, our daily lives are influenced by data mining applications. For example, almost every financial transaction is processed by a data mining application to detect fraud. Both data mining tools and data mining applications are valuable, however. Increasingly, organizations are using data mining tools and data mining applications together in an integrated environment for predictive analytics.

So what do data mining tools add? Data mining tools are used to ensure flexibility and the greatest accuracy possible. Essentially, data mining tools increase the effectiveness of data mining applications. Since no two organizations or data sets are alike, no single technique delivers the best results for everyone. Not only do data mining tools deliver in-depth techniques, but data mining tools also deliver flexibility to use combinations of techniques to improve predictive accuracy.

Because data mining tools are so flexible, a set of data mining guidelines and a data mining methodology have been developed to help

guide the process. For example, the Cross-Industry Standard Process for Data Mining (CRISP-DM) ensures the organization's results with data mining tools are timely and reliable. This methodology was created in conjunction with practitioners and vendors to supply data mining practitioners with checklists, guidelines, tasks, and objectives for every stage of the data mining process.

While large-scale information technology has been evolving separate transaction and analytical systems, data mining provides the link between the two. Data mining software analyzes relationships and patterns in stored transaction data based on open-ended user queries. Several types of analytical software are available: statistical, machine learning, and neural networks. Generally, any of four types of relationships are sought:

- Classes: Stored data is used to locate data in predetermined groups. For example, a restaurant chain could mine customer purchase data to determine when customers visit and what they typically order. This information could be used to increase traffic by having daily specials.
- Clusters: Data items are grouped according to logical relationships or consumer preferences. For example, data can be mined to identify market segments or consumer affinities.
- Associations: Data can be mined to identify associations. The beer-diaper example is an example of associative mining.
- Sequential patterns: Data is mined to anticipate behavior patterns and trends. For example, an outdoor equipment retailer could predict the likelihood of a backpack being purchased based on a consumer's purchase of sleeping bags and hiking shoes.

Data mining consists of five major elements:

- Extract, transform, and load transaction data onto the data warehouse system.
- Store and manage the data in a multidimensional database system.
- Provide data access to business analysts and information technology professionals.
- Analyze the data by application software.
- Present the data in a useful format, such as a graph or table.

Different levels of analysis are available:

- Artificial neural networks: Non-linear predictive models that learn through training and resemble biological neural networks in structure.

- Genetic algorithms: Optimization techniques that use processes such as genetic combination, mutation, and natural selection in a design based on the concepts of natural evolution.
- Decision trees: Tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset. Specific decision tree methods include Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID). CART and CHAID are decision tree techniques used for classification of a dataset. They provide a set of rules that you can apply to a new (unclassified) dataset to predict which records will have a given outcome. CART segments a dataset by creating 2-way splits while CHAID segments using chi square tests to create multi-way splits. CART typically requires less data preparation than CHAID.
- Nearest neighbor method: A technique that classifies each record in a dataset based on a combination of the classes of the *k* record(s) most similar to it in a historical dataset (where *k* 1). Sometimes called the *k*-nearest neighbor technique.
- Rule induction: The extraction of useful if-then rules from data based on statistical significance.
- Data visualization: The visual interpretation of complex relationships in multidimensional data. Graphics tools are used to illustrate data relationships.

4 Conclusions

Data mining provides very useful methods to realize efficient economic analysis that classical methods were not able to provide. Data mining is quickly developing and recent advances have led to the newest and hottest trends in data mining - text mining and Web mining. These two data mining technologies open a rich vein of customer data in the form of textual comments from survey research and log files from Web servers, which were previously unusable. Applying data mining to these data adds a richness and depth to the patterns already uncovered through data mining efforts.

Reference

*** http://www.anderson.ucla.edu

*** http://www.crisp-dm.org

*** http://www.megacomputer.com

*** http://www.statsoft.com