Commercially Available Data Mining Tools used in the Economic Environment

Mihai ANDRONIE[1], Daniel CRIŞAN[2]
[1]Academy of Economic Studies, Bucharest, Romania
[2]Eidgenössische Technische Hochschule Zürich, Switzerland
mihai_a380@yahoo.com, crisand@student.ethz.ch

This paper presents some of the most common commercially available data mining tools, with their most important features, side by side, and some considerations regarding the evaluation of data mining tools by companies that want to acquire such a system. Among some of the most important factors that a company has to take into account are the amounts of data available, how it is stored and the data mining tasks that must be performed, but there are also others. Not the last it should be mentioned that the cost of a data mining system is important for a company, having a limiting effect on the expansion of the data mining products market towards small companies.
Keywords: *Data Mining, Data Analysis, Information System, Data Mining Tools*

1 Introduction

Following the significant advances in the information technology during the last decades, our society faced an increase of the volume of data produced and stored in various fields of activity. For this reason there were developed various advanced technologies to process the huge volumes of data suggestively called data mining techniques (an analogy to the mining processes that are carried on to extract precious metals from the ground).

Data mining answers the need to process huge volumes of data to gain useful information and knowledge in various fields such as market analysis, production control, marketing, scientific research and others.

In the economic field there are many benefits of using data mining techniques in integrated information systems. In the following there will be mentioned some of the most important techniques used in conjunction with their range of applicability in the economy.

There will also be presented some of the most representative commercial available data mining tools with their most important features, side by side so that their key features to be outlined to the reader.

Not the least important of the material presented in this paper are the factors that a company should take into account before deciding to buy such a commercially available data mining tool, in order to use it for its management to improve its decision making capabilities.

Besides using data mining techniques for company management support, among the most useful applications of data mining in economics is the area of retail distribution of merchandise. There exist traditional or online stores that record information in their databases regarding their sales, customers, etc. All this information gathered over a long stretch of time can be analyzed using data mining techniques.

Data mining tools can be used even for the improvement of a company's information system, this having measurable economic consequences even if it isn't an economic purpose in its own.

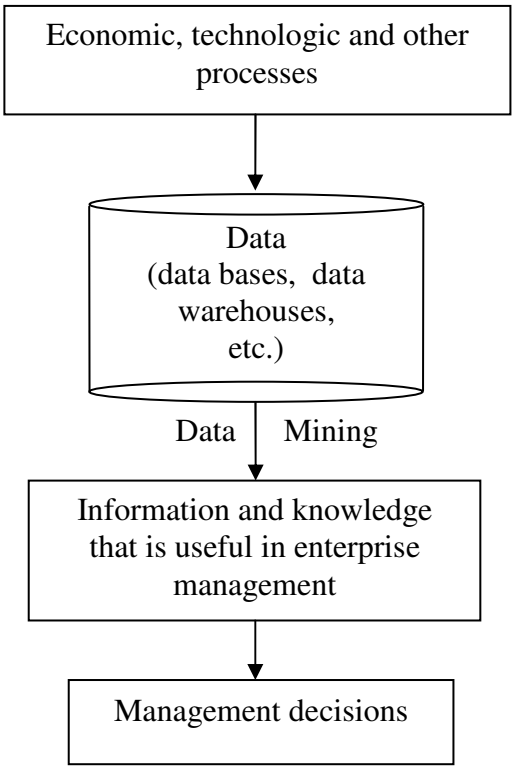

Fig. 1. The role of the data mining techniques in taking management decisions

We have for the cases previously described different applications of the data mining techniques [4] such as:

- The well known marketing problem [3] of shopping basket analysis from which there can be obtained different relationships between the products which were sold together in the past;
- Using data mining techniques to assist in designing data warehouses for a specific application. In this way one can more easily determine what dimensions should be included in such a model, according to the data available;
- Determining the effectiveness of sales campaigns – it is another economic problem to which a solution can be found using different data mining techniques; if it is found that sales campaigns do not have the desired results these can be improved or canceled;
- Analysis of customer behavior; customers may or may not be loyal to the company and they may respond more or less on certain factors which depend on the company's policy; knowledge of the customer's behavior can be very useful in adopting a business strategy to maximize profits.
- Another very important application of data mining in economics is to use advanced data analysis techniques for strategic management of enterprises. For a company is vitally important that decisions that are taken by the top management may be taken on an informed basis and not based solely on the talent and experience of the manager. This application of data mining techniques became possible by making predictions based on the data that the company has access to (data from its own databases or even external sources of data). This is perhaps the area where data mining techniques are the most important because the data gathered in time cannot be managed and analyzed traditionally being in huge quantities.

2. Factors that influenced the development of data mining techniques

Early in the development of data analysis there have been various statistical techniques for extracting useful knowledge from data. These techniques didn't require a computing power similar with what we encounter today.

The first data mining techniques were derived from statistical methods and are called classical techniques (such as statistical analysis or cluster analysis which are even today among the data mining tasks performed by modern data mining systems).

The classical techniques were originally associated with data collections and were later adapted to analyze huge volumes of data. They remain the most used data mining techniques because specialists are very familiar with them, being usually applied on databases and data warehouses.

Once the technology has evolved there appeared more diverse data types, so that new generation data mining techniques (also called next generation techniques) started to

be developed.

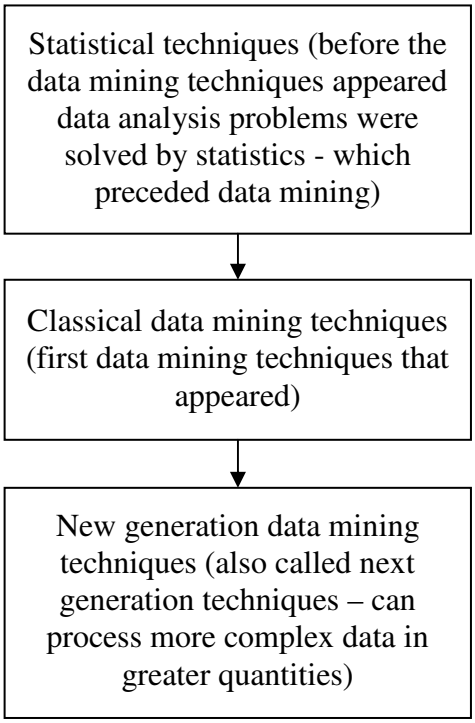

Fig. 2. The development of data mining techniques since their beginning

New generation data mining techniques are able to process a much larger volume of data (compared to the first databases and data warehouses) in a relatively much shorter period of time. Also, these techniques can easily process the unstructured data types that may prove a valuable source of data but are little exploited.

Taking into account the current evolutions in the data mining field we can anticipate that in the near future it will be possible to exploit data such as the content of papers that until now were mostly on printed format like books (scanned prior to being put into electronic format), maps, and others.

Also we expect that the Internet will become one of the most important sources of data publicly available and it will be exploited at levels not seen until now. In this case the only major obstacles that have to be overcome are the high degree of heterogeneity of web-based data and choosing the relevant information from the millions of web pages that are available. Not the least it should be taken into consideration that not all the information found on the Internet is correct, the analysis of incorrect data leading to errors in the results.

Last but not least it should be mentioned that it is not conceivable an advanced data analysis without a prior development of effective methods of data storage, retrieval and processing, this being facilitated by technologies in the fields of databases and data collections. For this reason it is not a coincidence that most of the data mining performed until today had as basis relational databases so relational databases came quickly to be a preferred option for storing and managing large volumes of data. Among the factors that contributed to the development, acceptance and maturity of relational database technologies must be mentioned online transaction processing databases (OLTP).

3. The role of data mining techniques

Although the field of data mining is relatively recent, as a result of the the research in the last years have appeared a large variety of algorithms, methods and techniques which allow users to perform a wide range of functions using these techniques. These functions are important for an enterprise that has to choose a data mining system to implement so that they will be mentioned below.

Data mining techniques, data mining algorithms and data mining methods each have a purpose for which they were designed, performing a specific data mining task [5]. Among the most common tasks the data mining systems perform include clustering, association, classification, regression, forecasting, analysis of data sequences (series of discrete values) or outlier analysis (analysis of exceptions) (Fig. 3.).

A data mining system can be specialized for one task or can be general enough to perform multiple data mining tasks. Most of the available data mining tools described below implement a number of algorithms

and are able to perform most of these tasks, being general enough to cover the needs of the customers.

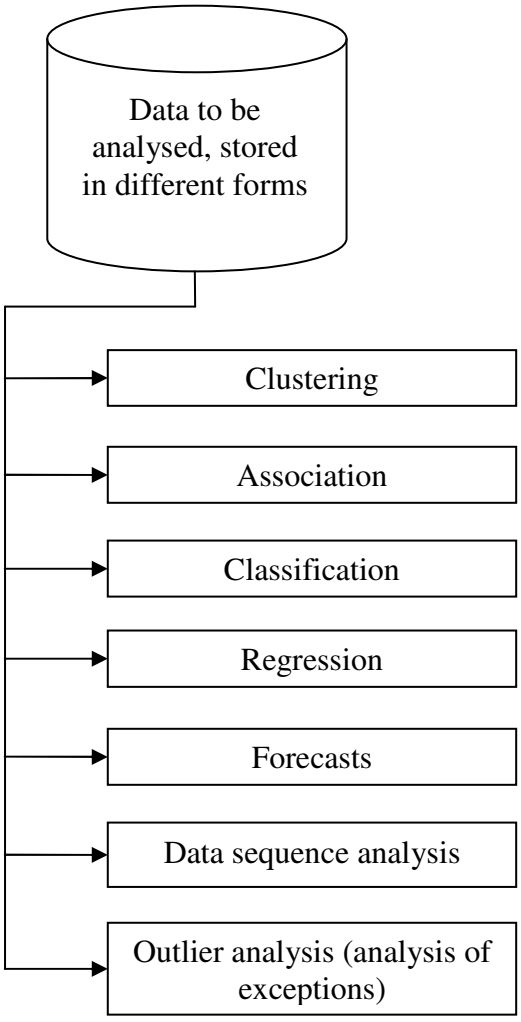

Fig. 3. The main data mining tasks that can be used to analyze a large amount of data

In the next paragraphs will be summarized these data mining tasks and will be highlighted the specific characteristics for each of them. Some of them can be used to some extent for similar problems, but there are also notable differences that distinguish one from another [6]:

- *Association* – aims to identify the most common sets of "objects" that appear together, sometimes the end user can choose which objects he wants to be analyzed or how frequent he wants them to be; the user can also set rules of association to better control the analysis; Among the most common applications of association are the identification of products most frequently sold together (market basket analysis), identifying areas where more than two products are sold together most often or the identification of time periods during which the sales are growing;
- *Classification* – It is one of the most common data mining tasks, responding to issues such as risk analysis; unlike clustering, where categories are not predetermined, the objects are inserted into categories, called classes, according to predetermined characteristics; classification algorithms need to work in a supervised mode, requiring some criteria in order to categorize objects; for this reason usually there must be previously established criteria for classification which can be obtained as a result of the analysis of historical data held; some classification techniques include decision trees, neural networks, etc. As examples can be mentioned the classification of bank customers in terms of how they pay debts, insurance risk analysis in order to determine the insurance premium for various specific cases or establishing property taxes based on certain criteria (value, area, size);
- *Clustering* – helps to find categories of "objects" based on their attributes; groups are formed containing "objects" similar in many respects (having similar attributes); a difference between clustering and classification is that clustering algorithms can work in an unsupervised mode, taking as input the attributes of the objects and offering as outputs the groups obtained (clusters); Two examples of clustering are grouping the customers of a company according to their purchase value and grouping available products by their features;
- *Forecast* – aims to make a prediction while taking into account past values usually in the form of series of events conducted over time and unlike regression can take into account other factors such as periodical fluctuations of events; As examples of forecasting can be mentioned the seasonal

forecasts in different fields of activity (if the variation is cyclical the regression is not suitable for analysis) or the forecast of sales of goods for the future based on the data available from the past;

- *Outlier analysis* – this type of analysis is intended to detect a number of "objects" that behave very differently from the rest; as an example it is most frequently used for fraud detection, intrusion detection or finding errors;
- *Regression* – is a data mining task that has its origin in statistics where it was used extensively; it resembles from some points of view with classification the difference being the fact that it works on continuous-valued attributes; regression can be used for predictions also, when the data analyzed is time related, but not for cyclic predictions (linear or polynomial non periodic predictions); a frequent use of regression is the calculation of intermediate values by interpolation;
- *Sequence analysis* – it is used to identify patterns in a series of discrete values; the main difference between sequence analysis and association is that the first is searching for the order of events and the transitions between different states while the second only searches for correlations between supposed independent objects; an example of sequence analysis can be the identification of models followed by a company's sales over time.

As it can be seen from those stated before, there are some similarities between some of the data mining tasks. Thus for example it can be said that both the classification and the clustering have the effect of dividing "objects" into groups or classes, but the difference between them is how this action is done. The classification deals with pre-existing classes that are defined before the analysis is performed, while the clustering groups "objects" into groups which are constructed taking into account the characteristics of "objects" under consideration.

Another difference worth mentioning is that between association and sequence analysis. If the first simply notes some rules and patterns in the data analyzed, the latter is looking for similar models on some data ordered in time, trying to determine a way in which these models are ordered, assuming that there are some correlations between the events.

It also should be noted the difference between the tasks of regression and forecasting. Regression plots a general trend, if carried out in time, can be used among others to make simple forecasts also. However, the forecast task is more complex and can take into account other factors, working on cyclical events, etc. For example forecasts can be used for data that varies from a period of time to another like sales around the seasons of the year.

A correlation between different data mining tasks that is worth mentioning is that between outlier analysis and other tasks like clustering or classification. The latter two are not intended to find exceptions in the data available (outliers) but can also be used for these tasks because an exception can be classified in a different class (or cannot be attributed to a certain class at all) and it can also form an individual cluster, very different from the others. Anyway, it is not needed to use classification and clustering to analyze exceptions because there are specialized techniques and algorithms for this task.

4. Considerations Regarding the Evaluation of Data Mining Tools

For a system with embedded data mining techniques to be profitable for a company's business it is necessary for it to have certain features. The characteristics that a computer system with embedded data mining techniques must have should be in connection with the business where it needs to be used and its requirements.

First of all, for a company operating in the economic field, whatever its domain of activity is, it is necessary to have an existing

computer system with data to be analyzed prior to the data analysis. Only after this requirement is met, a data mining system can be installed and used to extract useful information from the available data.

Another fact that should be considered is that in many cases a company's computer system consists of a collection of applications that are interconnected and have access to data that is often stored in relational databases. The users of the enterprise's computer system, usually employees of the company, have access to this data using these applications, not directly (as shown in Fig. 4). To integrate a data mining system in the computer system of a company it must be able to retrieve data from the computer system's database using their analysis techniques and performing tasks such as clustering, association, prediction, regression, etc. [3]. Following this analysis on the existing data there can be extracted useful information and knowledge that can be used at various levels within the company, from the basic business processes to the top management for decision making.

Another thing to consider before deciding to integrate a data mining system in a company's computer system is the importance of the data that the company has. The data that the company owns or has access to must be relevant and contain potential knowledge to be searched for using data mining techniques. For this it can be made an assessment of the data at which a company has access. In most cases the use of data mining techniques brings benefits to companies, regardless of their field of activity.

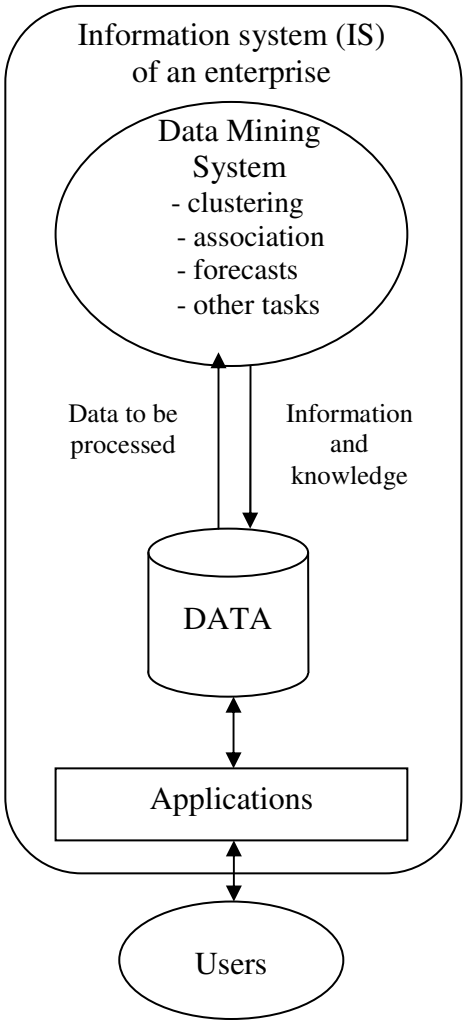

Fig. 4. The integration of a data mining system into an enterprise's information system

Not the least important it must be taken into account the costs of implementing a data mining system. For small companies the high costs of data mining products can be an important factor that can slow down the penetration of data mining tools on the market. On the other hand for large companies which are more financially powerful, the use of these products has already shown benefits.

Companies nowadays are provided with a large variety of data mining commercially available tools. The only problems they face is identifying their needs and evaluate the available tools in order to choose the data mining system that best fits its needs and budget.

Because the data mining products available

on the market tend to have different features in most of their aspects, a company should plan to consider a number of criteria against which to evaluate a data mining tool before deciding which data mining system to acquire [4]. Among the most important criteria to be considered are:

- The amount of data that is available to be analyzed. Is it necessary to buy a more powerful data mining tool which is more expensive, or it is enough something simpler?
- The amount of preprocessing the available data needs before being mined. If the data is stored in relational databases it is easier to analyze it, and most of the data mining systems will work. But if for example the data is printed text, first it must be scanned or introduced in the computer and only after it should be analyzed using a data mining tool that can take text as an input.
- The way the available data is stored. If the data is stored in relational databases it is needed a data mining system that can work on databases, but if the data comes from large data streams it is necessary a more specialized data mining tool that can make real time analysis.
- How complex the analysis must be. For simpler analysis the data mining system can be a more affordable one but for more complex analysis a specialized data mining system is necessary.
- What data mining tasks the company wants to be performed by the data mining system: association, clustering, classification, outlier analysis, regression, forecasts or others.
- The data mining system that is chosen must cover all the future needs of the company. If it is intended to make other types of analysis in the future those bust be covered too by the system.
- Flexibility. A data mining system must be flexible to be adapted to different types of analysis. For each data mining task there can be implemented more than one data analysis algorithm.
- The coupling between the data mining system and the database management system that the company is using (in the case of analyzing data stored in databases). A data mining system that is more coupled with a database management system has access to its internal functions and is more efficient. A data mining system that isn't coupled with a database management system uses its external functions to access data (like SQL queries for example) and for this reason is not so effective, but it is more flexible.
- API interfaces available. Some data mining systems offer API function libraries that make possible the integration of data mining functions in the software that a company is already using. This is a great advantage because it eliminates the need of running more applications at the same time, one for current use, and one for data analysis.
- Scalability of the system – it is very important if the company's database extends and becomes necessary to analyze a data volume higher than originally planned.
- How user friendly a data mining system is. It is important because most of the times the persons using the system are not IT specialists. Visualization tools are important because they make the presentation of the obtained results to be more suitable to be grasped by the human end user.

5. Features of Some Commercially Available Data Mining Tools

Data mining systems available on the market are usually the products of companies coming from the databases, hardware, statistical analysis or other related fields. Some of the most popular products on the market are presented in Tab. 1 [6]. where there are tresented the main features of each. Among these there are included: Intelligent Miner, produced by IBM, the data mining tools included in Microsoft SQL Server 2005 package and later, Oracle Data Mining, working with Oracle 10g database, and

others. Each of these solutions implement several data mining functions, each function in turn using several techniques and methods of performing the data analysis.

Table 1. Some data mining products commercially available

Data Mining Product	**Main Features**
SAS Enterprise Miner	• It comes from statistics; • Easy to use graphical interface; • Rich set of algorithms including algorithms for data mining: decision trees, neural networks, regression, association, etc. • Ability to analyze text.
SPSS	• It comes from statistics; • Includes among others, decision tree data mining algorithms (Answer Tree); • Allows users to perform data cleansing and data transformation.
IBM Intelligent Miner	• It comes from the database field; • Features advanced visualization tools and data presentation; • Compatible with PMML language (Predictive Modeling Markup Language) for exporting the data models found; • Can work with DB2 database management system.
Microsoft SQL Server 2005	• It comes from the database field of activity; • Among others, it offers algorithms for decision trees, prediction and clustering; • Implements the OLE DB standard for Data Mining, which defines a data mining language similar to SQL; • Features an easy to use API interface for facilitating the integration of data mining facilities into the user applications.
Oracle Data Mining (from Oracle 10g)	• It comes from the database; • It started with algorithms such as association and Naive Bayes (version 9i) and with the 10g version it includes a great variety of algorithms; • Integrates Java Data Mining API, a Java package for including the data mining facilities into the user's applications.
Angoss Knowledge STUDIO	• Presents algorithms for building decision trees, cluster analysis (grouping) and predictive models; • Allows users to exploit data in different forms; • Offers powerful visualization tools of the results that make it very user friendly; • It is compatible with other databases such as Microsoft SQL Server, and can interact with them at datamining level.
KXEN	• Has algorithms for regression, time series analysis, classification, etc. • Implements procedures for working with OLAP data cubes; • It can retrieve data from spreadsheet programs like Microsoft Excel.

By analysing the previous table it can be seen that each product described has different features and for this reason for a certain case of a company it has to find

which one best meets its requirements.

6. Data mining tools used in the business environment

A study realized by Daniel Andersson and Hannes Fries [1] tries to analyze if companies use data mining tools to improve decision-making capacity. The study presented here is based on a series of 95 large companies from Sweden, that are chosen at random in order to see how many of them use a data mining system and how their managers see it.

The results of the study show that only 30% of the analyzed companies used a data mining system at that time (Fig. 5.) while more than 50% of their managers were seeking with interest to the developments in this area and intended to acquire such a solution in the near future (Fig. 6.). The same study predicted that future data mining systems market would grow by 100% over the next four years.

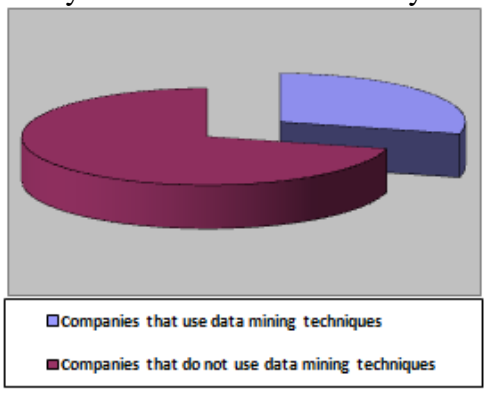

Fig. 5. Percentage of companies using data mining techniques

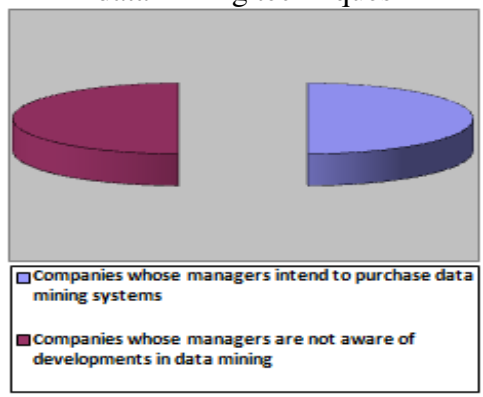

Fig. 6. Percentage of companies whose management is aware of developments in data mining and intend to implement such a system.

We believe that the results of this study can not be extended to other countries others than Sweden without taking into account the differences in technological development.

Therefore we can assume that in countries like those from Eastern Europe which are less technologicaly advanced the percentage of using data mining technology is much lower than the previously mentioned study indicates, but also the market potential could be greater. Also we must consider the fact that no small firms were considered in the study and that for many of them the use of data mining products is seen more as an option and not as something that would bring immediate profit. This leads us to believe that the data mining products market is still in its infancy and big developments in this domain can be expected during the next period of time [2]. Manufacturers of such systems will need to aim their products also towards small firms that have a great potential and are in an increased number this meaning they could be an important market.

7. Conclusions

For a business that wants to use a data mining system to optimize its activities there are available on the market various data mining commercial systems that can be integrated into that company's computer system, each of them having its own features.

By putting side by side different commercially available data mining products one can conclude that a company must first evaluate its needs regarding the data mining analysis and find the product that best covers them and fits the available budget.

There are a number of factors to be considered by a company before deciding which data mining system to use that are presented previously in this paper, like the amount of data available, the way the data is stored or the data mining tasks to be performed.

Because until recently only large companies had access to large volumes of data and financially afforded to use data mining tools,

the data mining market was mainly aimed to them. Despite this, as studies show, only a small percentage of the big companies use such tools to optimize their activities, but the managers are generally aware of the advantages brought by technology so it is expected that in the future the data mining tools market will grow.

The data mining tools market for small companies is not as developed as that for large companies but it has a great potential. The only things that are needed for its further development are the appearance of data mining tools affordable enough and the awareness of the small companies' managers.

References

[1] D. Andersson, H Fries: *Data Mining Maturity. A Quantitative Study of Large Companies in Sweden*, Jonkoping University, Master's Thesis in Informatics, 2008;

[2] M. ANDRONIE - *Modern data mining techniques,* The Ninth International Conference on Informatics in Economy IE, Bucureşti 2009, p. 753-757, ISBN 978-606-505-172-2;

[3] D. Hand, H. Mannila, P. Smith: *Principles of Data Mining*, Prentice Hall India, 2006;

[4] J. Han, M. Kamber: *Data Mining. Concepts and Techniques*, *Second Edition,* Elsevier, 2006;

[5] M. Andronie, M. ANDRONIE: *Analiza datelor stocate in depozite mari de date*, Sesiunea de comunicari stiintifice a cadrelor didactice din facultãtile economice ale Universitãtii Spiru Haret, Bucureşti, 2008, ISBN 978-973-163-230-8;

[6] Z. Hui Tang, J. MacLennan: *Data Mining with SQL Server 2005,* Wiley, 2005.

Mihai Andronie is a graduate of the Faculty of Automatic Control and Computers, of the Politehnica University Bucharest in 2006. He is currently a PHD student at the Academy of Economic Studies of Bucharest. His domains of work are: informatics systems and databases. He participated as an author at the book "Administrarea bazelor de date" (Database management) (2008) and published several papers on domains like data mining, economic process optimization and others. He also participated at national and international conferences on his domain of activity.

Daniel Crişan received a M.Sc. degree in Informatics from Ecole Polytechnique, France in 2008 and in Computer Science from EPF Lausanne, Switzerland in 2010. He is currently a PhD student in the Department of Information Technology and Electrical Engineering of ETH Zürich, Switzerland. His research interests include datacenter networking, routing and congestion management. He contributed to a novel routing scheme that efficiently exploits path diversity in datacenter networks. He co-authored a paper describing this method which was ranked #1 at IEEE Symposium on High-Performance Interconnects 2010.