The Need to Use Data Mining Techniques in E-Business

Claudia Elena Dinucă[1], Dumitru Ciobanu[2]

Abstract: The number of Internet users rose from 400 million in 2000 to just over 2 billion in early 2011. This means that approximately one third of the world's population uses the internet. Taking these conditions into consideration, we can say that businesses have changed their way. Many companies that, over the last century could not even dream that could have a certain volume of activity or they could face competition with industry giants, have succeeded in giving to enjoy great success. For example: Amazon.com, founded in 1995, had in 1999 a turnover of at least 13 times higher than other prestigious names in the U.S., such as Barnes & Noble and Borders Books & Music. E-business is the key to make life easier for the people. Knowledge of e-business environment is essential for doing business in this century. More must be understood and new technologies applied to extract knowledge from data.

Keywords: data mining; clustering; regression; association rule; e-business

JEL Classification: M15; M21; M20

1. Introduction

Historically, the notion of determining patterns (understandable information) of data was given a variety of names by statisticians and community professionals working with databases and data mining (data mining), knowledge of data mining, discovery information, harvesting information, data archeology and processing forms (patterns) of data. The Knowledge Discovery System which is able to operate on large scale database system is called Knowledge Discovery in Databases System-KDD.

KDD term first appeared in 1989. By definition, KDD is „a non-trivial process of extracting information, previously unknown and potentially useful data” but as „the science of extracting useful information from massive data or databases” according

[1] PhD Student, University of Craiova, Faculty of Economic and Business Administration, Romania, Address: A. I. Cuza, no. 13, Craiova, 200585, Romania, Tel: +4(251) 411317, Corresponding author: clauley4u@yahoo.com.
[2] PhD Student, University of Craiova, Faculty of Economic and Business Administration, Romania, Address: A. I. Cuza, no. 13, Craiova, 200585, Romania, Tel: +4(251) 411317, e-mail: ciobanubebedumitru@yahoo.com.

to Fayyad and others, 1996. In this context, the data is a collection of facts, and the model is a higher level of expression that describes the data or a subset there of. The data analysis features of models that KDD identifies must be valid, novelty, without repetitions**,** useful and ultimately understandable. A model is correctly describing the data with some degree of safety. Finally, it is desirable that the models found to be understood so being further analyzed to study the causes and effects. Because Data Mining is the central part of the process of knowledge discovery from databases (KDD), the terms data mining and knowledge discovery in databases were used alternately for many researchers in the field. Lately, however, is a clear distinction between the two terms. The distinction is related to that of knowledge discovery in databases (KDD) can be considered as the extraction of useful and interesting information from the database. The authors distinguish between DM and KDD as KDD is considered an iterative and interactive complex process that includes DM. KDD refers to the process of discovering useful knowledge from data, while data mining refers to a particular step in this process. Data mining is the application of specific algorithms for extracting patterns (models) of data.

As a consequence of the dismissal of large reservoirs of data has developed data mining. Collecting data in various formats, digitization began in the 60s allowing a retrospective analysis of data by computer. In the 80s came relational databases with Structured Query Language (SQL) and application that allows dynamic data analysis. 90s are characterized by an explosion of data. To store them it began to use data warehouses. In response to the challenges faced by the community of specialists in database data mining appeared, dealing with massive amounts of data, applying statistical analysis and search techniques specific to artificial intelligence on the data. The role of data mining is the extraction of new knowledge, implicit and direct action of large data collections, discovering things that are not obvious from the data, which cannot be extracted manually, representing useful information that can improve the current action.

2. Knowledge Discovery in Database Process

According to Fayyad and his colleagues (1996), KDD is the process of using database along with the steps required as select, pre-processing, transformation of data to apply data-mining methods (algorithms) in order to obtain patterns of data and evaluate data mining process to identify the subset of patterns listed as knowledge. KDD process is divided into seven steps as follows:

1. *Domain analysis* **-** the nature of field data are analyzed and defined target discovery. If they are previous knowledge in this area, these are evaluated.

2. *Selection* - or segmentation of data in accordance with certain criteria, which may mean removing some fields or rows of data, or both.

3. *Preprocessing* - data cleansing stage where certain information is removed, also determine ways of working with missing data fields.

4. *Transformation* - the data is processed. A representation of the data that is compatible with data-mining algorithm that is to be implemented is done at this stage. The data is analyzed to determine the characteristics to represent data in accordance with the target to be reached.

5. *Data mining* - this step takes care of extracting the data model. For this purpose we use a data-mining algorithm properly. The quality of this phase depends heavily on the previous stages.

6. *Interpretation and evaluation* - identified system models, following the algorithm applied, interpreted in the knowledge that can be used to support decisions made by humans, such predictions and classification problems, summarizing database content and explaining observed phenomena.

7. *Enhancing Knowledge Discovered* - models (patterns) found is put in use. A plausible way to use is the incorporation of knowledge obtained in another system for further action, documentation and transmission of models to stakeholders and reapply KDD database using this new knowledge as a basis.

Data mining is thus materialized by applying algorithms to extract patterns from data. Additional steps of the process of discovering knowledge from data such as data preparation, data selection, cleaning phase, the integration of previous knowledge required are in fact an essential step to ensure that will extract useful knowledge from data.

3. Data Mining Techniques

There are two fundamental classes of learning methods:

- *predictive* (based on supervised learning), which uses a set of variables (called predictors) through which predictions are made relative to the values (continuous or discrete) of other variables (called decision variables);

- *descriptive* (based on unsupervised learning), for extraction of patterns (structures understandable) of data.

Predictive models are built based on artificial intelligence in a training phase, in which the model learns to predict the right answer (decision) when the input values is formed with different sets of predictors. After consuming training phase, prediction model can be used to solve, as applicable to classification problems (if the decision variable is nominal or discrete) or regression problems (if the decision variable is continuous).

Descriptive data mining methods form the second largest category of data mining. Unlike predictive models, in descriptive methods (such as clustering) the variables are treated uniformly, without distinguishing between predictors and response (decision) as such is not supervised learning (in terms of learning from examples, that of providing responses in the training phase). Descriptive methods allow the description and explanation of the characteristic phenomena of the system studied based on the patterns found.

Association Rules

Crowds frequent articles / links can be determined if we consider the key principle of monocity or a priority which says that if a set of items (links) L is frequent (at least in the l-part of the site/click), then any subset is frequent. It uses the term frequent sets of items (frequent items etc.) for a set of articles S appearing in at least **s** part of the shopping basket/links, where **s** is a chosen constant, usually 0.01.

To determine frequent sets of articles/links must go through stages:

- Proceed at the level we find the first articles/links frequent sets of size 1, then frequent pairs, triplets common.

- Find all frequent sets of maximum articles / links (sets M so that any set strictly including M is not frequent) in one or more pass.

The method can be applied in any sector which requires finding the possible groups of products or services: banking, telecommunications services. It can be applied to study medical complications due to the combination of drugs or fraud, in which case looks for unusual combinations.

Association rules are defined as follows. Let I = {i1, i2 ,..., im} be a set of symbols, called elements. D is considered a set of transactions where each transaction T is a subset of I. Consider only the present (represented binary) elements in the transaction and does not consider other quantitative or qualitative characteristics thereof. Each transaction gets an identifier (TID). Key measures in the mining association rules are support and confidence. Support refers to the proportion in which a relationship occurs in data.

The confidente/trust of the association rule relates to the probability of finding an antecedent having a consequence. The confidence in the rule r is determined as the ratio of transactions T in D that if $X \subseteq T$ then $Y \subseteq T$ ($X \cup Y \subseteq T$), representing the conditional probability that a transaction contains Y if it contains on X.

Determination of association rules is done in two steps:

- Determination of frequent sets of elements, those that have sufficient support;

- Determination of rules of association between these sets of rules determining the strong rules. This step resolves as follows: for each frequent set X and every subset of X, $Y \subset X$ determine the parameters of the rule $X \setminus Y \rightarrow Y$ considering the outcome of the meeting left with the right side must be a frequent set, in this case $X \setminus Y \cup Y = X$. Association rules are used to find frequent sets of articles in databases that contain consumer transactions, the problem known as the market basket analysis. Market basket analysis consists of finding associations between items purchased, displayed on the receipt. It studies how the customers are doing shopping to get information on the products which tend to be purchased at the same time. In this case, the database of consumer transactions is represented by a sequence of transactions T = (t1, ... tn), and each transaction is a set of articles. For example, in the case of shopping cart it requires that trust to be significantly higher than if the items were placed randomly in cart. It can be found a rule {milk, butter} => bread on the principle that many people buy bread, but the example of beer / diaper found in the U.S. show that the rule {diapers} => {beer} is checked with a significantly higher confidence than multitude of baskets containing beer. The result of this study helps retailers in the settlement of the articles in shelves and controls how a typical buyer crosses store.

- In the case of click stream analysis, a database with server sessions, the links of a user on the website in its current session, is a sequence S = (s1 ,..., sn) of sessions, where each session is a set of links visited by the user. Determination of frequent links and association rules is essential for the click stream analysis problem, how users navigate the Internet and accessing various sites.

Sequential Association Rules

Often, transactions are recorded taking into account a temporal sequence. For example, transactions for loyalty card holders correspond to sales receipts sequence. Transactions that record navigation paths followed by a web user are associated with a temporal sequence of sessions. In such situations, analysts are keen to extract association rules that take into account temporal dependencies. Sequential analysis is used to determine patterns of data using a temporal sequence of states. The problem of discovering sequential rules was first introduced by Agrawal and Srikant in (Agrawal & Srikant, 1995, pp. 3-14).

Classification and regression are forms of supervised learning. Classification and regression are the largest category of applications, consisting of building models to forecast the membership to a set of class (classification) or to forecast of some values (regression). There are several techniques devoted to solving problems of classification and regression, including decision trees, Bayesian techniques, neural networks and k-NN enjoy wide recognition. Supervised learning techniques aim to generate automatic induction mechanisms with predictive power by extracting information contained in the database and their transformation into a knowledge base.

There are two main classes of algorithms for induction:

- Classification algorithms - when the variable is done about that prediction is qualitative (nominal or ordinal) or quantitative with discrete values;

- Regression algorithms - when the variable about which the prediction is made is still quantitative continue (it takes real values). Classification is the process of seeking common properties from objects sets of class data being classified into different classes according to a classification model. Classification allows you to create models to predict class members. The purpose of classification is primarily driven analysis based on these data and development of a model, an exact description of each class using the features of the available data. In order to be used the classifier must first learn a mapping from a set of input variables and their values to predict output values for decision variables. Classifier can be used to predict output variables values using input values once the pattern has been learned through the training data. Classification is often used in business data mining applications. For example, the classification meets in detecting fraud, where classification is trying to identify if the transaction is legal or suspect. Other examples of using the method of classification are to define customer profile analysis of ineffective treatments, medical diagnosis, and credit approvals.

Clustering *is a form of unsupervised learning which involves searching databases for input differences found between the items, and found, in the process of differentiation, groups (clusters) of objects in the input data*. Clusters are often used to change and detect of deviation aimed at finding items have data that does not fit the norm, or group (cluster). Objects in the same cluster should have similar profiles (intra-cluster homogeneity) and objects in different clusters have distinct profiles (inter-cluster heterogeneity). Change and deviation detection is applied in many fields, such as is detecting fraudulent transactions (fraud phones or bank cards), detect inappropriate drug treatment before it is too late and detect new market trends. In e-business clustering is useful because it can work with large collections of data and uses the achievement of different groups based on common objects features. It can be used before applying the method of Classification. For example, if we use the clustering method for a list of user profiles, a framework of different types of clients can be built. This clustering method has various applications in marketing, customer support and determination of fraud (if the behavior of a cell phone user immediately jumps from one cluster to another, this may indicate a phone robbery or cloning).

Clustering process involves stages of solving the following problems:

- *Lay the subject of clustering process* is a main stage which sometimes includes setting the number of classes / groups, type and scale characteristics / attributes available clustering algorithm.

- *Feature extraction* is the process of identifying the most useful attributes/ features used in the clustering. It refers to a way to make one or more transformations of input data in order to obtain new dominant features.

- *Defining a measure of proximity in a crowd.* The proximity of elements is measured by the distance function defined on pairs of elements. Similarity measures can be used to characterize the conceptual similarity between two or more items.

- *Clustering process can be accomplished in several ways.* Output data can be hard (separation of elements in clearly defined groups) or fuzzy (in which each element has a variable degree of membership of each group results)

- *Extraction of results* is the process of obtaining results in a simpler form and representative. Extraction results are a concise description of each group obtained, usually in the form of representative elements. All clustering algorithms should lead to the achievement of groups / classes for any set of inputs. If in the process of using a clustering algorithm does not get group items, then apply another algorithm that can provide better results than the previous.

- *Validity analysis* group performed an evaluation of clustering process, usually a criterion for optimization. It is checked if the results of spatial clustering are correct.

4. Applications of Data Mining Methods in E-Business

Direct Marketing. Due to the size and complexity of the current market, mass marketing has become increasingly expensive, unprofitable, so being replaced by direct marketing, which is based on selecting target groups of clients and establishing individual correlations with them on multiple channels. Thus, companies strategic are repositioned, product-centric orientation quickly transforms to a client centric.

Customer Relationship Management (CRM) target is to develop strategies to attract new customers, maintain existing ones and regaining those who migrated to other bidders. From operational point, CRM includes all activities relating to direct contact with the consumer. At the analytical level, CRM provides a number of methods for analyzing customer behavior by analyzing data obtained through transaction processing systems.

Analytical customer relationship management has three major objectives:

- *Market segmentation,* which is the division of customers into homogeneous groups based on the internal as manifested similarities (habits, tastes, affinities), this group is more heterogeneous among themselves. Thus, the firm may treat different segments of customers personalized and can be concentrated on certain target groups that correspond to some criteria of profitability.

- *Consumer profiling* involves modeling consumer behavior based on a wide range of attributes such as the geographical, cultural and ethnic, economic conditions, frequency of purchase, frequency of complaints, preferences and their degree of satisfaction, age, education, lifestyle, media used, method of recruitment that the customer response.

- *Positioning the product* in the preferences of potential customers is a marketing tool focused on identifying the most attractive features of a product to maximize customer temptation of buying it. Hence the so-called problem of shopping carts. Determine the probability that certain products are purchased together.

Association rules extracted from web logs can be used for loading and preloading web pages with a high probability of being visited [YAZL01]. For each sequence of web pages the algorithm chooses a rule left-hand combination of type-side (LHS) that matches the sequence and has the longest of all applicable rules.

In web usage mining it is necessary to understand the pattern navigate of web pages and the frequency with which different combinations of web pages are visited by a particular individual in a single session or consecutive sessions. The list of visited pages visited a session are recorded as a transaction, possibly identified by serial number and the time of visit.

It is interesting to identify regular patterns, possibly hidden in the data that allows the association of one or more pages that are visited with visits to other pages. Discovered rules may take the form such as "if an individual visits the site timesonline.co.uk it will visit within a week the site economics.com with a probability of 0.87". (Vercellis, 2009)

Association rules of this type may influence the structure and links between pages for making the navigation easier and for recommending specific ways of navigation and also placement of banners and other promotional messages. (Vercellis, 2009)

Purchases made by credit card are analyzed using association rules. Thus, association rules are used to analyze the purchases made by credit card holder to make any promotions. In this case, the transaction is made of purchases and payments made by the customer. In this situation, products and services can be accessed by a credit card holder are virtually endless.

Fraud detection. In fraud detection, operations consist of incident reports and claims for damages. There may reveal specific combinations of fraudulent behavior and thus warrants a more thorough investigation by the insurance company.

Discovery of sequential rules is an important topic in data mining with a wide variety of applications, such as buying patterns of customers predicting careers in the financial sector, to determine patterns of access on web analyzing web clicks, diseases, natural disasters, DNA sequences, etc.

5. Conclusions

In today's business world, computer use for business process and data recording has become ubiquitous. With this electronic age comes an invaluable product-data (information). Virtually every large company records all its transactions.

Data mining is the process used to make this huge volume of data and turning them into useful knowledge. Data Mining refers to the process of selection of previously unknown relationships in order to obtain a clean and useful result to that which holds the database.

As a result, a data mining system has several phases. Phases begin to turn data set and ends with knowledge extraction that occurred as a result of carrying out the steps: selection, preprocessing, transformation, data mining, interpretation and evaluation.

The origins of data mining techniques were designed as coming from three areas of learning and research: statistical, machine learning and artificial intelligence. The first foundation of data mining methods was in statistics. Statistics is the most technology that relies on data mining. Many of the statistics domains such as regression analysis, standard distributions, standard deviations and variations, the group analysis are construction techniques that advanced statistical techniques underlying data mining.

To differentiate into the Internet economy, companies must realize that winning e-business means more than simple transactions of purchase / sale, appropriate strategies are the key to improve competitive power. This can be done using data mining techniques and other statistical analysis on historical data from e-business activities.

References

Adam, Jolly (2003). *The Secure Online Business*. Kogan Page and Contributors.

Agrawal, R., Srikant, R. (1995). Mining sequential patterns. *International Conference on Data Engineering (ICDE'95).* Taiwan: Taipei, pp. 3-14.

Award, Elias (2002). *Electronic Commerce from Vision to Fulfillment*. Pearson Education. New Jersey: Upper Saddle River.

Berry, M., Linoff, G. (1997). *Data Mining Techniques for Marketing, Sales and Customer Support*. Chichester: John Wiley and Sons.

Dunham, M.H. (2003). *Data Mining: Introductory and Advanced Topics*. Prentice Hall, Pearson Education Inc.

Gunjam Santami (2002). *B2B Integration –A Practical Guide to Collaborative E-commerce*. London: Imperial College Press.

Harmon, P; Rosen, M; Guttman, M (2001). *Developing E-Business Systems & Architectures- A Manager's Guide*; SUA: Academic Press.

Janice Reynolds (2004). *The Complete E-Commerce Book: Design, Build, & Maintain a Successful Web-based Business*. Second Edition. CMP Books.

Jatinder N.D. Gupta and Sushil K. Sharma Ball and other (2004). *Intelligent Enterprises of the 21st Century* SUA: Idea Group.

Jiawei Han, Micheline Kamber (2006). *Data Mining Concepts and Techniques.* Second Edition, USA: Elsevier.

Mike Havey (2005). *Essential Business Process Modeling* SUA: O'Reilly.

Nong, Y. (2003). *The handbook of Data Mining, Lawrence Erlbaum Associates*. New Jersey: Publishers Mahwah.

Vercellis, C. (2009). *Business Intelligence: Data Mining and Optimization for Decision Making*. UK: John Wiley & Sons.

Porter, Michael E. *Competitive Strategy: Techniques for Analyzing Industries and Competitors.*

Raisinghani, M (2004). *Business Intelligence in the Digital Economy: Opportunities, Limitations, and Risks.* SUA: Idea Group Publishing.

Razvan Serbu (2004). *Comertul Electroni/Electrionic Commerce.* Sibiu: Continent.

Turban, Efraim; King, David (2003). *Introduction to E-commerce*. *Pearson Education*. New Jersey: Upper Saddle River.