DATA MINING SYSTEM AND APPLICATIONS: A REVIEW

Mr. S. P. Deshpande[1] and Dr. V. M. Thakare[2]

[1] Department of MCA, D.C.P.E, H.V.P.Mandal Amravati, India

shrinivasdeshpande68@gmail.com

[2]Post Graduate Deptt. of Computer Science,SGB, Amravati University, Amravati, India

vilthakare@yahoo.co.in

ABSTRACT:

In the Information Technology era information plays vital role in every sphere of the human life. It is very important to gather data from different data sources, store and maintain the data, generate information, generate knowledge and disseminate data, information and knowledge to every stakeholder. Due to vast use of computers and electronics devices and tremendous growth in computing power and storage capacity, there is explosive growth in data collection. The storing of the data in data warehouse enables entire enterprise to access a reliable current database. To analyze this vast amount of data and drawing fruitful conclusions and inferences it needs the special tools called data mining tools. This paper gives overview of the data mining systems and some of its applications.

Keywords:

Data mining system architecture, Data mining application

1. INTRODUCTION

To generate information it requires massive collection of data. The data can be simple numerical figures and text documents, to more complex information such as spatial data, multimedia data, and hypertext documents. To take complete advantage of data; the data retrieval is simply not enough, it requires a tool for automatic summarization of data, extraction of the essence of information stored, and the discovery of patterns in raw data. With the enormous amount of data stored in files, databases, and other repositories, it is increasingly important, to develop powerful tool for analysis and interpretation of such data and for the extraction of interesting knowledge that could help in decision-making. The only answer to all above is 'Data Mining'.

Data mining is the extraction of hidden predictive information from large databases; it is a powerful technology with great potential to help organizations focus on the most important information in their data warehouses [1,2,3,4]. Data mining tools predict future trends and behaviors, helps organizations to make proactive knowledge-driven decisions [2]. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by

DOI : 10.5121/ijdps.2010.1103

retrospective tools typical of decision support systems. Data mining tools can answer the questions that traditionally were too time consuming to resolve. They prepare databases for finding hidden patterns, finding predictive information that experts may miss because it lies outside their expectations.

Data mining, popularly known as Knowledge Discovery in Databases (KDD), it is the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases[3,5]. Though, data mining and knowledge discovery in databases (or KDD) are frequently treated as synonyms, data mining is actually part of the knowledge discovery process[1,3,5].

2 THE DATA MINING TASKS:

The data mining tasks are of different types depending on the use of data mining result the data mining tasks are classified as[1,2]:

1. Exploratory Data Analysis: It is simply exploring the data without any clear ideas of what we are looking for. These techniques are interactive and visual.

2. Descriptive Modeling: It describe all the data, It includes models for overall probability distribution of the data, partitioning of the p-dimensional space into groups and models describing the relationships between the variables.

3. Predictive Modeling: This model permits the value of one variable to be predicted from the known values of other variables.

4. Discovering Patterns and Rules: It concern with pattern detection, the aim is spotting fraudulent behavior by detecting regions of the space defining the different types of transactions where the data points significantly different from the rest.

5. Retrieval by Content: It is finding pattern similar to the pattern of interest in the data set. This task is most commonly used for text and image data sets.

3. TYPES OF DATA MINING SYSTEMS:

Data mining systems can be categorized according to various criteria the classification is as follows[3]:

- Classification of data mining systems according to the type of data source mined: This classification is according to the type of data handled such as spatial data, multimedia data, time-series data, text data, World Wide Web, etc.

- Classification of data mining systems according to the data model: This classification based on the data model involved such as relational database, object-oriented database, data warehouse, transactional database, etc.

- Classification of data mining systems according to the kind of knowledge discovered: This classification based on the kind of knowledge discovered or data mining functionalities, such as characterization, discrimination, association, classification, clustering, etc. Some systems tend to be comprehensive systems offering several data mining functionalities together.

• Classification of data mining systems according to mining techniques used: This classification is according to the data analysis approach used such as machine learning, neural networks, genetic algorithms, statistics, visualization, database oriented or data warehouse-oriented, etc.

The classification can also take into account the degree of user interaction involved in the data mining process such as query-driven systems, interactive exploratory systems, or autonomous systems. A comprehensive system would provide a wide variety of data mining techniques to fit different situations and options, and offer different degrees of user interaction.

4. DATA MINING LIFE CYCLE:

The life cycle of a data mining project consists of six phases[2,4]. The sequence of the phases is not rigid. Moving back and forth between different phases is always required. It depends on the outcome of each phase. The main phases are:

1. Business Understanding: This phase focuses on understanding the project objectives and requirements from a business perspective, then converting this knowledge into a data mining problem definition and a preliminary plan designed to achieve the objectives.

2 Data Understanding: It starts with an initial data collection, to get familiar with the data, to identify data quality problems, to discover first insights into the data or to detect interesting subsets to form hypotheses for hidden information.

3. Data Preparation: It covers all activities to construct the final dataset from the initial raw data.

4. Modeling: In this phase, various modeling techniques are selected and applied and their parameters are calibrated to optimal values.

5. Evaluation: In this stage the model is thoroughly evaluated and reviewed. The steps executed to construct the model to be certain it properly achieves the business objectives. At the end of this phase, a decision on the use of the data mining results should be reached.

6. Deployment: The purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that the customer can use it. The deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process across the enterprise.

5. THE DATA MINING MODELS:

The data mining models are of two types[1,2,6,45]: Predictive and Descriptive.

The predictive model makes prediction about unknown data values by using the known values. Ex. Classification, Regression, Time series analysis, Prediction etc.

The descriptive model identifies the patterns or relationships in data and explores the properties of the data examined. Ex. Clustering, Summarization, Association rule, Sequence discovery etc.

Many of the data mining applications are aimed to predict the future state of the data. Prediction is the process of analyzing the current and past states of the attribute and prediction of its future state. Classification is a technique of mapping the target data to the predefined groups or classes, this is a supervise learning because the classes are predefined before the examination of the target data. The regression involves the learning of function that map data item to real valued prediction variable. In the time series analysis the value of an attribute is examined as it varies over time. In time series analysis the distance measures are used to determine the similarity between different time series, the structure of the line is examined to determine its behavior and the historical time series plot is used to predict future values of the variable.

Clustering is similar to classification except that the groups are not predefined, but are defined by the data alone. It is also referred to as unsupervised learning or segmentation. It is the partitioning or segmentation of the data in to groups or clusters. The clusters are defined by studying the behavior of the data by the domain experts. The term segmentation is used in very specific context; it is a process of partitioning of database into disjoint grouping of similar tuples. Summarization is the technique of presenting the summarize information from the data. The association rule finds the association between the different attributes. Association rule mining is a two-step process: Finding all frequent item sets, Generating strong association rules from the frequent item sets. Sequence discovery is a process of finding the sequence patterns in data. This sequence can be used to understand the trend.

6. THE KNOWLEDGE DISCOVERY PROCESS:

Data mining is one of the tasks in the process of knowledge discovery from the database. The steps in the KDD process contains:[1,3]

1. Data cleaning: It is also known as data cleansing; in this phase noise data and irrelevant data are removed from the collection.

2. Data integration: In this stage, multiple data sources, often heterogeneous, are combined in a common source.

3. Data selection: The data relevant to the analysis is decided on and retrieved from the data collection.

4. Data transformation: It is also known as data consolidation; in this phase the selected data is transformed into forms appropriate for the mining procedure.

5. Data mining: It is the crucial step in which clever techniques are applied to extract potentially useful patterns.

6. Pattern evaluation: In this step, interesting patterns representing knowledge are identified based on given measures.

7. Knowledge representation: It is the final phase in which the discovered knowledge is visually presented to the user. This essential step uses visualization techniques to help users understand and interpret the data mining results.

7. DATA MINING METHODS:

The data mining methods are broadly categories as: On-Line Analytical Processing (OLAP), Classification, Clustering, Association Rule Mining, Temporal Data Mining, Time Series Analysis, Spatial Mining, Web Mining etc. These methods use different

types of algorithms and data. The data source can be data warehouse, database, flat file or text file. The algorithms may be Statistical Algorithms, Decision Tree based, Nearest Neighbor, Neural Network based, Genetic Algorithms based, Ruled based, Support Vector Machine etc. The selection of data mining algorithm is mainly depends on the type of data used for mining and the expected outcome of the mining process. The domain experts play a significant role in the selection of algorithm for data mining.

A knowledge discovery (KD) process involves preprocessing data, choosing a data-mining algorithm, and post processing the mining results. There are very many choices for each of these stages, and non-trivial interactions between them. Therefore both novices and data-mining specialists need assistance in knowledge discovery processes.

The Intelligent Discovery Assistants [7] (IDA), helps users in applying valid knowledge discovery processes. The IDA can provide users with three benefits:

1. A systematic enumeration of valid knowledge discovery processes;

2. Effective rankings of valid processes by different criteria, which help to choose between the options;

3. An infrastructure for sharing knowledge, which leads to network externalities.

Several other attempts have been made to automate this process and design of a generalized data mining tool that posses intelligence to select the data and data mining algorithms and up to some extent the knowledge discovery.

8. DATA MINING APPLICATION:

The data mining applications can be generic or domain specific. The generic application is required to be an intelligent system that by its own can takes certain decisions like: selection of data, selection of data mining method, presentation and interpretation of the result. Some generic data mining applications cannot take its own these decisions but guide users for selection of data, selection of data mining method and for the interpretation of the results. The multi agent based data mining application[8,10] has capability of automatic selection of data mining technique to be applied. The Multi Agent System used at different levels[8]: First, at the level of concept hierarchy definition then at the result level to present the best adapted decision to the user. This decision is stored in knowledge Base to use in a later decision-making. Multi Agent System Tool used for generic data mining system development[10] uses different agents to perform different tasks.

A multi-tier data mining system is proposed to enhance the performance of the data mining process[9]. It has basic components like user interface, data mining services, data access services and the data. There are three different architectures presented for the data mining system namely One-tire, Two-tire and Three-tire architecture.

Generic system required to integrates as many learning algorithms as possible and decides the most appropriate algorithm to use. CORBA (Common Object Request Broker Architecture) has features like:

Integration of different applications coded in any programming language considerably easy.

It allows reusability in a feasible way and finally it makes possible to build large and scalable system.

The data mining system architecture based on CORBA is given by Object Management Group[10] has all characteristics to accomplish a distributed and object oriented computation.

A data-centric focus and automated methodologies makes data mining accessible to non-experts[11]. The use of high-level interfaces can implement the automated methodologies that hide the data mining concepts away from the users. A data-centric design hides away all the details of mining methodology and exposes them through high-level tasks that are goal-oriented. These goal-oriented tasks are implemented using data-centric APIs. This design makes data mining task like other types of queries that users perform on the data.

In data mining better results could be obtained if large data is available. It leads to the merging and linking of local databases. A new data-mining architecture based on Internet technology addressed this problem.[12]

The context factor plays vital role in the success of data mining. The importance and meaning of same data in the different context is different. A data in one context is very important may not be much important in other context. A context-aware data-mining framework filters useful and interesting context factors, and can produce accurate and precise prediction using those factors[46].

The domain specific applications are focused to use the domain specific data and data mining algorithm that targeted for specific objective. The applications studied in this context are aimed to generate the specific knowledge. In the different domains the data generating sources generate different type of data. Data can be from a simple text, numbers to more complex audio-video data. To mine the patterns and thus knowledge from this data, different types of data mining algorithms are used. The collection and selection of context specific data and applying the data mining algorithm to generate the context specific knowledge is thus a skillful job. In many domain specific data mining applications the domain experts plays vital role to mine useful knowledge.

In the identification of foreign-accented French the audio files were used and the best 20 data mining algorithms were applied[13] the Logistic Regression model found the most robust algorithm than other algorithm.

In language research and language engineering many time extra linguistic information is needed about a text. A linguistic profile that contains large number of linguistic features can be generated from text file automatically using data mining[14]. This technique found quite effective for authorship verification and recognition. A profiling system using combination of lexical and syntactic features shows 97% accuracy in selecting correct author for the text. The linguistic profiling of text effectively used to control the quality of language and for the automatic language verification.[15] This method verifies automatically the text is of native quality. The results show that language verification is indeed possible.

In medical science there is large scope for application of data mining. Diagnosis of diesis, health care, patient profiling and history generation etc. are the few examples. Mammography is the method used in breast cancer detection. Radiologists face lot of difficulties in detection of tumors. Computer-aided methods could assist medical staff and improve the accuracy of detection[16]. The neural networks with back-propagation and association rule mining used for tumor classification in mammograms. The data mining effectively used in the diagnosis of lung abnormality that may be cancerous or benign[17]. The data mining algorithms significantly reduce patient's risks and diagnosis costs. Using the prediction algorithms the observed prediction accuracy was 100% for 91.3% cases. The use of data mining in health care is the widely used application of data mining. The medical data is complex and difficult to analyze. A

REMIND (Reliable Extraction and Meaningful Inference from Non-structured Data) system[21] integrates the structured and unstructured clinical data in patient records to automatically create high quality structured clinical data. The high quality of structuring allows existing patient records to be mined to support guidelines compliance and to improve patient care.[21]

Data mining in distance learning automatically generate useful information to enhance the learning process based on the vast amount of data generated by the tutors and student's interactions with web based distance-learning environment.[18] The Data Mining Applications transfers the data into information and feedback to the e-learning environment. This solution transforms large amounts of useless data into an intelligent monitoring and recommendation system applied to the learning process.

In Web-based Education[42] the data mining methods are used to improve courseware. The relationships are discovered among the usage data picked up during students' sessions. This knowledge is very useful for the teacher or the author of the course, who could decide what modifications will be the most appropriate to improve the effectiveness of the course.

The data mining methods are also used to provide learners with real-time adaptive feedback on the nature and patterns of their on-line communication while learning collaboratively[41]. This makes it possible to increase the awareness of learners. The application of data mining methods to educational chats is both feasible and can bring the improvement in learning environments.

Data mining facilitates software maintenance engineers to comprehend the structure of a software system and assess its maintainability.[24] The clustering algorithm effectively used to produce overviews of systems by creating mutually exclusive groups of classes, member data or methods, according to their similarities and hence reduces the time required to understand the overall system. This method also helps in discovering programming patterns and "unusual" or outlier cases which may require attention.

The anomaly detection in the Network is very difficult and needs a very close watch on the data traffic. The intrusion detection plays an essential role in computer security. The classification method of data mining is used to classify the network traffic normal traffic or abnormal traffic.[26]. If any TCP header does not belong to any of the existing TCP header clusters, then it can be considered as anomaly.

A malicious executable is threat to system's security, it damage a system or obtaining sensitive information without the user's permission. The data mining methods used to accurately detect malicious executables before they run[25]. Classification algorithms RIPPER, Naive Bayes, and a Multi-Classifier system are used to detect new malicious executables. This classifier had shown detection rate 97.76%.

Sports are ideal for application of data mining tools and techniques. In the sports world the vast amounts of statistics are collected for each player, team, game, and season. Data mining can be used by sports organizations in the form of statistical analysis, pattern discovery, as well as outcome prediction. Patterns in the data are often helpful in the forecast of future events. Data mining can be used for scouting, prediction of performance, selection of players, coaching and training and for the strategy planning[34]. The data mining techniques are used to determine the best or the most optimal squad to represent a team in a team sport in a season, tour or game.[44] The 'Cy Young Award'[30] has been presented annually to the best pitcher in the major league of baseball. The award is based largely on statistics compiled over the course of the baseball season. A Bayesian classifier is developed to predict Cy Young Award winners in American major league baseball.

The Intelligence Agencies collect and analyze information to investigate terrorist activities. One challenge to law enforcement and intelligent agencies is the difficulty of analyzing large volume of data involve in criminal and terrorist activities. Data mining makes it easy, convenient and practical to explore very large databases for organizations. The different data mining techniques are used in crime data mining.[19,33,37,43] Entity extraction used to automatically identify person, address, vehicle, narcotic drug, and personal properties from police narrative reports. Clustering techniques used to automatically associate different objects such as persons, organizations, vehicles etc. in crime records. Deviation detection is applied in fraud detection, network intrusion detection, and other crime analyses that involve tracing abnormal activities. Classification is used to detect email spamming and find authors who send out unsolicited emails. String comparator is used to detect deceptive information in criminal record. Social network analysis used to analyze criminals' roles and associations among entities in a criminal network.

The data mining system implemented at the Internal Revenue Service to identify high-income individuals engaged in abusive tax shelters[23] show significantly good results. The major lines of investigation included visualization of the relationships and data mining to identify and rank possibly abusive tax avoidance transactions.

Bankruptcy is the major threat to the banking sector[36] it increases the cost of lending. The data mining algorithms effectively used for prediction of the personal bankruptcy. Predicting bankruptcy has become the province of computer science rather than statistics. The data mining method least squares regression; neural nets and decision trees are proved to be the suitable for prediction of bankruptcy.

To enhance the quality of product data mining techniques can be used effectively. The data mining technology SAS/EM is used to discover the rules those are unknown before and it can improve the quality of products and decrease the cost. A regression model and the neural network model when applied for this purpose given accuracy above 80%.[31] The neural network model found better than the regression model.

E-commerce is also the most prospective domain for data mining[39]. It is ideal because many of the ingredients required for successful data mining are easily available: data records are plentiful, electronic collection provides reliable data, insight can easily be turned into action, and return on investment can be measured. The integration of e-commerce and data mining significantly improve the results and guide the users in generating knowledge and making correct business decisions. This integration effectively solves several major problems associated with horizontal data mining tools including the enormous effort required in pre-processing of the data before it can be used for mining, and making the results of mining actionable.

The Digital Library retrieves, collects, stores and preserves the digital data. The advent of electronic resources and their increased use in libraries has brought about significant changes in Library[40]. The data and information are available in the different formats. These formats include Text, Images, Video, Audio, Picture, Maps, etc. therefore digital library is a suitable domain for application of data mining.

Retailers have been collecting large amount of data like customer information and related transactions, product information etc. This significantly improves the applications like product demand forecasting, assortment optimization, product recommendation and assortment comparison across retailers and manufacturers[22]. To update the product details database is thus the main issue. The text mining application for extraction of implicit attributes and explicit attributes from product descriptions documents is the main task in such applications. Naive Bayes and Expectation-Maximization these two methods of data mining are used in this context.

In another application to design effective user interfaces for consumer information system data mining can be used effectively[35]. Consumers use compensatory and non-compensatory decision strategies when formulating their purchasing decisions. Compensatory decision-making strategies are used when the consumer fully rationalizes their decision outcome whereas non-compensatory decision-making strategies are used when the consumer considers only that information which has most meaning to them at the time of decision. These decision-making strategies are considered while designing online shopping support tools, and personalizing the design of the user interface. The data mining methods cluster analysis and rough sets, are used to obtain consumer information needed in support of designing customizable and personalized user interface enhancements.

Group work has an important role in many aspects of life. This makes it important for people to learn to be effective team members. Data mining is used for identifying patterns that characterize successful groups from less successful ones[20]. The data mining algorithms are used that can properly account for the temporal nature of the data and the character of group interaction. There are two way processes involved, where theories of effective group behavior can drive the data mining and, in the opposite direction, that the data mining should provide results that are meaningful to groups wishing to improve their effectiveness. A frequent sequential pattern-mining algorithm is used, which addresses the problem of discovering frequent sequences in a database with a minimum frequency called support.

The data mining tools based on compositional analyses of the protein sequences are used to analyze the genomes of representatives of three kingdoms of life, namely, archaea, eubacteria and eukaryota.[27] The exploratory Principal Component Analysis carried out to classify the proteins into clusters. Gene mapping is another application of data mining[29]. A large number of computational approaches are used for data mining, and they tend to share certain attractive characteristics for genetic association analysis. The data mining approaches mostly used are (1) classification methods that directly aim to find markers and other features that help to predict the disease status; (2) clustering techniques for finding subgroups of subjects, based on their genotypic and phenotypic similarity, and analysis of their disease association; and (3) methods based on the discovery of typical haploid types and analysis of their associations with the disease.

The Internet contains a large number of online documents available thus required an automated text and document classification systems that are capable of automatically organizing and classifying documents. There are several different data mining methods for text classification, including statistical-based algorithms, Bayesian classification, distance-based algorithms, k-nearest neighbors, decision tree-based methods etc.[28] Text classification techniques are used in many applications on web, including e-mail filtering, mail routing, Spam filtering, news monitoring, sorting through digitized paper archives, automated indexing of scientific articles, classification of news stories and searching for interesting information on the WWW.

Data mining algorithms used effectively for classification of Arabic documents[28]. Developing text classification systems for Arabic documents is a challenging task due to the complex and rich nature of the Arabic language. The N-gram frequency statistics technique employing a dissimilarity measure called the “Manhattan distance”, and Dice’s measure of similarity are used for classifying Arabic text documents.

The pharmaceutical industry is well known for performing quantitative analysis for clinical research and market research[32]. In the marketing departments data mining applications are used for sales force planning and direct marketing to doctors and consumers. Data mining techniques used quite well to a variety of critical business decisions in the pharmaceutical industry. It also used for forecasting production schedules for the manufacturing plants, determining market

potential in critical go/no decisions on continuing work on development compounds, or making financial projections for stock holders and investors on Wall Street.

The prediction in engineering applications was treated effectively by a data mining approach[17]. The prediction problems like the cost estimation problem in engineering, the problem of engineering design that involves decisions where parameters, actions, components, and so on are selected. This selection is often made based on prior data, information, or knowledge. Numerous models and algorithms have been developed for autonomous predictions based on data corresponding to different characteristics. The data mining algorithm applied on the test file with nine features has produced 100% correct predictions. Several other applications studied in this context.

9. CONCLUSION:

Most of the previous studies on data mining applications in various fields use the variety of data types range from text to images and stores in variety of databases and data structures. The different methods of data mining are used to extract the patterns and thus the knowledge from this variety databases. Selection of data and methods for data mining is an important task in this process and needs the knowledge of the domain. Several attempts have been made to design and develop the generic data mining system but no system found completely generic. Thus, for every domain the domain expert's assistant is mandatory. The domain experts shall be guided by the system to effectively apply their knowledge for the use of data mining systems to generate required knowledge. The domain experts are required to determine the variety of data that should be collected in the specific problem domain, selection of specific data for data mining, cleaning and transformation of data, extracting patterns for knowledge generation and finally interpretation of the patterns and knowledge generation.

Most of the domain specific data mining applications show accuracy above 90%. The generic data mining applications are having the limitations. From the study of various data mining applications it is observed that, no application called generic application is 100 % generic. The intelligent interfaces and intelligent agents up to some extent make the application generic but have limitations. The domain experts play important role in the different stages of data mining. The decisions at different stages are influenced by the factors like domain and data details, aim of the data mining, and the context parameters. The domain specific applications are aimed to extract specific knowledge. The domain experts by considering the user's requirements, and other context parameters guide the system. The results yield from the domain specific applications are more accurate and useful. Therefore it is conclude that the domain specific applications are more specific for data mining. From above study it seems very difficult to design and develop a data mining system, which can work dynamically for any domain.

References:

[1]Introduction to Data Mining and Knowledge Discovery, Third Edition ISBN: 1-892095-02-5, *Two Crows Corporation, 10500 Falls Road, Potomac, MD 20854 (U.S.A.), 1999.*

[2]Larose, D. T., "Discovering Knowledge in Data: An Introduction to Data Mining"*, ISBN 0-471-66657-2, John Wiley & Sons, Inc, 2005.*

[3]Dunham, M. H., Sridhar S., "Data Mining: Introductory and Advanced Topics", *Pearson Education, New Delhi, ISBN: 81-7758-785-4, 1st Edition, 2006.*

[4].Chapman, P., Clinton, J., Kerber, R., Khabaza, T.,Reinartz, T., Shearer, C. and Wirth, R.. "CRISP-DM 1.0 : Step-by-step data mining guide, NCR Systems Engineering Copenhagen (USA and Denmark), DaimlerChrysler AG (Germany), SPSS Inc. (USA) and OHRA Verzekeringenen Bank Group B.V (The Netherlands), 2000".

[5]. Fayyad, U., Piatetsky-Shapiro, G., and Smyth P., "From Data Mining to Knowledge Discovery in Databases," *AI Magazine, American Association for Artificial Intelligence, 1996.*

[6]. Tan Pang-Ning, Steinbach, M., Vipin Kumar. "Introduction to Data Mining"*, Pearson Education, New Delhi, ISBN: 978-81-317-1472-0, 3rd Edition, 2009.*

[7]. Bernstein, A. and Provost, F., "An Intelligent Assistant for the Knowledge Discovery Process", *Working Paper of the Center for Digital Economy Research, New York University and also presented at the IJCAI 2001 Workshop on Wrappers for Performance Enhancement in Knowledge Discovery in Databases.*

[8]. Baazaoui, Z., H., Faiz, S., and Ben Ghezala, H., "A Framework for Data Mining Based Multi-Agent: An Application to Spatial Data, volume 5, ISSN 1307-6884," *Proceedings of World Academy of Science, Engineering and Technology, April 2005.*

[9]. Rantzau, R. and Schwarz, H., "A Multi-Tier Architecture for High-Performance Data Mining**,** A Technical Project Report of ESPRIT project, The consortium of CRITIKAL project, Attar Software Ltd. (UK), Gehe AG (Denmark); Lloyds TSB Group (UK), Parallel Applications Centre, University of Southampton (UK), BWI, University of Stuttgart (Denmark), IPVR, University of Stuttgart (Denmark)".

[10]. Botia, J. A., Garijo, M. y Velasco, J. R., Skarmeta, A. F., "A Generic Data mining System basic design and implementation guidelines*", A Technical Project Report of CYCYT project of Spanish Government. 1998. Web Site: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.53.1935*

[11] Campos, M. M., Stengard, P. J., Boriana, L. M., "Data-Centric Automated Data Mining", , Web Site.: www.oracle.com/technology/products/bi/odm/pdf/automated_data_mining_paper_1205.pdf

[12].Sirgo, J., Lopez, A., Janez, R., Blanco, R., Abajo, N., Tarrio, M., Perez, R., "A Data Mining Engine based on Internet, Emerging Technologies and Factory Automation," *Proceedings ETFA '03, IEEE Conference, 16-19 Sept. 2003. Web Site :* www.citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.11.8955

[13] Bianca V. D.,Philippe Boula de Mareüil and Martine Adda-Decker, "Identification of foreign-accented French using data mining techniques, Computer Sciences Laboratory for Mechanics and Engineering Sciences (LIMSI)". Web Site : *www.limsi.fr/Individu/bianca/article/Vieru&Boula&Madda_ParaLing07.pdf*

[14]Halteren, H. van, "Linguistic Profiling for Author Recognition and Verification", *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics USA**,** Barcelona, Spain, Article No. 199, Year of Publication: 2004.*

[15]. Halteren, H. V., Oostdijk N., "Linguistic profiling of texts for the purpose of language verification, The ILK research group, Tilburg centre for Creative Computing and the Department of Communication and Information Sciences of the Faculty of Humanities, Tilburg University, The Netherlands." WebSite: *www.ilk.uvt.nl/~antalb/textmining/LingProfColingDef.pdf*

[16]. Antonie, M. L., Zaiane, O. R.,Coman, A., "Application of Data Mining Techniques for Medical Image Classification", *Proceedings of the Second International Workshop on Multimedia Data Mining (MDM/KDD 2001) in conjunction with ACM SIGKDD conference, San Francisco, August 26, 2001.*

[17].Kusiak, A., Kernstine, K.H., Kern, J.A., McLaughlin, K.A., and Tseng, T.L., "Data Mining: Medical and Engineering Case Studies". *Proceedings of the Industrial Engineering Research 2000 Conference, Cleveland, Ohio, pp. 1-7,May 21-23, 2000.*

[18].Luis, R., Redol, J., Simoes, D., Horta, N., "Data Warehousing and Data Mining System Applied to E-Learning**,** Proceedings of the II International Conference on Multimedia and Information & Communication Technologies in Education, Badajoz, Spain, December 3-6th 2003.

[19].Chen, H., Chung, W., Qin, Y., Chau, M., Xu, J. J., Wang, G., Zheng, R., Atabakhsh, H., "Crime Data Mining: An Overview and Case Studies", *A project under NSF Digital Government Programme, USA, "COPLINK Center: Information and Knowledge Management for Law Enforcement,", July 2000 – June 2003*.

[20].Kay, J., Maisonneuve, N., Yacef, K., Zaiane O., "Mining patterns of events in students' teamwork data", *Proceedings of the ITS (Intelligent Tutoring Systems) 2006 Workshop on Educational Data Mining*, pages 45-52, Jhongli, Taiwan, 2006.

[21]Rao, R. B., Krishnan, S. and Niculescu, R. S., "Data Mining for Improved Cardiac Care" , *SIGKDD Explorations Volume 8, Issue 1.*

[22]Ghani, R., Probst, K., Liu, Y., Krema, M., Fano, A., "Text Mining for Product Attribute Extraction**",** *SIGKDD Explorations Volume 8, Issue 1.*

[23]DeBarr, D., Eyler-Walker, Z., "Closing the Gap: Automated Screening of Tax Returns to Identify Egregious Tax Shelters". *SIGKDD Explorations Volume 8, Issue 1.*

[24]Kanellopoulos, Y., Dimopulos, T., Tjortjis, C., Makris, C. "Mining Source Code Elements for Comprehending Object-Oriented Systems and Evaluating Their Maintainability", *SIGKDD Explorations Volume 8, Issue 1.*

[25]Schultz, M. G., Eskin, Eleazar, Zadok, Erez, and Stolfo, Salvatore, J., "Data Mining Methods for Detection of New Malicious Executables". *Proceedings of the 2001 IEEE Symposium on Security and Privacy, IEEE Computer Society Washington, DC, USA , ISSN:1081-6011, 2001.*

[26] Cai, W. and Li L., "Anomaly Detection using TCP Header Information, STAT753 Class Project Paper, May 2004.". Web Site:*http://www.scs.gmu.edu/~wcai/stat753/stat753report.pdf.*

[27].Nandi, T., Rao, C. B. and Ramchandran, S., "Comparative genomics using data mining tools, Journal of Bio-Science, Indian Academy of Sciences, Vol. 27,No. 1, Suppl. 1, page No. 15-25, February 2002".

[28]Khreisat, L., "Arabic Text Classification Using N-Gram Frequency Statistics A Comparative Study". *Proceedings of The 2006 International Conference on Data Mining, DMIN'06, pp 78-82, Las Vegas, Nevada, USA, June 26-29, 2006.*

[29] Onkamo, P. and Toivonen, H., "A survey of data mining methods for linkage dis-equilibrium mapping", *Henry Stewart Publications 1473 – 9542. Human Genomics. VOL 2, NO 5, Page No. 336–340, MARCH 2006.*

[30] Smith, L., Lipscomb, B., and Simkins, A., "Data Mining in Sports: Predicting Cy Young Award Winners". *Journal of Computer Science, Vol. 22, Page No. 115-121,April 2007.*

[31]Deng, B., Liu, X., "Data Mining in Quality Improvement". *Proceedings of the Twenty-Seventh Annual SAS Users Group International Conference 2002 by SAS Institute Inc., Cary, NC, USA. ISBN 1-59047-061-3. Web Site :http://www2.sas.com/proceedings/sugi27/Proceed27.pdf*

[32]Cohen, J. J., Olivia, C., Rud, P., "Data Mining of Market Knowledge in The Pharmaceutical Industry". *Proceeding of 13th Annual Conference of North-East SAS Users Group Inc., NESUG2000, Philadelphia Pennsilvania, September 24-26 2000.*

[33]Elovici, Y., Kandel, A., Last, M., Shapira, B., Zaafrany, O., "Using Data Mining Techniques for Detecting Terror-Related Activities on the Web". Web Site: *www.ise.bgu.ac.il/faculty/mlast/papers/JIW_Paper.pdf*

[34] Solieman, O. K., "Data Mining in Sports: A Research Overview, A Technical Report, MIS Masters Project, August 2006". Web Site: *http://ai.arizona.edu/hchen/chencourse/Osama-DM_in_Sports.pdf*

[35] Maciag, T., Hepting, D. H., Slezak, D., Hilderman, R. J., "Mining Associations for Interface Design". *Lecture Notes in Computer Science, Springer Berlin / Heidelberg, Volume 4481, pp. 109-117, June 26, 2007.*

[36]Foster, D. P. and Stine, R. A., "Variable Selection in Data Mining: Building a Predictive Model for Bankruptcy". *Journal of the American Statistical Association, Alexandria, VA, ETATS-UNIS, vol. 99,* ***ISSN*** *0162-1459, pp. 303-313 January 15, 2004*

[37]Kraft, M. R., Desouza, K. C., Androwich, I., "Data Mining in Healthcare Information Systems: Case Study of a Veterans' Administration Spinal Cord Injury Population". *IEEE, Proceedings of the 36th Hawaii International Conference on System Sciences, 0-7695-1874-5/03, 2002.*

[38]Kusiak, A., Kernstine, K. H., Kern, J. A., McLaughlin, K. A., and Tseng, T. L., "Data Mining: Medical and Engineering Case Studies". *Proceedings of the Industrial Engineering Research 2000 Conference, Cleveland, Ohio, pp. 1-7,May 21-23, 2000.*

[39]Ansari, S., Kohavi, R., Mason, L., and Zheng, Z., "Integrating E-Commerce and Data Mining: Architecture and Challenges". *Proceedings of IEEE International Conference on Data Mining, 2001.*

[40] Jadhav, S. R., and Kumbargoudar, P., "Multimedia Data Mining in Digital Libraries: Standards and Features".*Proceedings of conference Recent advances in Information Science and Technology READIT – 2007, pp 54-59, Organized by Madras Library Association - Kalpakkam Chapter & Scintific information Resource Division, Indira Gandhi Center for Atomic research, Department of Atomic Energy, Kalpakkam, Tamilnadu,India. 12-13 July 2007.*

[41]Anjewierden, A., Koll¨offel, B., and Hulshof C., "Towards educational data mining: Using data mining methods for automated chat analysis to understand and support inquiry learning processes". *International Workshop on Applying Data Mining in e-Learning, ADML'07, Vol-305, Page No 23-32, Sissi, Lassithi - Crete Greece, 18 September, 2007.*

[42]Romero, C., Ventura, S. and De-Bra, P. "Knowledge Discovery with Genetic Programming for Providing Feedback to Courseware Authors, Kluwer Academic Publishers, Printed in the Netherlands, 30/08/2004".

[43]Chen, H., Chung, W., Xu Jennifer, J., Wang, G., Qin, Y., Chau, M., "Crime Data Mining: A General Framework and Some Examples". *Technical Report, Published by the IEEE Computer Society, 0018-9162/04, pp 50-56, April 2004.*

[44]Chodavarapu Y., "Using data-mining for effective (optimal) sports squad selections". *Web Site: http://insightory.com/view/74//using_data-mining_for_effective_(optimal)_sports_squad_selections*

[45]Jensen, Christian, S., "Introduction to Temporal Database Research," Web site: http://www.cs.aau.dk/~csj/Thesis/pdf/chapter1.pdf

[46]Vajirkar, P., Singh, S., and Lee, Y., "Context-Aware Data Mining Framework for Wireless Medical Application". *Lecture Notes in Computer Science (LNCS), Volume 2736, Springer-Verlag. ISBN 3-540-40806-1, pp. 381 – 391.*

* * * * *