A REVIEW OF DATA MINING SYSTEM AND ITS APPICATION

G.SATHYA[#1] and S.ANANDANARAYANAN[*2]

[#] *Research Scholar, PG and Research Department of Computer Science Shanmuga Industries Arts and Science College Tiruvannamalai*

[*] *H.O.D., Shanmuga Industries Arts and Science College, Tiruvannamalai*

***Abstract*— Information plays a main role in every human life. It is very important to gather data from different data sources and maintain the data, generate information, and also 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 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 data mining systems ant its some application.**

***Index Terms*— Data mining system, Data mining application**

I. Introduction

To generate information it requires large 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. The data retrieval is simply not enough but it requires a tool for automatic summarization of data, extraction of the essence of information stored, and the discovery of patterns in raw data. The enormous amount of data stored in files, databases and other repositories, to develop powerful tool for analysis and interpretation of such data and for extraction of interesting knowledge that could help in decision making. Answer of all above is “Data Mining”.

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

II. 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. 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.

III. TYPES OF DATA MINING SYSTEMS:

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

- 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.

IV. THE DATA MINING MODELS:

The data mining models are of two types: 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...

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. 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. 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.

V. DATA MINING LIFE CYCLE:

The life cycle of a data mining project consists of six phases. 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

VI. 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 contain:

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.

VII. 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 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 (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 possess intelligence to select the data and data mining algorithms and up to some extent the knowledge discovery.

VIII. DATA MINING APPLICATION:

A multi-tier data mining system is proposed to enhance the performance of the data mining process. 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 integrate 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 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. The neural networks with back-propagation and association rule mining used for tumor classification in mammograms. 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 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.

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 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.

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. 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.

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.

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. E-commerce is also the most prospective domain for data mining. 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.

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. 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.

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. 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.

IX. CONCLUSION:

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. 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 is 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. International Journal of Distributed and Parallel systems (IJDPS) Vol.1, No.1, September 2010

[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".