SURVEY OF DATA MINING AND WAREHOUSING

Prem kumar[#1], Dhanabakyam[*2]

[#1] Prem kumar Department of Computer Applications,
K.S.G College of Arts and Science,
Bharathiar University
Coimbatore ,Tamilnadu, India
[1]premkumar.mss@gmail.com

[*]Dhanabagyam
Assitant Professor
Department of Computer Applications,
Sri Jeyendra Saraswathy Maha Vidyalaya College of Arts And Science,
Bharathiar University
Coimbatore ,Tamilnadu, India
[2]dhana.ganesh09@gmail.com

***Abstract*— Data and Information or Knowledge has a significant role on human activities. Data mining is the knowledge discovery process by analyzing the large volumes of data from various perspectives and summarizing it into useful information. Due to the importance of extracting knowledge/information from the large data repositories, data mining has become an essential component in various fields of human life including business, education, medical, scientific etc..**

***Keywords*— Introduction, definition, six steps in data mining, types of data mining, trends in data mining, applications of data mining, needs for data mining, conclusion.**

I.INTRODUCTION

The data collected from different applications require proper mechanism of extracting knowledge/information from large repositories for better decision making. Knowledge discovery in databases (KDD), often called data mining, aims at the discovery of useful information from large collections of data[1]. Data mining (sometimes called data or knowledge discovery) is the process of analysing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. Data mining software is one of a number of analytical tools for analysing data. It allows users to analyse data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases.

II.DEFINITION OF DATA MINING

"Data Mining represents a process developed to examine large amounts of data routinely collected [2]. The term also refers to a collection of tools used to perform the process. Data mining is used in most areas where data are collected-marketing, health, communications, etc."

Example, one Midwest grocery chain used the data mining capacity of ORACLE software to analyse local buying patterns. They discovered that when men bought diapers on Thursdays and Saturdays, they also tended to buy beer. Further analysis showed that these shoppers typically did their weekly grocery shopping on International Journal of On Thursdays, however, they only bought a few items. The retailer concluded that they purchased the beer to have it available for the upcoming weekend. The grocery chain could use this newly discovered information in various ways to increase revenue. For example, they could move the beer display closer to the diaper display. And, they could make sure beer and diapers were sold at full price on Thursdays.

III.THE SIX-STEPS IN DATA MINING PROCESS

The six-step DMKD process[5] is described as follows:

A. *Understanding the problem domain.*

In this step one works closely with domain experts to define the problem and determine the project goals, identify key people, and learn about current solutions to the problem. It involves learning domain-specific terminology. A description of the problem including its restrictions is done. The project goals then need to be translated into the DMKD goals, and may include initial selection of potential DM tools.

B. *Understanding the data.*

This step includes collection of sample data, and deciding which data will be needed including its format and size. If a background knowledge does exist some attributes may be ranked as more important. Next, we need to verify usefulness

of the data in respect to the DMKD goals. Data needs to be checked for completeness, redundancy, missing values, plausibility of attribute values, etc.

C. *Preparation of the data.*

This is the key step upon which the success of the entire knowledge discovery process depends; it usually consumes about half of the entire project effort. In this step, we decide which data will be used as input for data mining tools of step 4. It may involve sampling of data, running correlation and significance tests, data cleaning like checking completeness of data records, removing or correcting for noise, etc. The cleaned data can be further processed by feature selection and extraction algorithms (to reduce dimensionality), by derivation of new attributes (say by discretization), and by summarization of data (data granularization). The result would be new data records, meeting specific input requirements for the planned to be used DM tools.

D. *Data mining.*

This is another key step in the knowledge discovery process. Although it is the data mining tools that discover new information, their application usually takes less time than data preparation. This step involves usage of the planned data mining tools and selection of the new ones. Data mining tools include many types of algorithms, such as rough and fuzzy sets, Bayesian methods, evolutionary computing, machine learning, neural networks, clustering, pre-processing techniques, etc.

E. *Evaluation of the discovered knowledge.*

This step includes understanding the results, checking whether the new information is novel and interesting, interpretation of the results by domain experts, and checking the impact of the discovered knowledge. Only the approved models (results of applying many data mining tools) are retained. The entire DMKD process may be revisited to identify which alternative actions could have been taken to improve the results. A list of errors made in the process is prepared.

F. *Using the discovered knowledge.*

This step is entirely in the hands of the owner of the database. It consists of planning where and how the discovered knowledge will be used. The application area in the current domain should be extended to other domains. A plan to monitor the implementation of the discovered knowledge should be created, and the entire project documented.

IV. TYPES OF DATA MINING

There are two types of data mining, the are

- Descriptive data mining
- Predictive data mining

G. *Descriptive data mining*

1) Summarization and visualization

Before you can build good predictive models, you must understand your data. Start by gathering a variety of numerical summaries (including descriptive statistics such as averages, standard deviations and so forth) and looking at the distribution of the data. You may want to produce cross tabulations (pivot tables) for multi-dimensional data.
Data can be continuous, having any numerical value (e.g., quantity sold) or categorical, fitting into discrete classes (e.g., red, blue, green). Categorical data can be further defined as either ordinal, having a meaningful order (e.g., high/medium/low), or nominal, that is unordered (e.g., postal codes).

Graphing and visualization tools are a vital aid in data preparation and their importance to effective data analysis cannot be overemphasized. Data visualization most often provides the Aha! leading to new insights and success. Some of the common and very useful graphical displays of data are histograms or box plots that display distributions of values. You may also want to look at scatter plots in two or three dimensions of different pairs of variables. The ability to add a third, overlay variable greatly increases the usefulness of some types of graphs.

Visualization works because it exploits the broader information bandwidth of graphics as opposed to text or numbers. It allows people to see the forest and zoom in on the trees. Patterns, relationships, exceptional values and missing values are often easier to perceive when shown graphically, rather than as lists of numbers and text.

The problem in using visualization stems from the fact that models have many dimensions or variables, but we are restricted to showing these dimensions on a two-dimensional computer screen or paper. For example, we may wish to view the relationship between credit risk and age, sex, marital status, own-or-rent, years in job, etc. Consequently, visualization tools must use clever representations to collapse n dimensions into two. Increasingly powerful and sophisticated data visualization tools are being developed, but they often require people to train their eyes through practice in order to understand the information being conveyed. Users who are colour-blind or who are not spatially oriented may also have problems with visualization tools.

2) Clustering

Clustering divides a database into different groups. The goal of clustering is to find groups that arevery different from each other, and whose members are very similar to each other. Unlike classification (see Predictive Data Mining, below), you don't know what the clusters will be when you start, or by which attributes the data will be clustered. Consequently, someone who is knowledgeable in the business must interpret the clusters. Often it is necessary to modify the clustering by excluding variables that have been employed to group instances, because upon examination the user identifies them as irrelevant or not meaningful. After you have found clusters

that reasonably segment your database, these clusters may then be used to classify new data. Some of the common algorithms used to perform clustering include Kohonen feature maps and K-means.

Don't confuse clustering with segmentation. Segmentation refers to the general problem of identifying groups that have common characteristics. Clustering is a way to segment data into groups that are not previously defined, whereas classification is a way to segment data by assigning it to groups that are already defined.

3) Link analysis

Link analysis is a descriptive approach to exploring data that can help identify relationships among values in a database. The two most common approaches to link analysis are association discovery and sequence discovery. Association discovery finds rules about items that appear together in an event such as a purchase transaction. Market-basket analysis is a well-known example of association discovery. Sequence discovery is very similar, in that a sequence is an association related over time.

Associations are written as A => B, where A is called the antecedent or left-hand side (LHS), and B is called the consequent or right-hand side (RHS). For example, in the association rule "If people buy a hammer then they buy nails," the antecedent is "buy a hammer" and the consequent is "buy nails."

It's easy to determine the proportion of transactions that contain a particular item or item set: simply count them. The frequency with which a particular association (e.g., the item set "hammers and nails") appears in the database is called its support or prevalence. If, say, 15 transactions out of 1,000 consist of "hammer and nails," the support for this association would be 1.5%. A low level of support (say, one transaction out of a million) may indicate that the particular association isn't very important — or it may indicated the presence of bad data (e.g., "male and pregnant").

H. *Predictive data mining*

1) Classification

Classification problems aim to identify the characteristics that indicate the group to which each case belongs. This pattern can be used both to understand the existing data and to predict how new instances will behave. For example, you may want to predict whether individuals can be classified as likely to respond to a direct mail solicitation, vulnerable to switching over to a competing longdistance phone service, or a good candidate for a surgical procedure.

Data mining creates classification models by examining already classified data (cases) and inductively finding a predictive pattern. These existing cases may come from an historical database, such as people who have already undergone a particular medical treatment or moved to a new long distance service. They may come from an experiment in which a sample of the entire database is tested in the real world and the results used to create a classifier. For example, a sample of a mailing list would be sent an offer, and the results of the mailing used to develop a classification model to be applied to the entire database. Sometimes an expert classifies a sample of the database, and this classification is then used to create the model which will be applied to the entire database.

2) Regression

Regression uses existing values to forecast what other values will be. In the simplest case, regression uses standard statistical techniques such as linear regression. Unfortunately, many real-world problems are not simply linear projections of previous values. For instance, sales volumes, stock prices, and product failure rates are all very difficult to predict because they may depend on complex interactions of multiple predictor variables. Therefore, more complex techniques (e.g., logistic regression, decision trees, or neural nets) may be necessary to forecast future values.

The same model types can often be used for both regression and classification. For example, the CART (Classification And Regression Trees) decision tree algorithm can be used to build both classification trees (to classify categorical response variables) and regression trees (to forecast continuous response variables). Neural nets too can create both classification and regression models.

3) Time series

Time series forecasting predicts unknown future values based on a time-varying series of predictors. Like regression, it uses known results to guide its predictions. Models must take into account the distinctive properties of time, especially the hierarchy of periods (including such varied definitions as the five- or seven-day work week, the thirteen-"month" year, etc.), seasonality, calendar effects such as holidays, date arithmetic, and special considerations such as how much of the past is relevant.

V.TRENDS IN DATA MINING

As different types of data are available for data mining tasks, so data mining approaches poses many challenging research issues in data mining. The design of a standard data mining languages, the development of effective and efficient data mining methods and systems, the construction of interactive and integrated data mining environments, and the applications of data mining to solve large applications large application problems are important tasks for data mining researches and data mining system and application developers. Here we will discuss some of the trends in data mining that reflect the pursuit of these challenges:

I. *Application Exploration*

Earlier data mining was mainly used for business purpose, to overcome the competitors. But as data mining is becoming more popular it is gaining wide acceptance in other fields also such as biomedicine, stock market, fraud detection,

telecommunication and many more. And many new explorations are being done for this purpose. In addition for data mining for business continues to expand as e-commerce and marketing becomes mainstream elements of the retail industry.

J. *Scalable data mining methods*

The current data mining methods capable of handling only a particular type of data and limited amount of data, but as data is expanding at a massive rate, there is a need to develop new data mining methods which are scalable and can handle different types of data and large volume of data. The data mining methods should be more interactive and user friendly. One important direction towards improving the repair efficiency of the timing process while increasing user interaction is constraint-based mining. This provide user with more control by allowing the specification and use of constraints to guide data mining systems in their search for interesting patterns.

K. *Combination of data mining with database systems, data warehouse systems, and web database systems*

Database systems, data warehouse systems, and WWW are loaded with huge amounts of data and have thus become the major information processing systems. It is important to make sure that data mining serves as essential data analysis component that can be easily included in to such an information-processing environment. The desired architecture for data mining system is the tight coupling with database and data warehouse systems. Transaction management query processing, online analytical processing and online analytical mining should be integrated into one unified framework.

L. *Standardization of data mining language:*

Today few data mining languages are commercially available in the market like Microsoft's SQL server 2005, IBM Intelligent Miner, SAS Enterprise Miner, SGI Mine set, Clementine , DBMiner and many more but a standard data mining language or other standardization efforts will provide the orderly development of data mining solutions, improved interpretability among multiple data mining systems and functions.

M. *Visual data mining*

It is rightly said a picture is worth a thousand words. So if the result of the mined data can be shown in the visual form it will further enhance the worth of the mined data. Visual data mining is an effective way to discover knowledge from huge amounts of data. The systematic study and development of visual data mining techniques will promote the use for data mining analysis. New methods for mining complex types of data The complex types of data like geospatial, multimedia, time series, sequence and text data poses an important research area in field of data mining. There is still a huge gap between the needs for these applications and the available technology.

N. *Web mining*

The World Wide Web is huge collection of globally distributed collection of news, advertisements, consumer records, financial, education, government, e-commerce and many other services. The WWW also contains huge and dynamic collection hyper linked information, providing a huge source for data mining. Based on the above facts, the web also poses great challenges for efficient resource and knowledge discovery.

O. *Business Trends*

Today's business must be more profitable, react quicker and offer high quality services that ever before. With these types of expectations and constraints, data mining becomes a fundamental technology in enabling customer's transactions more accurately. Data mining techniques of classification, regression, and cluster analysis are used for in current business trends. Most of the current business data mining applications utilize the classification and prediction techniques for supporting business decisions. In business environment data mining has evolved to Decision Support Systems (DSS) and very recently it has grown to Business Intelligence (BI) systems.

P. *Historical trends of data mining*

Data mining is useful in various disciplines, which includes database management systems (DBMS), Statistics, Artificial Intelligence (AI), and Machine Learning (ML). The era of data mining applications was conceived in the year1980 primarily by research-driven tools focused on single tasks. The early day's data mining trends are as under.

Q. *Data Trends*

In initial days, data mining algorithms work best for numerical data collected from a single data base, and various data mining techniques have evolved for flat files, traditional and relational databases where the data is stored in tabular representation. Later on, with the confluence of Statistics and Machine Learning techniques, various algorithms evolved to mine the non numerical data and relational databases.

R. *Computing Trends*

The field of data mining has been greatly influenced by the development of fourth generation programming languages and various related computing techniques. In, early days of data mining most of the algorithms employed only statistical techniques. Later on they evolved with various computing techniques like AI, ML and Pattern Reorganization. Various data mining techniques (Induction, Compression and Approximation) and algorithms developed to mine the large volumes of heterogeneous data stored in the data warehouse.

S. *Future trends*

Due to the enormous success of various application areas of data mining, the field of data mining has been establishing itself as the major discipline of computer science and has shown interest potential for the future developments. Ever

increasing technology and future application areas are always poses new challenges and opportunities for data mining, the typical future trends of data mining includes:

- Standardization of data mining languages
- Data pre-processing
- Complex objects of data
- Computing resources
- Web mining
- Scientific Computing
- Business data

VI. APPLICATIONS OF DATA MINING

As data mining matures, new and increasingly innovative applications for it emerge. Although a wide variety of data mining scenarios can be described [3]. For the purpose of this paper the applications of data mining are divided in the following categories:

• Healthcare
• Finance
• Retail industry
• Telecommunication
• Text Mining & Web Mining
• Higher Education

T. *Healthcare*

The past decade has seen an explosive growth in biomedical research, ranging from the development of new pharmaceuticals and in cancer therapies to the identification and study of human genome by discovering large scale sequencing patterns and gene functions. Recent research in DNA analysis has led to the discovery of genetic causes for many diseases and disabilities as well as approaches for disease diagnosis, prevention and treatment.

U. *Finance*

Most banks and financial institutions offer a wide variety of banking services (such as checking, saving, and business and individual customer transactions), credit (such as business, mortgage, and automobile loans), and investment services (such as mutual funds). Some also offer insurance services and stock services. Financial data collected in the banking and financial industry is often relatively complete, reliable and high quality, which facilitates systematic data analysis and data mining. For example it can also help in fraud detection by detecting a group of people who stage accidents to collect on insurance money.

V. *Retail Industry*

Retail industry collects huge amount of data on sales, customer shopping history, goods transportation and consumption and service records and so on. The quantity of data collected continues to expand rapidly, especially due to the increasing ease, availability and popularity of the business conducted on web, or e-commerce. Retail industry provides a rich source for data mining. Retail data mining can help identify customer behavior, discover customer shopping patterns and trends, improve the quality of customer service, achieve better customer retention and satisfaction, enhance goods consumption ratios design more effective goods transportation and distribution policies and reduce the cost of business.

W. *Telecommunication*

The telecommunication industry has quickly evolved from offering local and long distance telephone services to provide many other comprehensive communication services including voice, fax, pager, cellular phone, images, e-mail, computer and web data transmission and other data traffic. The integration of telecommunication, computer network, Internet and numerous other means of communication and computing are underway. Moreover, with the deregulation of the telecommunication industry in many countries and the development of new computer and communication technologies, the telecommunication market is rapidly expanding and highly competitive. This creates a great demand from data mining in order to help understand business involved, identify telecommunication patterns, catch fraudulent activities, make better use of resources, and improve the quality of service

X. *Text Mining and Web Mining*

Text mining is the process of searching large volumes of documents from certain keywords or key phrases. By searching literally thousands of documents various relationships between the documents can be established. Using text mining however, we can easily derive certain patterns in the comments that may help identify a common set of customer perceptions not captured by the other surveys questions. An extension of text mining is web mining. Web mining is an exciting new field that integrates data and text mining within a website. It enhances the web site with intelligent behaviour, such as suggesting related links or recommending new products to the consumer. Web mining is especially exciting because it enables tasks that were previously difficult to implement. They can be configured to monitor and gather data from a wide variety of locations and can analyse the data across one or multiple sites. For example the search engines work on the principle of data mining.

Y. *Higher Education*

An important challenge that higher education faces today is predicting paths of students and alumni. Which student will enrol in particular course programs? Who will need additional assistance in order to graduate? Meanwhile additional issues, enrolment management and time-to degree, continue to exert pressure on colleges to search for new and faster solutions. Institutions can better address these students and alumni through the analysis and presentation of data. Data mining has quickly emerged as a highly desirable tool for using current reporting capabilities to uncover and understand hidden patterns in vast databases.

VII. NEED FOR DATA MINING

The massive growth of data from terabytes to perabytes is due to the wide availability of data in automated form from various sources as WWW, Business, science, Society and many more. But we are drowning in data but deficient of knowledge Data is useless, if it cannot deliver knowledge. That is why data mining is gaining wide acceptance in today's world. A lot has been done in this field and lot more need to be done.

VIII. CONCLUSIONS

Since data mining is a young discipline with wide and diverse applications, there is still a nontrivial gap between general principles of data mining and domain specific, effective data mining tools for particular applications. A few application domains of Data Mining (such as finance, the retail industry and telecommunication) and Trends in Data Mining which include further efforts towards the exploration of new application areas and new methods for handling complex data types, algorithms scalability, constraint based mining and visualization methods, the integration of data mining with data warehousing and database systems, the standardization of data mining languages, and data privacy protection and security.

REFERENCES

1.*Data Mining Techniques, Arun k pujari 1st Edition.*

2. Abiteboul, S., Quass, D., McHugh, J., Widom, J., and Wiener, J., The Lorel Query Language for Semistructured Data, International Journal on Digital Libraries, 1:1, pp.68-88, 1997.

3. Data Mining Concepts and Techniques – Jiawei Han & Micheline Kamber

4. Modern Data Warehousing, Mining and Visualization Core Concepts by George M. Marakas.

5. Cios, K.J., Teresinska, A., Konieczna, S., Potocka, J., Sharma, S., Diagnosing Myocardial Perfusion from PECT Bull's-eye Maps - A Knowledge Discovery Approach, IEEE Engineering in Medicine and Biology Magazine, Special issue on Medical Data Mining and Knowledge Discovery, 19:4, pp. 17-25, 2000

6.Charu C.Aggarwal IBM T.J. Watson Research Center, USA and Philip S. "Privacy preserving data mining: Models and algorithms" Yu University of Illinois at Chicago, USA.

Authors

Mr. S.Prem kumar MSc.,(MPhil),
Assistant Professor
KSG College of Arts and Science
Bharathiar University
Coimbatore – 641 046.
premkumar.mss@gmail.com

Mrs.D. Dhanabagyam
Assistant Professor
Sri Jeyendra Saraswathy Maha Vidyalaya
College of Arts And Science,
Bharathiar University
Coimbatore – 641 046.
vijimohan_2000@yahoo.com