A survey on Data Mining: Tools, Techniques, Applications, Trends and Issues.

Vikas Gupta, Prof. Devanand

Abstract--Data mining is a process which finds useful patterns from large amount of data. It is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. It uses machine learning, statistical and visualization techniques to discovery and present knowledge in a form which is easily comprehensible to humans. This review of literature focuses on data mining techniques, trends, issues, tools, crucial concepts and applications. In this paper we have focused a variety of techniques, approaches and different areas of the research which are helpful and marked as the important field of data mining Technologies. This paper imparts more number of applications of the data mining and also focuses on trends in the data mining which will helpful in the further research.

Keywords: Data Mining (DM), Tools. Techniques, Applications, Research Issues, classification, clustering, decision tress.

1. Introduction

The field of data mining and knowledge discovery is emerging as a new, fundamental research area with important applications to science, engineering, medicine, business, and education. Data mining attempts to formulate analyze and implement basic induction processes that facilitate the extraction of meaningful information and knowledge from unstructured data [1]. Data mining uses mathematical analysis to derive patterns and trends that exist in data. 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 [2] [3]. Many organizations are combining data mining with such things as statistics, pattern recognition, and other important tools. Data mining can be used to find patterns and connections that would otherwise be difficult to find. This technology is popular with many businesses because it allows them to learn more about their customers and make smart marketing decisions [4]. The primary disadvantage to data mining is that users may discover things which are based on chances instead of a direct connection. In order for data mining to be used effectively, the users must be able to tell the difference between chance and a direct correlation. [3] Some disadvantages of DM are privacy and security issues. Information is collected through data mining intended for the ethical purposes can be misused. This information may be exploited by unethical people or businesses to take benefits of vulnerable people or discriminate against a group of people. Data mining is useless if you don't have any data to analyze. While most organizations already collect data to some extent, this is not enough if you want to use data mining successfully. The information must be specific and refined [6]. In a nutshell, data mining could be likened to finding a needle in a haystack. We live in a world that is full of information, and the biggest challenge is not only getting information, but searching through it to find connections and data that were not previously known.

2. Data Mining Process:

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:

2.1. 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.

2.2 Data Understanding:

- *Vikas Gupta is currently pursuing PhD degree program in Computer Science in Jammu University, India, PH-9419103174. E-mail: cvikas10@mail.com*
- *Professor Devanand is currently Dean Mathematical Sciences Jammu University, India, PH-9419103468. E-mail: devanand@jammuuniversity.in*

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.

2.3 Data Preparation:

In this stage, it collects all the different data sets and constructs the varieties of the activities basing on the initial raw data

2.4 Modeling:

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

2.5 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.

2.6 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. [7]

3. Data Mining Tools

Different types of data mining tools are available in the marketplace, each with their own strengths and weaknesses. Most data mining tools can be classified into one of three categories: traditional data mining tools, dashboards, and text-mining tools.

3.1 Traditional Data Mining Tools: Traditional data mining programs help companies establish data patterns and trends by using a number of complex algorithms and techniques. Some of these tools are installed on the desktop to monitor the data and highlight trends and others capture information residing outside a database. The majority are available in both Windows and UNIX versions, although some specialize in one operating system only. In addition, while some may concentrate on one database type, most will be able to handle any data using online analytical processing or a similar technology.

3.2 Dashboards: Installed in computers to monitor information in a database, dashboards reflect data changes and updates onscreen — often in the form of a chart or table — enabling the user to see how the business is performing. Historical data also can be referenced, enabling the user to see where things have changed (e.g., increase in sales from the same period last year). This functionality makes dashboards easy to use and particularly appealing to managers who wish to have an overview of the company's performance.

3.3 Text-mining Tools: The third type of data mining tool sometimes is called a text-mining tool because of its ability to mine data from different kinds of text — from Microsoft Word and Acrobat PDF documents to simple text files, for example. These tools scan content and convert the selected data into a format that is compatible with the tool's database, thus providing users with an easy and convenient way of accessing data without the need to open different applications. Scanned content can be unstructured (i.e., information is scattered almost randomly across the document, including e-mails, Internet pages, audio and video data) or structured (i.e., the data's form and purpose is known, such as content found in a database). Capturing these inputs can provide organizations with a wealth of information that can be mined to discover trends, concepts, and attitudes. [8]

Some of the popular domain based DM tools are as follows with brief description:

3.4 Aureka

Aureka is Thomson Reuter’s tool for analyzing and visualizing patent data and sharing it effectively inside the company. The best features of MicroPatent are its representative visualizations, basic statistics and interface.

3.5 OmniViz

OmniViz is BioWisdom’s powerful data mining tool. It is designed mainly for analyzing biological data, but is well suited to treating patent documents from other technology fields as well. The best features of OmniViz are its flexibility, efficiency, high degree of interactivity and supply of many different visualization techniques.

3.6 STN AnaVist

STN Anavist is the American Chemical Society's tool for flexible analysis of data retrieved from the STN data bank. The best features of STN AnaVist to be its representative visualizations, seamless interaction between different analyses, the ease with which various statistics could be prepared, and the user-friendly interface.

3.7 Thomson Data Analyzer VantagePoint

Thomson Data Analyzer is an analysis tool from Thomson Reuters which uses Search Technology's VantagePoint data mining software for the analysis. The best features of TDA VantagePoint were its flexibility and efficiency, as well as its macros for creating different reports and tools for comparing different groups made of the data.

3.8 TANAGRA TOOL

TANAGRA is free DATA MINING software for academic and research purposes. It proposes several data mining methods from exploratory data analysis, statistical learning, machine learning and databases area. This project is the successor of SIPINA which implements various supervised learning algorithms, especially an interactive and visual construction of decision trees. TANAGRA is more powerful, it contains some supervised learning but also other paradigms such as clustering, factorial analysis, parametric and nonparametric statistics, association rule, feature selection and construction algorithms. TANAGRA is an "open source project" as every researcher can access to the source code, and add his own algorithms, as far as he agrees and conforms to the software distribution license.

3.9 DBMINER TOOL

DBMiner, a data mining system for interactive mining of multiple-level knowledge in large relational databases, has been developed based on our years-of-research. The system implements a wide spectrum of data mining functions, including generalization, characterization, discrimination, association, classification, and prediction. By incorporation of several interesting data mining techniques, including attribute-oriented induction, progressive deepening for mining multiple-level rules, and meta-rule guided knowledge mining, the system provides a user-friendly, interactive data mining environment with good performance.

3.10 WITNESS MINER TOOL

WITNESS Miner is a graphical data mining tool comprising a collection of data structures and algorithms written specifically for the tasks required in knowledge discovery. Designed to be easy to use, it provides a visual method of constructing streams, containing data preparation and data mining tasks that form the knowledge discovery process. The key features of this tool are: decision trees, clustering, discretization, and rule induction using modern heuristic techniques, the ability to handle missing values, host of standard data processing tools, HTML output and in the case of the decision tree, XML output options, and feature subset selection.

3.11 ORANGE TOOL

Orange is a powerful free and open source component-based data mining and machine learning software suite. It contains complete set of components for data preprocessing, feature scoring and filtering, modeling, model evaluation, and exploration techniques. It is based on C++ components, that are accessed either directly (not very common), through Python scripts (easier and better), or through GUI objects called Orange Widgets. Orange is distributed free under GPL and can be downloaded from the download page.

3.12 WEKA TOOL

WEKA, formally called Waikato Environment for Knowledge Learning, is a computer program that was developed at the University of Waikato in New Zealand for the purpose of identifying information from raw data gathered from agricultural domains. WEKA supports many different standard data mining tasks such as data preprocessing, classification, clustering, regression, visualization and feature selection. The basic premise of the application is to utilize a computer application that can be trained to perform machine learning capabilities and derive useful information in the form of trends and patterns. [8][9]

Some of the Commercial data-mining software and applications are listed below as:

- **Angoss KnowledgeSTUDIO:** data mining tool provided by Angoss.
- **Clarabridge:** enterprise class text analytics solution.
- **E-NI (e-mining, e-monitor):** data mining tool based on temporal pattern.
- **IBM SPSS Modeler:** data mining software provided by IBM.
- **KXEN Modeler:** data mining tool provided by KXEN.
- **LIONsolver:** an integrated software application for data mining, business intelligence, and modeling that

implements the Learning and Intelligent OptimizatioN (LION) approach.

- **Microsoft Analysis Services:** data mining software provided by Microsoft.
- **Oracle Data Mining:** data mining software by Oracle.
- **SAS Enterprise Miner:** data mining software provided by the SAS Institute.
- **STATISTICA Data Miner:** data mining software provided by StatSoft. [10]

4. Data Mining Techniques

Various algorithms and techniques like Classification, Clustering, Regression, Artificial Intelligence, Neural Networks, Association Rules, Decision Trees, Genetic Algorithm, Nearest Neighbor method etc., are used for knowledge discovery from databases. Data mining techniques are discussed briefly below as:

4.1 Classification

Classification is the most commonly applied data mining technique, which employs a set of pre-classified examples to develop a model that can classify the population of records at large. Fraud detection and credit-risk applications are particularly well suited to this type of analysis. This approach frequently employs decision tree or neural network-based classification algorithms. The data classification process involves learning and classification. In Learning the training data are analyzed by classification algorithm. In classification test data are used to estimate the accuracy of the classification rules. If the accuracy is acceptable the rules can be applied to the new data tuples. For a fraud detection application, this would include complete records of both fraudulent and valid activities determined on a record-by-record basis. The classifier-training algorithm uses these pre-classified examples to determine the set of parameters required for proper discrimination. The algorithm then encodes these parameters into a model called a classifier. Types of classification models:

- Classification by decision tree induction
- Bayesian Classification
- Neural Networks
- Support Vector Machines (SVM)
- Classification Based on Associations

4.2 Clustering

Clustering is a data mining technique that makes meaningful or useful cluster of objects which have similar characteristics using automatic technique. The clustering technique defines the classes and puts objects in each class, while in the classification techniques, objects are assigned into predefined classes. To make the concept clearer, we can take book management in library as an example. In a library, there is a wide range of books in various topics available. The challenge is how to keep those books in a way that readers can take several books in a particular topic without hassle. By using clustering technique, we can keep books that have some kinds of similarities in one cluster or one shelf and label it with a meaningful name. If readers want to grab books in that topic, they would only have to go to that shelf instead of looking for entire library.

Types of clustering methods

- Partitioning Methods
- Hierarchical Agglomerative (divisive) methods
- Density based methods
- Grid-based methods
- Model-based methods

4.3 Prediction

The prediction, as it name implied, is one of a data mining techniques that discovers relationship between independent variables and relationship between dependent and independent variables. For instance, the prediction analysis technique can be used in sale to predict profit for the future if we consider sale is an independent variable, profit could be a dependent variable. Then based on the historical sale and profit data, we can draw a fitted regression curve that is used for profit prediction. Unfortunately, many real-world problems are not simply prediction. 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 networks too can create both classification and regression models.

Types of regression methods

- Linear Regression
- Multivariate Linear Regression
- Nonlinear Regression
- Multivariate Nonlinear Regression

4.4 Association

Association is one of the best known data mining technique. In association, a pattern is discovered based on a relationship between items in the same transaction. That's the reason why association technique is also known as relation technique. The association technique is used in market basket analysis to identify a set of products that customers frequently purchase together. Retailers are using association technique to research customer's buying habits. Based on historical sale data, retailers might find out that customers always buy crisps when they buy beers, and therefore they can put beers and crisps next to each other to save time for customer and increase sales. Association Rule algorithms need to be able to generate rules with confidence values less than one. However the number of possible Association Rules for a given dataset is generally very large and a high proportion of the rules are usually of little (if any) value.

Types of association rule

- Multilevel association rule
- Multidimensional association rule
- Quantitative association rule

4.5 Decision trees

Decision tree is one of the most used data mining techniques because its model is easy to understand for users. In decision tree technique, the root of the decision tree is a simple question or condition that has multiple answers. Each answer then leads to a set of questions or conditions that help us determine the data so that we can make the final decision based on it. Decision trees have become one of the most powerful and popular approaches in knowledge discovery and data mining, the science and technology of exploring large and complex bodies of data in order to discover useful patterns. The area is of great importance because it enables modeling and knowledge extraction from the abundance of data available. Both theoreticians and practitioners are continually seeking techniques to make the process more efficient, cost-effective and accurate. Decision trees, originally implemented in decision theory and statistics, are highly effective tools in other areas such as data mining, text mining, information extraction, machine learning, and pattern recognition. [12]

4.6 Neural networks

Neural network is a set of connected input/output units and each connection has a weight present with it. During the learning phase, network learns by adjusting weights so as to be able to predict the correct class labels of the input tuples. Neural networks have the remarkable ability to derive meaning from complicated or imprecise data and can be used to extract patterns and detect trends that are too complex to be noticed by either humans or other computer techniques. These are well suited for continuous valued inputs and outputs. For example handwritten character reorganization, for training a computer to pronounce English text and many real world business problems and have already been successfully applied in many industries. Neural networks are best at identifying patterns or trends in data and well suited for prediction or forecasting needs. [11]

Some of the other DM techniques are: Anomaly/outlier/change detection, Factor analysis, Regression analysis, Sequence mining, structured data analysis and Text mining.

5. Crucial Concepts in Data Mining

5.1 Feature Selection: One of preliminary stages in the process of data mining applicable when the data set includes more variables than could be included (or would be efficient to include) in the actual model building phase (or even in initial exploratory operations).

5.2 Bagging (Voting, Averaging) The concept of bagging (voting for classification, averaging for regression-type problems with continuous dependent variables of interest) applies to the area of predictive data mining, to combine the predicted classifications (prediction) from multiple models, or from the same type of model for different learning data. It is also used to address the inherent instability of results when applying complex models to relatively small data sets.

5.3 Boosting: The concept of boosting applies to the area of predictive data mining, to generate multiple models or classifiers (for prediction or classification), and to derive weights to combine the predictions from those models into a single prediction or predicted classification

5.4 Stacking (Stacked Generalization): The concept of stacking (short for Stacked Generalization) applies to the area of predictive data mining, to combine the predictions from multiple models. It is particularly useful when the types of models included in the project are very different.

5.5 Meta-Learning: The concept of meta-learning applies to the area of predictive data mining, to combine the predictions from multiple models. It is particularly useful when the types of models included in the project are very different. In this context, this procedure is also referred to as Stacking (Stacked Generalization).

5.6 Drill-Down Analysis: The concept of drill-down analysis applies to the area of data mining, to denote the interactive exploration of data, in particular of large databases. The process of drill-down analyses begins by considering some simple break-downs of the data by a few variables of interest (e.g., Gender, geographic region, etc.). Various statistics, tables, histograms, and other graphical summaries can be computed for each group.

5.7 Deployment: The concept of deployment in predictive data mining refers to the application of a model for prediction or classification to new data. After a satisfactory model of set of models have been identified (trained) for a particular application, one usually wants to deploy those models so that predictions or predicted classifications can quickly be obtained for new data.

5.8 Predictive Data Mining: The term Predictive Data Mining is usually applied to identify data mining projects with the goal to identify a statistical or neural network model or set of models that can be used to predict some response of interest. For example, a credit card company may want to engage in predictive data mining, to derive a (trained) model or set of models (e.g., neural networks, meta-learner) that can quickly identify transactions which have a high probability of being fraudulent.

5.9 Data Reduction: The term Data Reduction in the context of data mining is usually applied to projects where the goal is to aggregate or amalgamate the information contained in large datasets into manageable (smaller) information nuggets. Data reduction methods can include simple tabulation, aggregation (computing descriptive statistics) or more sophisticated techniques like clustering, principal components analysis, etc.

5.10 Machine Learning: Machine learning, computational learning theory, and similar terms are often used in the context of Data Mining, to denote the application of generic model-fitting or classification algorithms for predictive data mining. Unlike traditional statistical data analysis, which is usually concerned with the estimation of population parameters by statistical inference, the emphasis in data mining (and machine learning) is usually on the accuracy of prediction (predicted classification), regardless of whether or not the "models" or techniques that are used to generate the prediction is interpretable or open to simple explanation. A good example of this type of technique often applied to predictive data mining are neural networks or meta-learning techniques such as boosting, etc.[13]

6. DM Application domains

Data mining is an interdisciplinary field with wide and diverse applications there exist nontrivial gaps between data mining principles and domain-specific applications some application domains financial data analysis Retail industry Telecommunication industry Biological data analysis. There are a number of industries that are already using DM on a regular basis. Some of these organizations include retail stores, hospitals, banks, and insurance companies. Many of these organizations are combining data mining with such things as statistics, pattern recognition, and other important tools. Data mining can be used to find patterns and connections that would otherwise be difficult to find. This technology is popular with many businesses because it allows them to learn more about their customers and make smart marketing decisions. [14] 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 has capability of automatic selection of data mining technique to be applied. [8]

6.1 DM against terrorism: In the aftermath of the September 11 attacks, many countries approved new laws in the fight against terrorism. These laws allow intelligence services to gather all information deemed necessary to prevent new attacks and to swiftly identify potential terrorists. In this domain, the United States of America played a pioneer role with their Total Information Awareness program. The goal of this program was the creation of a huge central database that consolidates all the available information on the population. Similar projects were announced in Europe and the rest of the world. Although some of these programs were cancelled due to massive resistance of privacy organizations, most of these plans nevertheless seem to resurrect later under a slightly different name. For example, the "Total Information Awareness" program was conveniently relabeled to "Terrorist Information Awareness" program. [15]

6.2 Fraud or non-compliance anomaly detection: Data mining isolates the factors that lead to fraud, waste and abuse. The process of compliance monitoring for anomaly detection (CMAD) involves a primary monitoring system comparing some predetermined conditions of acceptance with the actual data or event. If any variance is detected (an anomaly) by the primary monitoring system then an exception report or alert is produced, identifying the specific variance. For instance credit card fraud detection monitoring, privacy compliance monitoring, and target auditing or investigative efforts can be done more effectively

6.3 Intrusion detection: It is a passive approach to security as it monitors information systems and raises alarms when security violations are detected. This process monitors and analyzes the events occurring in a computer system in order to detect signs of security problems. Intrusion detection systems (IDSs) may be either host based or network based, according to the kind of input information they analyze11. Over the last few years, increasing number of research projects (MADAM-ID, ADAM, Clustering project, etc.) have been applied data mining approaches (either host based or network based) to various problems (construction of operational IDSs, clustering audit log records, etc.) of intrusion detection

6.4 Lie detection (SAS Text Miner): SAS institute introduced lie-detecting software, called SAS Text Miner. Using intelligence of this tool, managers can be able to detect automatically when email or web information contains lies. Here data mining can be applied successfully to identify uncertainty in a deal or angry customers and also have many other potential applications13. Many other market mining tools are also available in real practice viz. Clementine, IBM's Intelligent Miner, SGI's MineSet, SAS's Enterprise Miner, but all pretty much the same set of tools.

6.5 Market basket analysis (MBA): Basically it applies data mining technique in understanding what items are likely to be purchased together according to association rules, primarily with the aim of identifying cross-selling opportunities. Sometimes it is also referred to as product affinity analysis. MBA gives clues as to what a customer might have bought if an idea had occurred to them. So, it can be used in deciding the location and promotion of goods by means of combo-package and also can be applied to the areas like analysis of telephone calling patterns, identification of fraudulent medical insurance claims, etc.

6.6 Aid to marketing or retailing: Data mining could help direct marketers by providing useful and accurate trends on purchasing behavior of their customers and also help them in predicting which products their customers may be interested in buying. In addition, trends explored by data mining help retail-store managers to arrange shelves, stock certain items, or provide a certain discount that will attract their customers. In fact data mining allows companies to identify their best customers, attract customers, aware customers via mail marketing, and maximize profitability by means of identifying profitable customers

6.7 Customer segmentation and targeted marketing: Data mining can be used in grouping or clustering customers based on the behaviors (like payment history, etc.), which in turn helps in customer relationship management (epiphany) and performs targeted marketing. Usually it becomes useful to define similar customers in a cluster, holding on good customers, weeding out bad customers, identify likely responders for business promotions.

6.8 Medicare and health care: Applying data mining techniques, it is possible to find relationship between diseases, effectiveness of treatments, to identify new drugs, market activities in drug delivery services, etc. However, a pharmaceutical company can analyze its recent sales to improve targeting of high-value physicians and determine which marketing activities will have the greatest impact in the next few months. The data needs to include competitor market activity as well as information about the local health care systems. Such dynamic analysis of the data warehouse allows best practices from throughout the organization to be applied in specific sale situation.

6.9 Corporate surveillance: Implies the monitoring of a person or group's behavior by a corporation, which is highly possible through data mining process. So it can be used as a form of business intelligence that enables the corporation to better tailor their products and services to be desirable by their customers. Normally the organizations that have enemies who wish to gather information about the group members or activities face the issue of infiltration usually followed a process of data mining. Thus, surveilling party may put pressure on certain members of the target organization to act as informants i.e. disclose the information they hold on the organization and its members.

6.10 Data Mining for Telecomm Industry: A rapidly expanding and highly competitive industry and a great demand for data mining Understand the business involved Identify telecommunication patterns Catch fraudulent activities Make better use of resources Improve the quality of service Multidimensional analysis of telecommunication data Intrinsically multidimensional: calling-time, duration, location of caller, location of callee, type of call, etc. Fraudulent pattern analysis and the identification of unusual patterns Identify potentially fraudulent users and their atypical usage patterns Detect attempts to gain fraudulent entry to customer accounts Discover unusual patterns which may need special attention Multidimensional association and sequential pattern analysis Find usage patterns for a set of communication services by customer group, by month, etc. Promote the sales of specific services Improve the availability of particular services in a region Use of visualization tools in telecommunication data analysis.

6.11 DNA Analysis: Examples Similarity search and comparison among DNA sequences Compare the frequently occurring patterns of each class (e.g., diseased and healthy) Identify gene sequence patterns that play roles in various diseases Association analysis: identification of co-occurring gene sequences Most diseases are not triggered by a single gene but by a combination of genes acting together Association analysis may help determine the kinds of genes that are likely to co-occur together in target samples Path analysis: linking genes to different disease development stages Different genes may become active at different stages of the disease Develop pharmaceutical interventions that target the different stages separately Visualization tools and genetic data analysis.

6.12 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. 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.

6.13 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 survey 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 behavior, 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 analyze the data across one or multiple sites. For example the search engines work on the principle of data mining.

6.14 Higher Education: An important challenge that higher education faces today is predicting paths of students and alumni. Which student will enroll in particular course programs? Who will need additional assistance in order to graduate? Meanwhile additional issues, enrollment 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.

There are many domains other than the already mentioned above where data mining plays or will play an important role. For example in astronomy, where new telescopes provide very detailed pictures of the universe, data mining techniques are used for an automatic classification of newly discovered celestial bodies. In text processing, many text mining projects emerge that create information from unstructured text documents. For example, the `Google News'-site (2005) uses clustering techniques to group news messages from several news providers according to the subject. Automatic summarizing of texts or detecting spam in e-mail messages are other applications in which text mining plays a vital role. The results of search engines depend strongly on the progress of text mining and natural language understanding. Search engines which can answer simple questions are already available and improve continuously, but it will still take many years before a computer can endure a Turing-test (Turing (1950)).

6.15 The Intelligence Agencies: 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. Now a days the intelligence agency are using the sophisticated data mining algorithms which makes it easy, to handle the very large data bases databases for organizations. The different data mining techniques are used in crime data mining. .Though the organization's have used large data bases but data mining helps us to generate the different types of information in the organization like personal details of the persons along with, vehicle details .In data mining the Clustering techniques is used (Association rule mining) for the different objects(like persons, organizations, vehicles etc.) in crime records. Not only data mining detects but also analyzes the crime data. The classification technique is also used to detect email spamming and also find person who has given the mail. String comparator is used to detect deceptive information in criminal record.

6.16 Internal Revenue Service: The data mining system implemented at the Internal Revenue Service to identify high-income individuals engaged in abusive tax shelters 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. 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%. The neural network model found better than the regression model.

6.17 Language research: Engineering much 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. 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. This method verifies automatically the text is of native quality. The results show that language verification is indeed possible.

6.18 Data Mining in Bioinformatics: Applications of data mining to bioinformatics include gene finding, protein function domain detection, function motif detection, protein function inference, disease diagnosis, disease prognosis, disease treatment optimization, protein and gene interaction network reconstruction, data cleansing, and protein sub-cellular location prediction. For example, microarray technologies are used to predict a patient's outcome. On the basis of patients' genotypic microarray data, their survival time and risk of tumor metastasis or recurrence can be estimated. Machine learning can be used for peptide identification through mass spectroscopy. Correlation among fragment ions in a tandem mass spectrum is crucial in reducing stochastic mismatches for peptide identification by database searching. An efficient scoring algorithm that considers the correlative information in a tunable and comprehensive manner is highly desirable. [7] [14] [16] [17] [18]

7. Trends in Data Mining

As different types of data are available for data mining tasks, 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:

7.1 Application exploration: Early data mining applications focused mainly on helping businesses gain a competitive edge.

7.2 Scalable and interactive data mining methods: In contrast with traditional data analysis methods, data mining must be able to handle huge amounts of data efficiently and, if possible, interactively.

7.3 Integration of data mining: with database systems, data warehouse systems, and Web database systems. Database systems, data warehouse systems, and the Web have become mainstream information processing systems.

7.4 Standardization of data mining language: A standard data mining language or other standardization efforts will facilitate the systematic development of data mining solutions, improve interoperability among multiple data mining systems and functions, and promote the education and use of data mining systems in industry and society.

7.5 Visual data mining: Visual data mining is an effective way to discover knowledge from huge amounts of data

7.6 Biological data mining: Although biological data mining can be considered under "application exploration" or "mining complex types of data," the unique combination of complexity, richness, size, and importance of biological data warrants special attention in data mining.

7.7 Data mining and software engineering: As software programs become increasingly bulky in size, sophisticated in complexity, and tend to originate from the integration of multiple components developed by different software teams, it is an increasingly challenging task to ensure software robustness and reliability.

7.8 Distributed data mining: Traditional data mining methods, designed to work at a centralized location, do not work well in many of the distributed computing environments present today (e.g., the Internet, intranets, local area networks, high-speed wireless networks, and sensor networks). Advances in distributed data mining methods are expected.

7.9 Real-time or time-critical data mining: Many applications involving stream data (such as e-commerce, Web mining, stock analysis, intrusion detection, mobile data mining, and data mining for counterterrorism) require dynamic data mining models to be built in real time. Additional development is needed in this area.

7.10 Raph mining, link analysis, and social network analysis: Graph mining, link analysis, and social network analysis are useful for capturing sequential, topological, geometric, and other relational characteristics of many scientific data sets (such as for chemical compounds and biological networks) and social data sets (such as for the analysis of hidden criminal networks)

7.11 Multi-relational and multi-database data mining: Most data mining approaches search for patterns in a single relational table or in a single database. However, most real world data and information are spread across multiple tables and databases.

7.12 New methods for mining complex types of data: mining complex types of data is an important research frontier in data mining. Although progress has been made in mining stream, time-series, sequence, graph, spatiotemporal, multimedia, and text data, there is still a huge gap between the needs for these applications and the available technology.

7.13 Privacy protection and information security in data mining: An abundance of recorded personal information available in electronic forms and on the Web, coupled with increasingly powerful data mining tools, poses a threat to our privacy and data security. [19] [20]

8. Issues / Challenges in DM

Data mining systems face a lot of problems and pitfalls. A system which is quick and correct on some small training sets, could behave completely different when applied to a larger database. A data mining system may work perfect for consistent data and perform significant worse when a little noise is added to the training set. In this section we take a look at what we mean are the most prominent problems and challenges of data mining systems today.

8.1 Noisy Data

In a large database, many of the attribute values will be inexact or incorrect. This may be due to erroneous instruments measuring some property, or human error when registering it. We will distinguish between two forms of noise in the data, both described below:

8.1.1 Corrupted Values: Sometimes some of the values in the training set are altered from what they should have been. This may result in one or more tuples in the database conflict with the rules already established. The system may then regard these extreme values as noise, and ignore them. Alternatively, one may take the values into account possibly changing correct patterns recognized. The problem is that one never knows if the extreme values are correct or not, and the challenge is how to handle ``weird'' values in the best manner.

8.1.2 Missing Attribute Values: One or more of the attribute values may be missing both for examples in the training set and for object which are to be classified. If attributes are missing in the training set, the system may either ignore this object totally, try to take it into account by for instance finding what is the missing attribute's most probable value, or use the value ``unknown'' as a separate value for the attribute. When an attribute value is missing for an object during classification, the system may check all matching rules and calculate the most probable classification.

8.2 Difficult Training Set

Sometimes the training set is not the ultimate training set due to several reasons. These are the following:

8.2.1 Not Representative Data: If the data in the training set is not representative for the objects in the domain, we have a problem. If rules for diagnosing patients are being created and only elderly people are registered in the training set, the result for diagnosing a kid based on these data probably will not be good. Even though this may have serious consequences, we would say that not representative data is mainly a problem of machine learning when the learning is based on few examples. When using large data sets, the rules created probably are representative, as long as the data being classified belongs to the same domain as those in the training set.

8.2.2 No Boundary Cases: To find the real differences between two classes, some boundary cases should be present. If a data mining system for instance is to classify animals, the property counting for a bird might be that it has wings and not that it can fly. This kind of detailed distinction will only be possible if e.g. penguins are registered.

8.2.3 Limited Information: In order to classify an object to a specific class, some condition attributes are investigated. Sometimes, two objects with the same values for condition attributes have a different classification. Then, the objects have some properties which are not among the attributes in the training set, but still make a difference. This is a problem for the system, which does not have any way of distinguish these two types of objects.

8.3 Mining methodology and user interaction issues:

These reflect the kinds of knowledge mined; the ability to mine knowledge at multiple granularities, the use of domain knowledge, ad hoc mining, and knowledge visualization.

8.3.1 Performance issues: These include efficiency, scalability, and parallelization of data mining algorithms.

- Efficiency and scalability of data mining algorithms: To effectively extract information from a huge amount of data in databases, data mining algorithms must be efficient and scalable.
- Parallel, distributed, and incremental mining algorithms: The huge size of many databases, the wide distribution of data, and the computational complexity of some data mining methods are factors motivating the development of algorithms that divide data into partitions that can be processed in parallel.

8.4 Issues relating to the diversity of database types:

Databases usually change continually. We would like rules which reflect the content of the database at all times, in order to make the best possible classification. Many existing data mining systems require that all the training examples are given at once. If something is changed at a later time, the whole learning process may have to be conducted again. An important challenge for data mining systems is to avoid this, and instead change its current rules according to updates performed.

- Handling of relational and complex types of data: Specific data mining systems should be constructed for mining specific kinds of data.

- Mining information from heterogeneous databases and global information systems: Local- and wide-area computer networks (such as the Internet) connect many sources of data, forming huge, distributed, and heterogeneous databases.

The size of databases seems to be ever increasing. Most machine learning algorithms have been created for handling only a small training set, for instance a few hundred examples. In order to use similar techniques in databases thousands of times bigger, much care must be taken. Having very much data is advantageous since they probably will show relations really existing, but the number of possible descriptions of such a dataset is enormous. Some possible ways of coping with this problem, are to design algorithms with lower complexity and to use heuristics to find the best classification rules. Simply using a faster computer is seldom a good solution. [21][22]

9. Domain Specific Research issues:

Aparna S. Varde in her article provides an overview of the dissertation problems (presented at a Ph.D. workshop in the ACM Conference on Information and Knowledge Management). In her article she has reported research issues in DM, Data bases and information retrieval. Following are some of the major research issues related to DM in different domains.

9.1 Text Mining: A text mining issue pertaining to topic models for text related applications was discussed by Ha Thuc et al. They covered text models such as aspect model or LDA. They discovered limitations of scalability issues in running the models in mining large corpora and the inability to model the important concept of relevance which prevents the models from being directly applied for text classification. To overcome these limitations, they introduced a one-scan topic model requiring only a single pass over a corpus for inference and also proposed relevance-based topic models that provided better results than state-of-the-art models.

9.2 KDD: The problem of concept search in discovering knowledge from non-English and non-European sources was discussed by Riaz. Urdu was chosen as an example language because of its unique nature, morphology and a large number of speakers. Named-entity identification was considered to be useful in determining the knowledge being sought by the user. A TREC like evaluation criteria was presented with relevance judgments, test collection and appropriate queries for knowledge discovery.

9.3 HCI: Wu reported the results of database learning in human computer interaction (HCI) errors. Earlier studies on HCI errors focused on improving the system performance. This work proposed an approach to identify the nature of HCI errors in interactions from the user perspective. It analyzed error episodes, recovery trials, and recovery actions, suggesting systematic methods for error recovery by users in different cases. Pilot studies corroborated the usefulness of the proposed approach.

9.4 Social Networks: Song et al. dealt with an NP-complete problem in the area of social network mining, proposing a 2-step method for community detection. This involved first analyzing vertex similarity of the network (a microscopic view) and putting a pair of vertices into the same community if they were similar; followed by incrementing modularity of the similarity-based communities. If the number of edges between 2 communities was greater than an expected number based on random choice, the communities were merged. They tested their method on over 20 data sets and did better than existing algorithms.

Social network mining was discussed in the work of Smith. Online communities are connecting hordes of individuals that generate rich social network data. The social capital residing in these networks is by far unknown and needs to be discovered. This work proposed to create a mathematical model of social capital to incorporate mobilization of social resources. It involved evaluating nodes based on their relationships and attributes, and also on their social resources. The result was a quantitative model for characterizing and providing decision support on how to maximize participation within social networks.

9.5 Knowledge Management: Knowledge management with respect to navigating humanities resources was the focus of Ryan Shaw. They emphasized that in the humanities, events and narratives though important, are not identified and disambiguated by most knowledge organization systems; and that the digitization of artifact collections and development of digital metadata make it imperative to address this issue. They described their research on gazetteers to depict such events; along with the relationships and best practices to use the gazetteers for improving digital resources and services. [23]

10. Conclusion:

In this paper we briefly reviewed the various data mining techniques, tools, applications issues and trends. This review would be helpful to researchers to focus on the various issues of data mining. Data mining has importance regarding finding the patterns, forecasting, and discovery of knowledge etc., in different business domains. Data mining techniques and algorithms such as classification, clustering etc., helps in finding the patterns to decide upon the future trends in businesses to grow. Data mining has wide application domain almost in every industry where the data is generated that's why data mining is considered one of the most important frontiers in database and information systems and one of the most promising interdisciplinary developments in Information Technology. 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. In future course, we will review the various classification algorithms and significance of evolutionary computing (genetic programming) approach in designing of efficient classification algorithms for data mining.

Acknowledgement: The Author is thankful to Dr Kumar Sourabh (PhD Computer Science) for inspirational and valuable comments on the draft of this paper.

References:

[1] M H Dunham, "Data Mining: Introductory and AdvancedTopics," Prentice Hall, 2002.

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

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

[4] Article: Exforsys Inc "Data Mining applications" Published on: 26th Jul 2006 Source: http://www.exforsys.com/tutorials/data-mining/data-mining-applications.html

[5] Article: Exforsys Inc" What is Data Mining" Published on: 27th Jul 2006 Source: http://www.exforsys.com/tutorials/data-mining/data-mining-overview.html

[6] Article: Exforsys Inc "Advantages of Data Mining" Published on: 26th Jul 2006 Source: http://www.exforsys.com/tutorials/data-mining/data-mining-advantages.html

[7] Neelamadhab Padhy, Dr. Pragnyaban Mishra , and Rasmita Panigrahi "The Survey of Data Mining Applications And Feature Scope" International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.2, No.3, June 2012 DOI : 10.5121/ijcseit.2012.2303 43

[8] Y. Ramamohan, K. Vasantharao, C. Kalyana Chakravarti, A.S.K.Ratnam "A Study of Data Mining Tools in Knowledge Discovery Process" International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-2, Issue-3, July 2012

[9] Laura Ruotsalainen "Data Mining Tools for Technology and Competitive Intelligence "VTT TIEDOTTEITA. RESEARCH NOTES 2451 pg 20-30. ISBN 978-951-38-7241-0 (URL: http://www.vtt.fi/publications/index.jsp) Web Source www.icsti.org/IMG/pdf/VTTDataMiningTools.pdf

[10] Article:"Commercial data-mining software and applications" Web Source: Wikipedia: http://en.wikipedia.org/wiki/Data_mining

[11] Mrs. Bharati M. Ramageri, "DATA MINING TECHNIQUES AND APPLICATIONS" Bharati M. Ramageri / Indian Journal of Computer Science and Engineering Vol. 1 No. 4 301-305

[12] Venkatadri.M Dr. Lokanatha C. Reddy "A Review on Data mining from Past to the Future" *International Journal of Computer Applications (0975 – 8887) Volume 15– No.7, February 2011*19

[13] Article: "Data Mining Techniques Crucial Concepts in Data Mining" Stat Soft Electronic Book Web Source: http://www.obgyn.cam.ac.uk/cam-only/statsbook/stdatmin.html#concepts

[14] Mr. S. P. Deshpande and Dr. V. M. Thakare "DATA MINING SYSTEM AND APPLICATIONS: A REVIEW" International Journal of Distributed and Parallel systems (IJDPS) Vol.1, No.1, September 2010 DOI: 10.5121/ijdps.2010.1103 32

[15] J. HUYSMANS, B. BAESENS, D. MARTENS,K. DENYS and J. VANTHIENEN "New Trends in Data Mining" Tijdschrift voor Economie en Management Vol. L, 4, 2005

[16] Jiban K Pal "Usefulness and applications of data mining in extracting information from different perspectives" Annals of Library and Information Studies Vol. 58, March 2011, pp. 7-16

[17] KHALID RAZA "APPLICATION OF DATA MINING IN BIOINFORMATICS" Khalid Raza / Indian Journal of Computer Science and Engineering Vol 1 No 2, 114-118

[18] Vivek Bhambri "Application of Data Mining in Banking Sector" IJCST Vol. 2, Issue 2, June 2011 ISSN : 2229-4333(Print) | ISSN: 0976-8491(Online)

[19] Article: "Practical Applications of Data Mining: Trends in Data Mining" Web Source: http://www.dataminingtools.net

[20] Krzysztof J. Cios and Lukasz A. Kurgan “Trends in Data Mining and Knowledge Discovery” Web Source: isds.bus.lsu.edu

[21] Article: Helge Grenager Solheim “Problems and Challenges in Data Mining” Web Source: http://www.pvv.ntnu.no/~hgs/project/report/node22.html

[22] Article: Morgan Kaufman, “Major issues in data mining” Source http://searchcrm.techtarget.com/tip/Major-issues-in-data-mining

[23] Aparna S. Varde “Challenging research issues in data mining, databases and information retrieval” ACM SIGKDD Explorations Newsletter Volume 11 Issue 1, June 2009 Pages 49-52 ACM New York, NY, USA.