IOSR Journal of Engineering (IOSRJEN)
e-ISSN: 2250-3021, **p**-ISSN: 2278-8719
Vol. 3, Issue 7 (July. 2013), ||V4 || PP 43-54

A Survey of Research on Data Mining

[1]Thudum Venkatesh, [2]Undrakunta Satyanarayana, [3]Kondabathula Durga Charan
M. Tech in Computational Engineering at RGUKT, Nuzvid.

***Abstract:* -** With an enormous amount of data stored in databases and data warehouses, it is increasinglyimportant to develop powerful tools for analysis of such data and mining interesting knowledgefrom it. Data mining is a process of inferring knowledge from such huge data. The mainproblem related to the retrieval of information from the World Wide Web is the enormousnumber of unstructured documents and resources, i.e., the difficulty of locating and trackingappropriate sources. In this survey of the research in the area of data mining andsuggest data mining categories and techniques. Furthermore, a data miningenvironment generator that allows naive users to generate a data mining environment specific to a given domain by providing a set of specifications.

***Keywords:* -** Association, *Clustering, Data Mining, Discrimination, Prediction*.

I. INTRODUCTION

Data mining is emerging as one of the key features of many homeland security initiatives. Often used as a means for detecting fraud, assessing risk, and productretailing, data mining involves the use of data analysis tools to discover previouslyunknown, valid patterns and relationships in large data sets. In the context ofhomeland security, data mining is often viewed as a potential means to identifyterrorist activities, such as money transfers and communications, and to identify andtrack individual terrorists themselves, such as through travel and immigration records. While data mining represents a significant advance in the type of analytical toolscurrently available, there are limitations to its capability. One limitation is thatalthough data mining can help reveal patterns and relationships, it does not tell theuser the value or significance of these patterns. These types of determinations mustbe made by the user. A second limitation is that while data mining can identifyconnections between behaviors and/or variables, it does not necessarily identify acausal relationship. To be successful, data mining still requires skilled technical andanalytical specialists who can structure the analysis and interpret the output that is created. Data mining is becoming increasingly common in both the private and publicsectors. Industries such as banking, insurance, medicine, and retailing commonly usedata mining to reduce costs, enhance research, and increase sales. In the publicsector, data mining applications initially were used as a means to detect fraud andwaste, but have grown to also be used for purposes such as measuring and improvingprogram performance. However, some of the homeland security data miningapplications represent a significant expansion in the quantity and scope of data to beanalyzed. Two efforts that have attracted a higher level of congressional interestinclude the Terrorism Information Awareness (TIA) project (now-discontinued) andthe Computer-Assisted Passenger Prescreening System II (CAPPS II) project (now-canceled and replaced by Secure Flight). As with other aspects of data mining, while technological capabilities areimportant, there are other implementation and oversight issues that can influence thesuccess of a project's outcome. One issue is data quality, which refers to theaccuracy and completeness of the data being analyzed. A second issue is theinteroperability of the data mining software and databases being used by differentagencies. A third issue is mission creep, or the use of data for purposes other thanfor which the data were originally collected. A fourth issue is privacy. Questionsthat may be considered include the degree to which government agencies should useand mix commercial data with government data, whether data sources are being usedfor purposes other than those for which they were originally designed, and possibleapplication of the Privacy Act to these initiatives.

II. EVOLUTION OF DATA MINING

Data mining is a tool that can extract predictive information from large quantities of data, and isdata driven. It uses mathematical and statistical calculations to uncover trends and correlationsamong the large quantities of data stored in a database. It is a blend of artificial intelligencetechnology, statistics, data warehousing, and machine learning. Data mining started with statistics. Statistical functions such as standard deviation, regressionanalysis, and variance are all valuable tools that allow people to study the reliability andrelationships between data. Much of what data mining does is rooted in statistics, making it oneof the cornerstones of data mining technology. In the 1970's data was stored using large mainframe systems and COBOL programmingtechniques. These simplistic beginnings gave way to very large databases called "datawarehouses", which store data in one standard format. The dictionary definition of a datawarehouse is "a generic term for storing, retrieving, and managing large amounts of data."These data warehouses "can now store and query terabytes and megabytes ofdata in sophisticated database management systems." These data stores are anessential part of data mining, because a cornerstone of the technology is that it needs very largeamounts of organized data to manipulate. In addition to basic statistics and large data warehouses, a major part of data mining technologyis artificial intelligence (AI). Artificial intelligence started in the 1980's with a set of algorithms that was designed to teach a computer how to "learn" by it. As they developed, thesealgorithms became valuable data manipulation tools and were applied to large sets of data.Instead of entering a set of pre-defined hypothesis; the data mining software, combined with AItechnology was able to generate its own relationships between the data. It was even able to analyze data and discover correlations between the data on its own, and develop models to helpthe developers interpret the relationships that were found.

Evolutionary Step	Business Question	Enabling Technologies	Product Providers	Characteristics
Data Collection (1960s)	"What was my total revenue in the last five years?"	Computers, tapes, disks	IBM, CDC	Retrospective, static data delivery
Data Access (1980s)	"What were unit sales in New England last March?"	Relational databases (RDBMS), Structured Query Language (SQL), ODBC	Oracle, Sybase, Informix, IBM, Microsoft	Retrospective, dynamic data delivery at record level
Data Warehousing & Decision Support (1990s)	"What were unit sales in New England last March? Drill down to Boston."	On-line analytic processing (OLAP), multidimensional databases, data warehouses	Pilot, Comshare, Arbor, Cognos, Microstrategy	Retrospective, dynamic data delivery at multiple levels
Data Mining (Emerging Today)	"What's likely to happen to Boston unit sales next month? Why?"	Advanced algorithms, multiprocessor computers, massive databases	Pilot, Lockheed, IBM, SGI, numerous startups (nascent industry)	Prospective, proactive information delivery

Figure 1. Data Mining Evolutionary Chart

AI gave way to machine learning. Machine learning is defined as "the ability of a machine toimprove its performance based on previous results." (dictionary.com) Machine learning is thenext step in artificial intelligence technology because it blends trial and error learning by thesystem with statistical analysis. This lets the software learn on its own and allows it to makedecisions regarding the data it is trying to analyze.

Later in the 1990's data mining became wildly popular. Many companies began to use the datamining technology and found that it was much easier than having actual people work with suchlarge amounts of data and attributes. This technology allows the systems to "think" forthemselves and run analysis that would provide trend and correlation information for the data inthe tables. In 2001, the use of data warehouses grew by over a third to 77%. Data mining is a very important tool for business and as time goes on, business is becomingmore and more competitive and everyone is scrambling for a competitive edge. Businesses need to gain a competitive edge, and can get it from the increased awareness they canget from data mining software that is available on the market right now.

III. KNOWLEDGE DISCOVERY PROCESS

There is huge gap from the stored data to the knowledge that could be constructed from the data, that's where data mining comes into picture. Knowledge Discovery in Database (KDD) refers to the overall process of discovering useful patterns from the data. Data mining is a major step in KDD process and at times synonym to KDD.

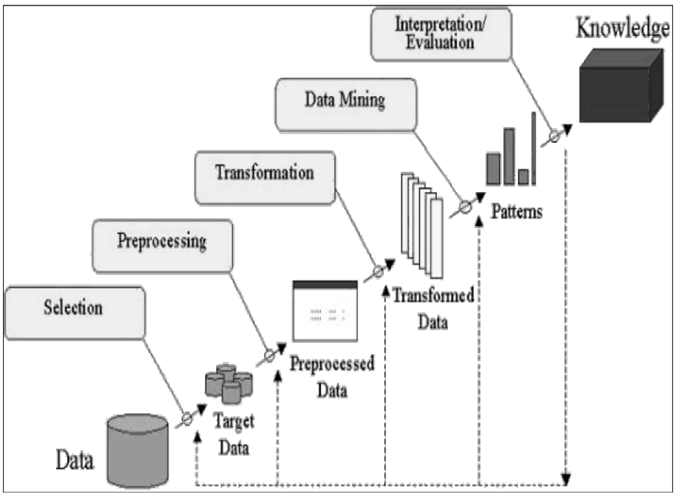

Figure 2. KDD Process

Knowledge Discovery Process Steps
The process of knowledge discovery using data mining can be divided into defined steps presented in above fig.

Selection
This step involves identification or extraction of relevant data for analysis.

Preprocessing
This involves preparing/cleaning the data set by resolving problems like missing data, skewed data, irrelevant fields, removal of outlying points, format conversion etc. This step might consist of following operations that need to be performed before a data mining technique is applied.

Data Cleaning
It consist of some basic operations like normalization, noise removal and handling of missing or inconsistent data. Data from real world sources are often erroneous, incomplete and inconsistent, may be due to operational error or implementation flaws.

Data integration
This includes integrating multiple, heterogeneous datasets generated from different sources.

Transformation
Consolidation of data into the form appropriate for mining. Eg. Performing aggregation or summary of data.

Reduction
This includes finding useful features to represent the data and using dimensionality reduction, feature discretization, and feature extraction/transformation methods.

Data Mining
This step involves application of knowledge discovery algorithms to the cleaned, transormed data in order to extract meaningful patterns from the data.

Pattern evaluation
This step involves evaluation of patterns for interestingness. One can evaluate the mined patterns automatically or semi automatically to identify the truly interesting or useful patterns for the user.

Knowledge presentation and Interpretation
This involves representation of discovered knowledge in proper format.

IV. DATA MINING DEFINITION

Data mining is defined as a non-trivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data. The term process implies that data mining consists of many steps, non-trivial means process is not straight forward and some search or inference is involved. The term pattern is an expression in some language describing a subset of data, finding structures from data, or , in general making any high level description of a set of data. Pattern should be novel and potentially useful, that is, it should lead to some benefits to the user or task. Ultimately pattern should be understandable, if not immediately then at a later stage after some post processing. Data mining is a highly inter disciplinary area spanning a range of disciplines; statistics, machine learning, database, pattern recognition and other areas.

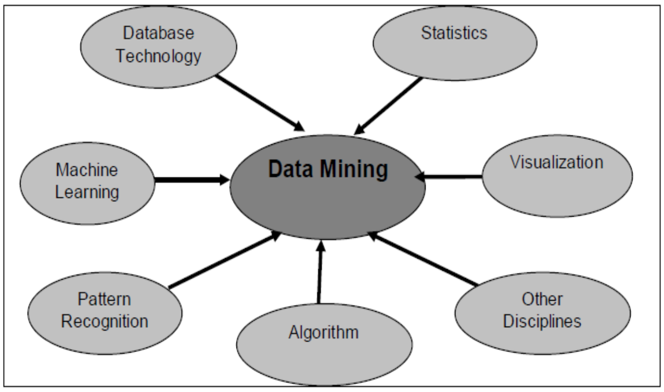

Figure 3. Data mining as a confluence of multiple disciplines

All of these fields are concerned with certain aspects of data analysis, so they have much in common but each has its own distinct flavor. Thus methods from these disciplines are welcome in data mining in their capacity to do the job. However the focus is different in various disciplines. In machine learning and statistics the stress is on the consistency of the algorithm, however in data mining it is the consistency of pattern that matters the most.

V. WHAT KIND OF DATA CAN BE MINED

Data mining should be applicable to any kind of information repository. However, algorithms and approaches may differ when applied to different types of data. Indeed, the challenges presented by different types of data vary significantly. Data mining is being put into use and studied for databases, including relational databases, object-relational databases and object-oriented databases, data warehouses, transactional databases, unstructured and semi-structured repositories such as the World Wide Web, advanced databases such as spatial databases, multimedia databases, time-series databases and textual databases, and even flat files.

1.1 Flat files

Flat files are actually the most common data source for data mining algorithms, especially at the research level. Flat files are simple data files in text or binary format with a structure known by the data mining algorithm to be applied. The data in these files can be transactions, time-series data, scientific measurements, etc.

1.2 Relational Databases

Briefly, a relational database consists of a set of tables containing either values of entity attributes, or values of attributes from entity relationships. Tables have columns and rows, where columns represent attributes and rows represent tuples. A tuple in a relational table corresponds to either an object or a relationship between objects and is identified by a set of attribute values representing a unique key.

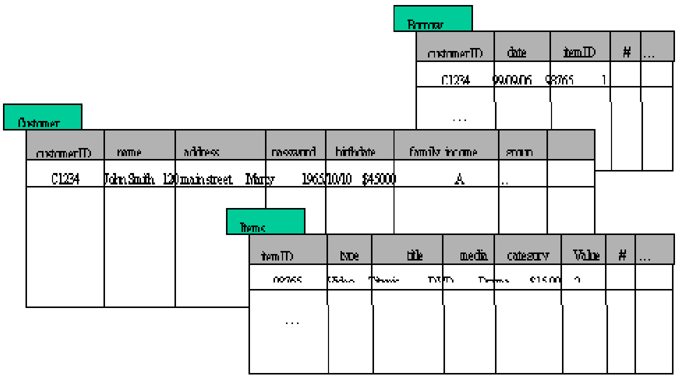

Figure 4. Relational Table

The most commonly used query language for relational database is SQL, which allows retrieval and manipulation of the data stored in the tables, as well as the calculation of aggregate functions such as average, sum, min, max and count. For instance, an SQL query to select the videos grouped by category would be:

SELECT count (*) FROM Items WHERE type=video GROUP BY category.

1.3 Data Warehouses

A data warehouse as a storehouse is a repository of data collected from multiple data sources (often heterogeneous) and is intended to be used as a whole under the same unified schema. A data warehouse gives the option to analyze data from different sources under the same roof.

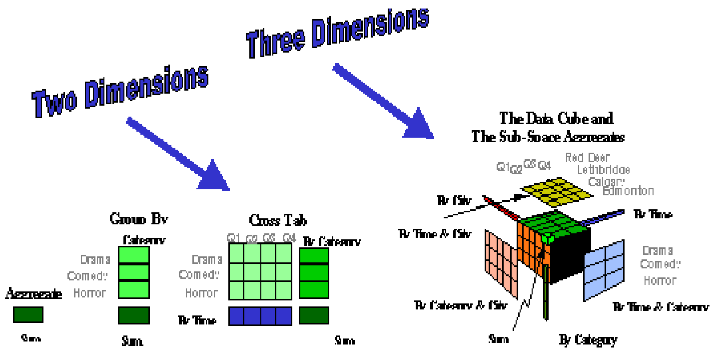

Figure 5. Multidimensional view

1.4 Transaction Databases

A transaction database is a set of records representing transactions, each with a time stamp, an identifier and a set of items. Associated with the transaction files could also be descriptive data for the items. For example, in the case of the video store, the rentals table such as shown in Figure 1.5, represents the transaction database. Each record is a rental contract with a customer identifier, a date, and the list of items rented (i.e. video tapes, games, VCR, etc.). Since relational databases do not allow nested tables (i.e. a set as attribute value), transactions are usually stored in flat files or stored in two normalized transaction tables, one for the transactions and one for the transaction items. One typical data mining analysis on such data is the so-called market basket analysis or association rules in which associations between items occurring together or in sequence are studied.

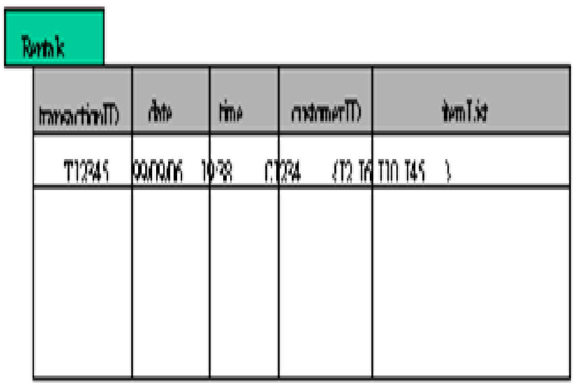

Figure 6. Transaction Data

1.5 Multimedia Databases

Multimedia databases include video, images, and audio and text media. They can be stored on extended object-relational or object-oriented databases, or simply on a file system. Multimedia is characterized by its high dimensionality, which makes data mining even more challenging. Data mining from multimedia repositories may require computer vision, computer graphics, image interpretation, and natural language processing methodologies.

1.6 Spatial Databases

Spatial databases are databases that, in addition to usual data, store geographical information like maps, and global or regional positioning. Such spatial databases present new challenges to data mining algorithms.

1.7 Time-Series Databases

Time-series databases contain time related data such stock market data or logged activities. These databases usually have a continuous flow of new data coming in, which sometimes causes the need for a challenging real time analysis. Data mining in such databases commonly includes the study of trends and correlations between evolutions of different variables, as well as the prediction of trends and movements of the variables in time.

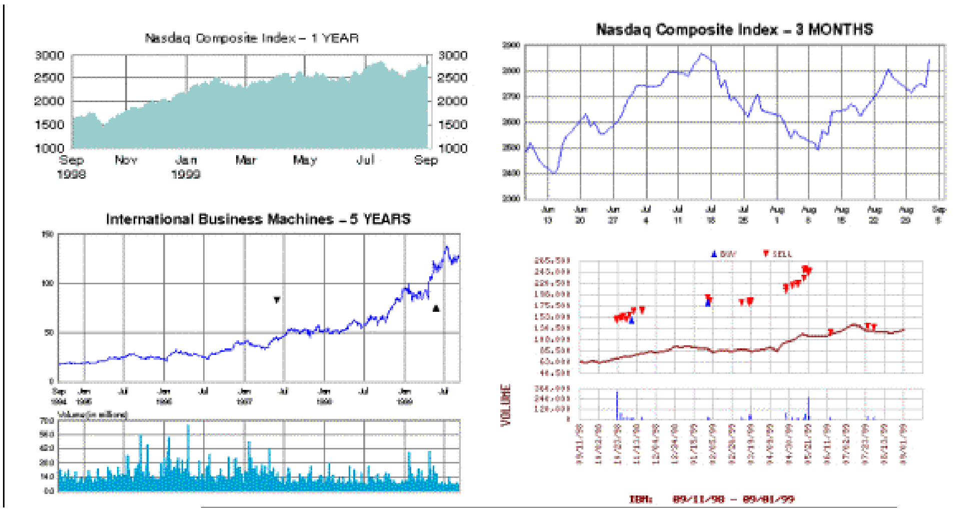

Figure 7. Time series data

1.8 World Wide Web

The World Wide Web is the most heterogeneous and dynamic repository available. A very large number of authors and publishers are continuously contributing to its growth and metamorphosis, and a massive number of users are accessing its resources daily. Data in the World Wide Web is organized in inter-connected documents. These documents can be text, audio, video, raw data, and even applications. Conceptually, the World Wide Web is comprised of three major components: The content of the Web, which encompasses documents available; the structure of the Web, which covers the hyperlinks and the relationships between documents; and the usage of the web, describing how and when the resources are accessed. A fourth dimension can be added relating the dynamic nature or evolution of the documents. Data mining in the World Wide Web, or web mining, tries to address all these issues and is often divided into web content mining, web structure mining and web usage mining.

VI. DATA MINING FUNCTIONALITIES

The kinds of patterns that can be discovered depend upon the data mining tasks employed. By and large, there are two types of data mining tasks: descriptive data mining tasks that describe the general properties of the existing data, and predictive data mining tasks that attempt to do predictions based on inference on available data. The data mining functionalities and the variety of knowledge they discover are briefly presented below.

1.9 Characterization

Data characterization is a summarization of general features of objects in a target class, and produces what is called characteristic rules. The data relevant to a user-specified class are normally retrieved by a database query and run through a summarization module to extract the essence of the data at different levels of abstractions.

1.10 Discrimination

Data discrimination produces what are called discriminant rules and is basically the comparison of the general features of objects between two classes referred to as the targetclass and the contrasting class.

1.11 Association analysis

Association analysis is the discovery of what are commonly called association rules. It studies the frequency of items occurring together in transactional databases, and based on a threshold called support, identifies the frequent item sets. Another threshold, confidence, which is the conditional probability than an item appears in a transaction when another item appears, is used to pinpoint association rules. Association analysis is commonly used for market basket analysis.

1.12 Classification

Classification analysis is the organization of data in given classes. Also known as supervised classification, the classification uses given class labels to order the objects in the data collection. Classification approaches normally use a training set where all objects are already associated with known class labels. The classification algorithm learns from the training set and builds a model. The model is used to classify new objects.

1.13 Prediction

Prediction has attracted considerable attention given the potential implications of successful forecasting in a business context. There are two major types of predictions: one can either try to predict some unavailable data values or pending trends, or predict a class label for some data. The latter is tied to classification. Once a classification model is built based on a training set, the class label of an object can be foreseen based on the attribute values of the object and the attribute values of the classes. Prediction is however more often referred to the forecast of missing numerical values, or increase/ decrease trends in time related data.

1.14 Clustering

Similar to classification, clustering is the organization of data in classes. However, unlike classification, in clustering, class labels are unknown and it is up to the clustering algorithm to discover acceptable classes. Clustering is also called unsupervised classification, because the classification is not dictated by given class labels. There are many clustering approaches all based on the principle of maximizing the similarity between objects in a same class (intra-class similarity) and minimizing the similarity between objects of different classes (inter-class similarity).

1.15 Outlier analysis

Outliers are data elements that cannot be grouped in a given class or cluster. Also known as exceptions or surprises, they are often very important to identify. While outliers can be considered noise and discarded in some applications, they can reveal important knowledge in other domains, and thus can be very significant and their analysis valuable.

1.16 Evolution and deviation analysis

Evolution and deviation analysis pertain to the study of time related data that changes in time. Evolution analysis models evolutionary trends in data, which consent to characterizing, comparing, classifying or clustering of time related data. Deviation analysis, on the other hand, considers differences between measured values and expected values, and attempts to find the cause of the deviations from the anticipated values.

VII. ARCHITECTURE OF DATA MINING SYSTEM

The architecture of a typical data mining system may have the following major components. They are Database, Data Warehouse, World Wide Web, etc.

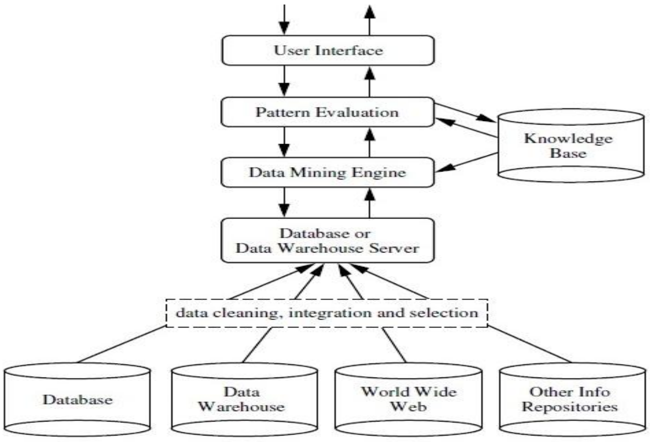

Figure 8. Data Mining Architecture

Database or Data Warehouse server

Database or data warehouse server is responsible for fetching the relevant data, based on the user's data mining request.

Knowledge base

This is the domain knowledge that is used to guide the search or evaluate the interestingness of resulting patterns. Such knowledge can include concept hierarchies used to organize attributes or attribute values into different levels of abstraction. Knowledge such as user beliefs, which can be used to assess a pattern's interestingness based on its unexpectedness, may also be included. Other examples of domain knowledge interestingness constraints or thresholds, and metadata (e.g., describing data from multiple heterogeneous sources).

Data Mining Engine

This is essential to the data mining system and ideally consists of a set of functional modules for tasks such as characterization, association and correlation analysis, classification, prediction, cluster analysis, outlier analysis, and evolution analysis.

Pattern evaluation module

This component typically employs interestingness measures and interacts with the data mining modules so as to focus the search toward interesting patterns. It may use interestingness thresholds to filter out discovered patterns. Alternatively, the pattern evaluation module may be integrated with the mining module, depending on the implementation of the data mining method used.For efficient data mining, it is highly recommended to push the evaluation of pattern interestingness as deep as possible into the mining process so as to confine the search to only the interesting patterns.

User interface

This module communicates between users and the data mining system, allowing the user to interact with the system by specifying a data mining query or task, providing information to help focus the search, and performing exploratory data mining based on the intermediate data mining results.Also, it allows the user to browse database and data warehouse schemas or data structures, evaluate mined patterns, and visualize the patterns in different form.

VIII. DATA MINING PROCESS

The data mining process must be reliable and repeatable by business people with little knowledge or no data mining background. In 1990, a cross-industry standard process for data mining (CRISP-DM) first published after going through a lot of workshops, and contributions from over 300 organizations.

1.17 The Cross-Industry Standard Process for Data Mining (CRISP-DM)

Cross-Industry Standard Process for Data Mining (CRISP-DM) consists of six phases intended as a cyclical process as the following Fig 9.

Business understanding

In the business understanding phase, first it is a must to understand business objectives clearly and make sure to find out what the client really want to achieve. Next, we have to assess the current situation by finding about the resources, assumptions, constraints and other important factors which should be considered. Then from the business objectives and current situations, we need to create data mining goals to achieve the business objective and within the current situation. Finally a good data mining plan has to be established to achieve both business and data mining goals. The plan should be as details as possible that have step-by-step to perform during the project including the initial selection of data mining techniques and tools.

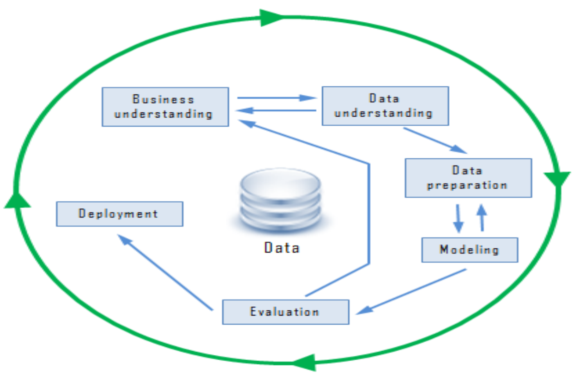

Figure 9. Cross-Industry Standard Process for Data Mining (CRISP-DM)

Data understanding

First, the data understanding phase starts with initial data collection that collects data from available sources to get familiar with data. Some important activities must be carried including data load and data integration in order to make the data collection successfully. Next, the "gross" or "surface" properties of acquired data need to be examined carefully and reported. Then, the data need to be explored by tackling the data mining questions, which can be addressed using querying, reporting and visualization. Finally, the data quality must be examined by answering some important questions such as "Is the acquired data complete?", "Is there any missing values in the acquired data?"

Data preparation

The data preparation normally consumes about 90% of the time. The outcome of the data preparation phase is the final data set. Once data sources available are identified, they need to be selected, cleaned, constructed and formatted into the desired form. The data exploration task at a greater depth may be carried during this phase to notice the patterns based on business understanding.

Modeling

First, modeling techniques have to be selected to be used for the prepared dataset. Next, the test scenario must be generated to validate the models' quality and validity. Then, one or more models are created by running the modeling tool on the prepared dataset. Last but not least, models need to be assessed carefully involving stakeholders to make sure that created models are meet business initiatives.

Evaluation

In the evaluation phase, the model results must be evaluated in the context of business objectives in the first phase. In this phase, new business requirements may be raised due to new patterns has been discovered in the model results or from other factors. Gaining business understanding is an iterative process in data mining. The go or no-go decision must be made in this step to move to the deployment phase.

Deployment

The knowledge or information that gain through data mining process needs to be presented in such a way that stakeholders can use it when they want it. Based on the business requirements, the deployment phase could be as simple as creating a report or as complex as a repeatable data mining process across the organization. In this phase, the deployment, maintained and monitoring plans have to be created for deployment and future supports. From project point of view, the final report of the project need to summary the project experiences and review the project to see what need to improved created learned lesson.

IX. DATA MINING APPLICATIONS

As data mining matures, new and increasingly innovativeapplications for it emerge. Although a wide variety of datamining scenarios can be described. For the purpose of thispaper the applications of data mining are divided in the following categories.

1.18 Healthcare

The past decade has seen an explosive growth in biomedicalresearch, ranging from the development of newpharmaceuticals and in cancer therapies to the identificationand study of human genome by discovering large scalesequencing patterns and gene functions. Recent research inDNA analysis has led to the discovery of genetic causes formany diseases and disabilities as well as approaches fordisease diagnosis, prevention and treatment.

1.19 Finance

Most banks and financial institutions offer a wide variety ofbanking services (such as checking, saving, and businessand individual customer transactions), credit (such asbusiness, mortgage, and automobile loans), and investmentservices (such as mutual funds). Some also offer insuranceservices and stock services. Financial data collected in thebanking and financial industry is often relatively complete, reliable and high quality, which facilitates systematic dataanalysis and data mining. For example it can also help infraud detection by detecting a group of people who stageaccidents to collect on insurance money.

1.20 Real Industry

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

1.21 Telecommunication

The telecommunication industry has quickly evolved fromoffering local and long distance telephone services toprovide many other comprehensive communication servicesincluding 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 andcomputing are underway. Moreover, with the deregulationof the telecommunication industry in many countries andthe development of new computer and communicationtechnologies, the telecommunication market is rapidlyexpanding and highly competitive. This creates a greatdemand from data mining in order to help understand business involved, identify telecommunication patterns,catch fraudulent activities, make better use of resources, andimprove the quality of service.

1.22 Text Mining and Web Mining

Text mining is the process of searching large volumes ofdocuments from certain keywords or key phrases. Bysearching literally thousands of documents variousrelationships between the documents can be established. Using text mining however; we can easily derive certainpatterns in the comments that may help identify a commonset of customer perceptions not captured by the other surveyquestions.

An extension of text mining is web mining. Web mining is an exciting new field that integrates data and text miningwithin a website. It enhances the web site with intelligentbehavior, such as suggesting related links or recommendingnew products to the consumer. Web mining is especiallyexciting because it enables tasks that were previouslydifficult to implement. They can be configured to monitorand gather data from a wide variety of locations and cananalyze the data across one or multiple sites. For examplethe search engines work on the principle of data mining.

1.23 Higher Education

An important challenge that higher education faces today ispredicting paths of students and alumni. Which student willenroll in particular course programs? Who will needadditional assistance in order to graduate? Meanwhileadditional issues, enrollment management and time-to-degree, continue to exert pressure on colleges to search fornew and faster solutions. Institutions can better addressthese students and alumni through the analysis andpresentation of data. Data mining has quickly emerged as ahighly desirable tool for using current reporting capabilitiesto uncover and understand hidden patterns in vast databases.

X. TRENDS IN DATA MINING

Following are the some of the trends in data mining.

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. Because the amount of data being collected continues to increase rapidly, scalable algorithms for individual and integrated data mining functions become essential. One important direction toward improving the overall efficiency of the mining process while increasing user interaction is constraint-based mining. This provides users with added control by allowing the specification and use of constraints to guide data mining systems in their search for interesting patterns.

Integration of data mining with database systems, data warehouse systems, and Web database systems

Database systems, data warehouse systems, and the Web havebecome mainstream information processing systems. It is important to ensure thatdata mining serves as an essential data analysis component that can be smoothlyintegrated into such an information processing environment. As discussed earlier,a data mining system should be tightly coupled with database and data warehousesystems. Transaction management, query processing, on-line analytical processing,and on-line analytical mining should be integrated into one unified framework. Thiswill ensure data availability, data mining portability, scalability, high performance,and an integrated information processing environment for multidimensional dataanalysis and exploration.

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. Recent efforts in this direction include Microsoft's OLE DB for Data Mining (the appendix of this book provides an introduction), PMML, and CRISP-DM.

Visual data mining
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 facilitate the promotion and use of data mining as a tool for data analysis.

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. Mining DNA and protein sequences, mining high-dimensional microarray data, biological pathway and network analysis, link analysis across heterogeneous biological data, and information integration of biological data by data mining are interesting topics for biological data mining research.

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. The analysis of the executions of a buggy software program is essentially a data mining process tracing the data generated during program executions may disclose important patterns and outliers that may lead to the eventual automated discovery of software bugs. We expect that the further development of data mining methodologies for software debugging will enhance software robustness and bring new vigor to software engineering.

Web mining
Given the huge amount of information available on the Web and the increasingly important role that the Web plays in today's society, Web content mining, Weblog mining, and data mining services on the Internet will become one of the most important and flourishing subfields in data mining.

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.

Real-time or time-critical data mining: Many applications involving stream data require dynamic data mining models to be built in real time. Additional development is needed in this area.

Graph mining, link analysis, and social network analysis: Graph mining, link analysis, and social network analysis are useful for capturing sequential, topological, geo-metric, and other relational characteristics of many scientific data sets and social data sets. Such modeling is also useful for analyzing links in Web structure mining. The development of efficient graph and linkage models is a challenge for data mining.

XI. TECHNIQUES IN DATA MINING

The most commonly used techniques include artificial neural networks, decision trees, and the nearest-neighbor method.

1.24 Artificial neural networks

Artificial neural networks are non-linear, predictive models that learn through training. Although they are powerful predictive modeling techniques, some of the power comes at the expense of ease of use and deployment. One area where auditors can easily use them is when reviewing records to identify fraud and fraud-like actions. Because of their complexity, they are better employed in situations where they can be used and reused, such as reviewing credit card transactions every month to check for anomalies.

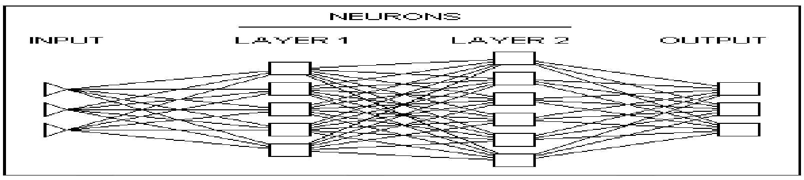

Figure 10. Neural Network Diagram

1.25 Decision Trees

Decision trees are tree-shaped structures that represent decision sets. These decisions generate rules, which then are used to classify data. Decision trees are the favored technique for building understandable models. Auditors can use them to assess, for example, whether the organization is using an appropriate cost-effective marketing strategy that is based on the assigned value of the customer, such as profit.

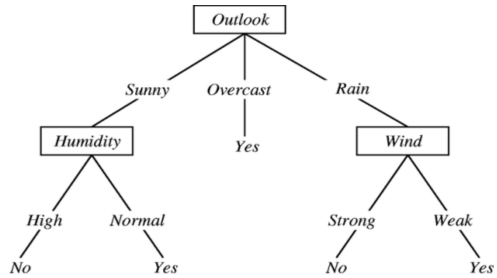

Figure 11. Decision tree example diagram

1.26 Nearest neighbor method

The nearest-neighbor method classifies dataset records based on similar data in a historical dataset. Auditors can use this approach to define a document that is interesting to them and ask the system to search for similar items. Each of these approaches brings different advantages and disadvantages that need to be considered prior to their use. Neural networks, which are difficult to implement, require all input and resultant output to be expressed numerically, thus needing some sort of interpretation depending on the nature of the data-mining exercise. The decision tree technique is the most commonly used methodology, because it is simple and straightforward to implement. Finally, the nearest-neighbor method relies more on linking similar items and, therefore, works better for extrapolation rather than predictive enquiries.

XII. CONCLUSION

In this paper I have given a brief overview of Data Mining. It is the information extraction activity from large amount of data bases. It is very useful in now a day because in olden days finding the information is time taking process but now a day we can extract the useful data within the seconds.

REFERENCES

[1]. Data Mining Concepts and Techniques by Jiawei Han and MichellineKamber Second Edition.
[2]. Introduction to Data Mining by PANG-NING TAN, MICHEL STEINBACH and VIPIN KUMAR.
[3]. Data Mining by Doug Alexamder. Data Mining Concepts, Models and Techniques by Florin Gorunescu, Volume 12.
[4]. B. Brumen, I. Golob, T. Welzer, I. Rozman, M. Druzovec, and H. Jaakkola. An Algorithm for Protecting Knowledge Discovery Data
[5]. Jiawei Han and MichelineKamber. Data Mining: Concepts and Techniques. Morgan Kaufmann. April 2000.
[6]. Michael Berry and Gordon Linoff, "Data Mining Techniques (For Marketing, Sales, and Customer Support), John Wiley & Sons, 1997.
[7]. K. Cios, W. Pedrycz, R. Swiniarski, L. Kurgan, Data Mining: A Knowledge Discovery Approach
[8]. Data Mining by Pieter Adriaans, Pearson publication.
[9]. Data Mining practical Machine Learning Tools and Techniques by Mark Hall, Ian Witten and Eibe Frank.
[10]. Michael Berry & Gordon Linoff, Mastering Data Mining, John Wiley & Sons, 2000.
[11]. Data Mining: Concepts and Techniques (The Morgan Kaufmann Series in Data Management Systems), Jiawei Han, Morgan Kaufmann, 2011.
[12]. Discovering knowledge in data-An introduction to Data Mining by DANIAL T.LAROSE.
[13]. Daniel Larose, Data Mining Methods and Models, Wiley-Interscience, Hoboken, NJ (to appear 2005).