International Journal of Innovative Research in Science and Engineering
Vol. No.3, Issue 01, January 2017
www.ijirse.com
ISSN (O) 2454-9665
ISSN (P) 2454-0663

PATTERN DISCOVERY: INSIGHT TO FORESIGHT

Dr. Kirti Singh Chundawat

Principal & HOD (CS), SDM PG Girls College,
Mahila Ashram Group of Institution, Bhilwara (Raj.)

ABSTRACT

The present paper focuses on the fast growing and indispensible technique which is known as Pattern Discovery. In fact, Pattern Discovery is a step in the process of Knowledge Discovery in Databases (KDD).Pattern Discovery provides the base for decision making and policy making for the corporate world. The process of knowledge discovery in data bases (KDD) includes a number of steps such as data cleaning, data integration, data selection, data transformation, pattern discovery, pattern evaluation and knowledge representation. Besides, Patterns are the methods of evaluation and interpretation which results in form of knowledge, decision factors or say managerial decision factors. The present paper covers the concept of pattern, pattern discovery, background of pattern discovery, significance of pattern discovery, pattern discovery with data mining techniques, characteristics of a good pattern discovery technique, classification of pattern discovery techniques, supervised and unsupervised pattern discovery techniques, recent trends and challenges in the field of pattern discovery.

Keywords: KDD,Pattern Discovery, Data Mining, ,Supervised Techniques, Unsupervised techniques, Clustering, Association, Classification, Regression, ,Partitioning , Hierarchical Clustering.

I. INTRODUCTION

A pattern is a recognizable regularity in the way in which something happens or is done. Pattern means a series of actions or events that together show how things normally happen. The elements of a pattern repeat in a predictable manner. Patterns should be original, unique, useful, innovative, and easy to understand and research oriented. Patterns are generated only when the elements have some common characteristics. A pattern represents some subgroup of the data and patterns are selected on the basis of the

support of the subgroup. Pattern Discovery in a data is the process of identifying patterns it contains and the relationships that exist in the data.

Pattern Discovery can be defined as a process of finding some very useful information from large volume of data i.e. database, which can then be turned into usable, predictive information, the information into action, and the action into profit. Pattern discovery is the study of methods and algorithms for categorizing data objects i.e. to find frequently occurring events in large databases. The general goal of pattern discovery may be described as finding frequent, unknown patterns or associations among the objects stored in a given data repository , both with or without the assumption of some domain-specific prior knowledge. Pattern Discovery can also be seen

ISSN (O) 2454-9665
ISSN (P) 2454-0663

as form of data reduction or data compression**.** While classical pattern discovery techniques used techniques of statistics and decision theory to discover patterns. The machine learning pattern discovery techniques are used to design practical systems. The Data Mining techniques are also used for pattern discovery. The pattern discovery stage is the centre of the entire data mining process. It is the stage where the hidden patterns and trends in the data are actually uncovered. Pattern Discovery is more popularly known as Data Mining. There are several approaches to the pattern discovery stage. These include association, classification, clustering, regression, time-series analysis and visualization. Each of the above approaches can be implemented through one of several competing methodologies, such as statistical data analysis, machine learning, neural networks and pattern recognition. The Pattern Discovery incorporates the methods of various branches like, statistics, data base technologies, artificial intelligence and visualization. It is because of the use of methodologies from several disciplines that Pattern Discovery is often viewed as a multidisciplinary field.

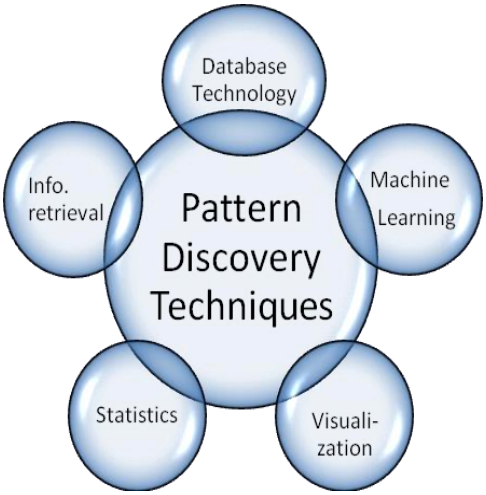

Figure 1 Pattern Discovery-A Multidisciplinary Field

II. PATTERN DISCOVERY: THEN & NOW

The need for discovering knowledge existed in the society for centuries. In 17 th century Bayes' theorem was used as a method for identifying patterns from the data. With the growth in the field of statistics, the regression analysis became a new approach for pattern identification in 18th century. With the extensive growth of computer technology the data collection, storage and manipulation ability has drastically increased across various fields. The conventional file processing system has been replaced with sophisticated and powerful database systems with the advancement in the field of computer science. The data base management system based on relational model developed with time. The new era in the database came with the advanced database systems, data warehousing and data mining for data analysis .The noticeable advancement in data analysis came after 1980's and the introduction of web & web based databases in 1990's change the information world altogether . With the beginning of the 21st century the drastic change in the IT was introduced with lots of new advanced data analysis techniques came into the industry. As data sets have grown in size, complexity and dimensions manual data analysis by experts has increasingly been replaced with automated data processing

ISSN (O) 2454-9665
ISSN (P) 2454-0663

using machines, aided by other discoveries in computer science, in the field of neural networks, cluster analysis, genetic algorithms, decision trees, and support vector machines. Pattern Discovery is the process of applying the above methods to large data sets for uncovering hidden patterns. The discovery algorithms are applied efficiently on the data stored in large data bases. Over the last few years, there has been an increasing interest in pattern discovery algorithms and the development of efficient and effective pattern discovery methods is a key area of study nowadays.

III. SIGNIFICANCE OF PATTERN DISCOVERY

Necessity is the mother of invention. Today in the world of competition, the companies require quick decision making to succeed in business. The Pattern Discovery generates decision factors on the basis of which effective and accurate decisions for the development of a company can be made. The decisions are based on timely presented facts, figures and information. The data is given highest priority as Garbage In will result in Garbage Out i.e. incomplete or inaccurate data will generate millions of false positives. Therefore validity, accuracy and completeness are prerequisites of any pattern discovery technique.

Today huge amount of data are being collected in the field of business, science, web, e-commerce etc. and are worked upon for generating beneficial outputs for future. With the rapidly growing volumes of data, there is an urgent need for a new automatic computational techniques and tools to assist humans in discovering patterns from these data. Low storage cost, high computational power and data base management system's capacity to handle large data bases has contributed to the requirement of Pattern Discovery.

Pattern Discovery automates the analysis of large volumes of data to detect relevant patterns in a database, using various data mining strategies and algorithms to look into current and historical data that can then be analyzed to predict future trends. Because data mining tools predict future trends and behaviors by reading through databases for hidden patterns, they allow organizations to make proactive, knowledge-driven decisions and answer questions that were previously too time-consuming to resolve. Knowledge Discovery is needed to discover sense from the data and use that information efficiently for better future. The good use of data in decision making has served as the driving force in Pattern Discovery.

Due to rapid changes in the business environment, the customers have become more demanding, markets are saturated and this requires company to use Pattern Discovery for developing effective policies to retain their customer share.

IV. CHARACTERISTICS OF GOOD PATTERN DISCOVERY TECHNIQUE

Following are some properties that every pattern discovery technique should possess –

- The technique should generate valid, useful and unique pattern.
- The technique should be scalable i.e. Data is growing continuously and it should work on massive amount of data.
- The technique should support easy integration into data base management system.
- The technique should support integration with search engine, data warehouses (DW) and cloud computing system as well.

ISSN (O) 2454-9665
ISSN (P) 2454-0663

- The technique should provide flexible data analysis i.e. it should work on all kinds of data such as spatio-temporal data, multimedia data, graph data, Web data etc.
- The technique should generate easily understandable results.

V. PATTERN DISCOVERY WITH DATA MINING TECHNIQUES

Data mining tasks are used to extract patterns from large data sets. Some researchers define Data Mining as a step in the KDD process which consists of applying data analysis and discovery algorithms that, under acceptable computational efficiency limitations, produce a particular enumeration of patterns over the data. Data Mining has been defined by many researchers in the following way

- DM ...is nothing else than torturing the data until it confesses…and if you torture it enough, you can get it to confess to anything (Fred Menger)
- "Data Mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner."
- (Hand, Mannila, Smyth. Principles of Data Mining. MIT Press. 2001.)
- Data mining is the use of statistical methods with computers to uncover useful patterns inside databases - JeromeH. Friedman [1].
- Data mining is the process of extracting previously unknown, comprehensible, and actionable information from large databases
  and using it to make crucial business decisions. – Zekulin [2]
- Data mining is a decision support process where we look in large databases for unknown and unexpected patterns of information. – Parsaye [2].
- KDD (Knowledge Discovery in Databases) was defined by (Fayyad, Piatetsky-Shapiro, & Smyth 1996) as knowledge Discovery in Databases is defined as the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data.[4]
- Knowledge discovery in databases is mainly concerned with identifying interesting patterns and describing them in a concise and meaningful manner [5].
- Data mining involves the use of sophisticated data analysis tools to discover previously unknown, valid patterns and relationships in large data sets. [6]

VI. CLASSIFICATION OF PATTERN DISCOVERY TECHNIQUES

Since Pattern Discovery is a multidisciplinary field, Researchers belonging to different communities have given their own classification of pattern discovery techniques. Therefore a variety of classification schemes exist for Pattern Discovery Techniques. Pattern Discovery Methods can be classified on the kind of databases on which the pattern discovery techniques will be applied, the type of knowledge discovered or on the basis of methods used for pattern discovery. The pattern discovery methods are broadly classified into the following based on the approach used for finding patterns-

- Verification Driven Pattern Discovery
- Discovery Driven Pattern Discovery

ISSN (O) 2454-9665
ISSN (P) 2454-0663

6.1 Verification Driven Pattern Discovery Techniques

Verification Driven technique is where the system is limited to verifying the user's hypothesis .In this technique, the statistical analysis and query & reporting tools are applied on the data to find patterns. The prior knowledge about the data exists and therefore hypothesis about the relationship between variables is framed and verified using this technique. The technique is not able to find new, unknown patterns that may exist in data. Therefore it can be said that no new knowledge is discovered in this branch of pattern discovery.

6.2 Discovery Driven Pattern Discovery Techniques

Discovery Driven technique is, where the system autonomously finds new patterns. The Discovery Driven technique may be classified on the basis of type of tasks performed by the patterns or by the methods used for finding patterns. The Discovery Driven may be subdivided into two broad areas based on the tasks performed by the patterns-

- Predictive, where the system finds patterns for the purpose of predicting the future behavior of some attributes given the values of other known attributes.
- Descriptive, where the system finds patterns for the purpose of presenting them to a user in a human-understandable form.

The Machine Learning community has categorized the pattern discovery techniques as follows based on the methods used for finding patterns

- Supervised Pattern Discovery Techniques- Most Pattern Discovery methods are supervised methods meaning that there is a identified target variable and the data mining algorithm is given many examples where the value of the target variable is provided, so that the algorithm may learn from the data which values of the target variable are associated with which values of the predictor variables and then will be used in doing prediction. Therefore can be called predictive technique as well.
- Unsupervised Pattern Discovery Techniques-In Unsupervised Pattern Discovery Techniques, all variables are independent variables and are used for model building. There is no output variable and this technique is used for finding unknown, hidden patterns that exist in data. Therefore can be called descriptive technique as well.

6.2.1 Supervised Pattern Discovery Techniques

Those techniques which are considered predictive are sometimes termed as supervised techniques. Predictive techniques finds patterns by making a prediction on behavior of some unknown attribute given the values of other known attributes.

Supervised learning, often also called as direct data mining, discovers patterns in the data that relate data attributes with a target attribute. These patterns are then used to predict the values of the target attribute in future data instances. Supervised methods are methods that attempt to discover the relationship between input attributes and a target attribute. Supervised learning requires the target variable to be well defined and a sufficient number of its values. The discovered relationship is represented in a structure referred to as a Model. Usually models describe and explain phenomena, which are hidden in the dataset and can be used for predicting the value of the target attribute knowing the values of the input attributes. The supervised methods can be

ISSN (O) 2454-9665
ISSN (P) 2454-0663

implemented on a variety of domains such as marketing, finance and manufacturing. The purpose of the analysis is to specify a relationship between the explanatory variables and the dependent variable as it is done in regression analysis. To apply directed data mining techniques the values of the dependent variable must be known for a sufficiently large part of the data set.

Pattern Discovery tasks that belongs to predictive model are

- Classification
- Regression

Classification Technique

Classification is a data mining task has been studied for many decades by the machine learning and statistics communities. Classification is in some way similar to the Clustering, but it requires prior knowledge about how classes are defined. Classification assigns items in a collection of data to target categories. The goal of this task is to accurately predict the value of a user-specified goal attribute based on the values of other attributes, called the predicting attributes. In the classification task the data being mined is divided into training and the test sets. The data mining algorithm tries to discover relationships between the attributes in training set,

which would make it possible to predict the output. Once the training process is finished and the algorithm has found a set of classification rules, these rules are evaluated on the test set. The algorithm analyzes the input test data and outputs a prediction. Classification can be used both to understand the existing data and to predict how new instances will behave.

Accuracy of classification model is judged by the percentage of correct predictions made by the model when compared with the actual classifications in the test data. Evaluation of the performance of a classification model is based on the counts of test records predicted correctly and incorrectly by the model. Popular Classification Techniques are:-

- Decision Tree based Methods
- Neural Networks
- Naïve Bayes
- Support Vector Machines

Regression Technique

Regression predicts new values based on the past inference and it computes the new values for a dependent variable based on the values of one or more measured attributes. The Regression task is similar to classification. Classification predicts categorical labels whereas Regression models continuous-valued functions. The main difference is that the predictable attribute is a continuous number i.e. Regression is used to predict missing or unavailable numerical data values. Regression Technology is very important in the field of statistics. Regression is a data mining function that analyze the dependency of some attribute values upon the values of other attributes in a set of data items and then produce a model that can predict these attribute values for new records. Linear Regression is used to approximate the relationship between a continuous response variable and a set of predictor variables. There can be number of independent variables which together produce a result i.e. a dependent variable. For example, given a data set of credit card transactions, regression builds a model that can

ISSN (O) 2454-9665
ISSN (P) 2454-0663

predict the likelihood of fraudulence for new transactions. Regression also encompasses the identification of distribution trends based on the available data.

6.2.2 Unsupervised Pattern Discovery Techniques

Descriptive model is used to determine the patterns and relationships in a sample data. According to [Kohavi and Provost (1998)][7] the term "Unsupervised learning" refers to "learning techniques that group instances without a pre-specified dependent attribute".

In Unsupervised learning, either the target variable is unknown or has been recorded only for a small number of cases. We need to explore the data to find inherent patterns in it. Unsupervised models do not predict a target value but focus on finding some kind of intrinsic structure and relations in the data. Unsupervised models are also called descriptive models.

Unsupervised Pattern Discovery tasks are

- Clustering
- Association

In unsupervised learning situations all variables are treated in the same way, there is no distinction between explanatory and dependent variables as it is in supervised learning. The most common form of unsupervised data mining method is clustering.

Clustering Technique

Clustering is the process of partitioning a set of data in a set of meaningful sub-classes, called clusters. A cluster is a collection of data objects that are similar to one another in some sense and thus can be treated collectively as one group. A good clustering method produces high-quality clusters to ensure that the inter-cluster similarity is minimum and the intra-cluster similarity is maximum. Clustering analysis identifies clusters embedded in the data. The terms segmentation and partitioning are sometimes used as synonyms for Clustering.

Given a set of data items, clustering partition this set into a set of sub-classes such that items with similar characteristics are grouped together. Clustering is best used for finding groups of items that are similar when no prior knowledge of data exists.

Clustering is an unsupervised learning task as no class values denoting a prior grouping of the data instances are given. Due to historical reasons, clustering is often considered synonymous with unsupervised learning. In fact, clustering is one of the most utilized data mining techniques. It has a long history and used in almost every field e.g. medicine, psychology, botany, sociology, biology, archeology, marketing, insurance, libraries etc. In recent years, due to the rapid increase of online documents, text clustering has become important.

There is no objectively correct clustering algorithm. A clustering algorithm may perform better in general although it may be more suitable for some specific types of data or applications. It is a challenging task to choose the “best” algorithm. Every algorithm has limitations and works well with certain data distributions. It is very hard to know what distribution the application data follow. The data may not fully follow any “ideal” structure or distribution required by the algorithms. One also needs to decide how to standardize the data, to choose a suitable distance function and to select other parameter values. Due to these complexities, the common practice is to run several algorithms using different distance functions and parameter settings and then carefully analyze and compare the results. The interpretation of the results must be based on insight into the meaning of the original data together with the knowledge of algorithms used. Clustering is highly application dependent and

ISSN (O) 2454-9665
ISSN (P) 2454-0663

to certain extent subjective i.e. it depends on the personal preferences of the user. Clustering is useful for exploring data. The quality of a clustering is very hard to evaluate because we do not know the correct clusters. Evaluation of Cluster is based on internal information like Intra-cluster cohesion and inters cluster separation. The Cohesion measures how near the data points in a cluster are to the cluster centroid and Separation means that different cluster centroid should be far away from one another.

Traditionally clustering techniques are broadly divided in *hierarchical* and *partitioning.*

- Hierarchical Clustering:-Hierarchical clustering builds a cluster hierarchical tree, also known as a dendrogram which displays both the cluster-sub cluster relationships and the order in which the clusters were merged (agglomerative) or split (divisive view). Every cluster node contains child clusters; sibling clusters partition the points covered by their common parent. Such an approach allows exploring data on different levels of granularity.

Types of hierarchical clustering

- Agglomerative (bottom up) clustering:
- Divisive (top down) clustering

Partitional Clustering

The Partitioning methods are used to find exclusive clusters of spherical shape.K-means is an example of partitional clustering algorithm. The *K*-Means algorithm [8][9] is by far the most popular clustering tool used in scientific and industrial applications. Data partitioning algorithm divides data into several subsets. Because checking all possible subset systems is computationally infeasible, certain greedy heuristics are used in the form of iterative optimization. Partitional methods are more useful for small to medium size data sets. The K-Means uses centroid to represent the cluster. The centroid is the centre point of the cluster and can be defined as mean or medoid of points assigned to the cluster.

Association Technique

The task of discovering association rules was first introduced in 1993. Association rule mining is also unsupervised Association analysis is a data mining function that discovers the probability of the co-occurrence of items in large data sets. The uncovered relationships between co-occurring items are expressed as association rules or sets of frequent items. Each rule has two measurements, support and confidence. Confidence is a measure of the rule's strength i.e. how likely do these items co-occur in the data; while support corresponds to statistical significance i.e. how often do these items co-occur in the data. An association rule is an implication of the form $X \rightarrow Y$ where X is a set of antecedent items and Y is the consequent item. The pattern is valid pattern if its support value is larger than a pre-defined threshold value. Association rules are considered interesting if they satisfy a minimum support threshold and a minimum confidence threshold [10]. Users or domain experts can specify such thresholds.

An Association model is often used for market basket analysis, which is widely used in data analysis for direct marketing, catalog design, sales promotions and other business decision-making processes. There are two key issues that need to be addressed when applying association analysis to market basket data. First, discovering patterns from a large transaction data set can be computationally expensive. Second, some of the discovered patterns lack authenticity or validity because they may happen simply by chance.

International Journal of Innovative Research in Science and Engineering
Vol. No.3, Issue 01, January 2017
www.ijirse.com
ISSN (O) 2454-9665
ISSN (P) 2454-0663

Sequential pattern mining is an important data mining task that has been studied extensively by many researchers. It was first introduced by Agrawal and Srikant in [11]: Setting min_support is a subtle task: A too small value may lead to the generation of thousands of patterns, whereas a too big value may lead to no answer. To come up with an appropriate number of patterns, one needs to have prior knowledge about the mining query and the task-specific data and be able to estimate beforehand how many patterns will be generated with a particular threshold.

A common strategy adopted by many association rule mining is to split the problem in to two major steps:

1. Generation of Frequent item sets
2. Generation of Rules

The efficiency of frequent itemset mining algorithms is determined mainly by three factors: the way candidates are generated, the data structure that is used and the implementation details. There are dozens of algorithms used for frequent item sets mining. Some of them, very well known, started a whole new era in pattern discovery. They made the concept of mining frequent item sets and association rules possible. Others are variations that bring improvements mainly in terms of processing time.

Popular Algorithms for Mining Frequent Itemsets are AIS ,APRIORI, FP-TREE GROWTH ALGORITHM, ECLAT, - Echivalence Class Clustering and Bottom-up Lattice Traversal, TREE-PROJECTION, ASCAL, RELIM, CLOSET, CHARM.

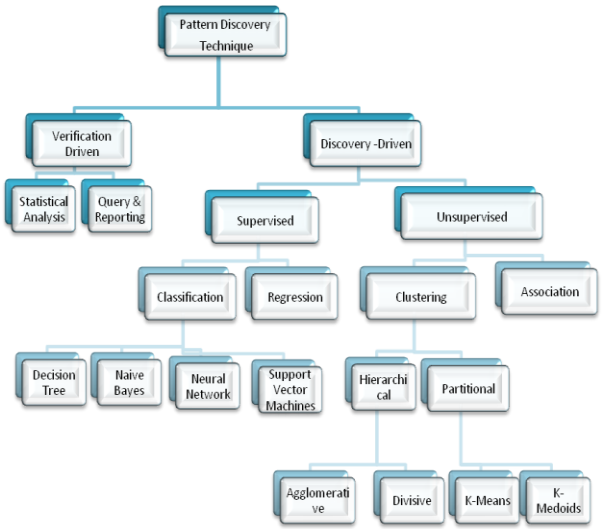

Figure 2 Classification of Pattern Discovery Techniques

VII. PATTERN DISCOVERY: REASONS

- Reduced cost of large storage devices and increasing ease of collecting data over networks.Lots of data is data is being collected and warehoused .Eg Web data, e-commerce,purchases at department/ grocery stores,Bank/Credit Card transactions
- Availability of Robust and Efficient machine learning algorithms to process data.
- Reduced cost of computational power — enabling use of computationally intensive methods for data analysis.

ISSN (O) 2454-9665
ISSN (P) 2454-0663

- Make use of your data assets. There is often information "hidden" in the data that is not readily evident .Human analysts may take weeks to discover useful information .
- There is a big gap from stored data to knowledge; and the transition won't occur automatically.
- Many interesting things you want to find cannot be found using database queries.Traditional techniques infeasible for raw and enormous data .
- Pattern Discovery may help scientists in classifying and segmenting data in Hypothesis Formation

VIII. RECENT TRENDS IN THE FIELD OF PATTERN DISCOVERY

The advancement in technology has open up various dimensions of pattern discovery. Initially the technique was utilized only for decision making & market basket analysis. The significance of pattern discovery was realized with time and the development of interactive, efficient and effective pattern discovery methods became a key area of study for researchers .The easy integration of pattern discovery methods into database management system is also an active area of research these days .More application specific pattern discovery methods are being developed for example methods specific for web and text data analysis, financial data analysis etc. The pattern discovery methods should support integration with Search Engines, Data Warehouses and Cloud computing System. Lot of research in this field is being done by research community. The recommender system is being developed to assist user in seeking information from the web and lot of work in the field of web personalization is also being done using pattern discovery techniques.

IX. PATTERN DISCOVERY: LIMITATIONS AND CHALLENGES

There are always some limitations and challenges involved with every technology and Pattern Discovery is no exception to this. The Pattern Discovery can discover patterns and relationships but it always require human intervention to judge the significance of the patterns as interestingness of pattern is a subjective issue and depends on the user's requirement of knowledge. Another limitation is that Pattern Discovery may retrieve all patterns that exist in data but it cannot judge the effect of additional external variables on the behavior of data. There are many challenging issues in pattern discovery .The areas include mining methodology, user interaction, efficiency and scalability, and dealing with diverse data types. The main challenges to the Pattern Discovery procedure may be presented as follows:

- Management of changing data and knowledge: Rapidly changing data, in a database may make previously discovered patterns invalid. Possible solutions include incremental methods for updating the patterns.
- Larger databases: Databases with hundreds of fields and tables, millions of records and multi-gigabyte size are quite common today. The huge size of the database, the wide distribution of data and the computational complexity of some Pattern Discovery methods motivate the development of parallel and distributed Pattern Discovery algorithms.
- High dimensionality: Not only is there often a very large number of records in the database, but there can also be a very large number of fields so that the dimensionality of the problem is high. In addition, it increases the chances that a Pattern Discovery algorithm will find patterns that are not valid in general.

ISSN (O) 2454-9665
ISSN (P) 2454-0663

- Handling of relational, complex and heterogeneous data: Because there are many kinds of data and databases used in different applications, one may expect that a knowledge discovery system should be able to perform effective Pattern Discovery on different kinds of data. Many applicable databases contain complex data types, such as structured data and complex data objects, hypertext and multimedia data, spatial and temporal data, transaction data, legacy data etc. A powerful system should be able to perform effective Pattern Discovery on such complex types of data along with relational data.
- Efficiency and scalability of data mining algorithms: To effectively extract information from a huge amount of data in databases, the pattern discovery algorithms must be efficient and scalable to large databases. That is, the speed of Pattern Discovery algorithm must be predictable and acceptable in large databases. Algorithms with exponential or even medium-order polynomial complexity will not be of practical use.
- Data Quality: The data can be missing and/or noisy. Noise and exceptional data should be handled tactfully in Pattern Discovery systems because it may generate false results. This also motivates a systematic study of measuring the quality of the discovered knowledge, including interestingness and reliability, by construc-tion of statistical, analytical and simulative models and tools.
- Expression of various kinds of Pattern Discovery requests and results: Different kinds of knowledge/information can be discovered from a large amount of data. Also, one may like to examine discovered knowledge from different viewpoints and present them in more understandable forms. This requires us to express both the Pattern Discovery requests and the discovered knowledge in high-level languages or graphical user interfaces so that the Pattern Discovery task can be specified by non experts and the discovered knowledge can be understandable and directly usable by users. This also requires the discovery system to adopt expressive knowledge representation techniques.
- Mining information from different sources of data: The widely available local and wide-area computer network, including Internet, connect many sources of data and form huge distributed heterogeneous data-bases. Mining knowledge from different sources of formatted or unformatted data with diverse data se-mantics poses new challenges to Pattern Discovery.
- Protection of privacy and data security: The urge for knowledge discovery requires data to be viewed at different angles and at different levels of abstraction. It is important to study when knowledge discovery may lead to an invasion of individual's personal information and security measures should be developed to prevent the disclosure of sensitive information.
- Integration: Pattern Discovery tools are often only a part of the entire decision making system. It is desirable that they integrate smoothly, both with the database and the final decision making procedure.

X. CONCLUSION

Pattern Discovery can help companies to discover purchasing pattern of customers, which can lead in reduction of advertising budget. With so many advantages in hand with Pattern Discovery there are some issues that need to be considered. Since Pattern Discovery software lacks the human experience and intuition to recognize the difference between a relevant and an irrelevant correlation, statistical analyst will remain in high demand. The other issue is related to privacy of consumer's data. Users of Pattern Discovery should start thinking about how the use of this technology will be impacted by legal issues related to privacy. The personal data should be

ISSN (O) 2454-9665
ISSN (P) 2454-0663

protected in terms of security, quality, purpose, use, openness, individual participation and accountability. The primary purpose of the collection must be clearly understood by the consumer and identified at the time of collection.

With so many advantages in hand, Data Mining in future will end up as standard tool built into database or data warehouse software product. The functionality of database marketing products will increase to integrate with relational database products and with DSS application. Database marketing software applications will have a tremendous impact on how business is done in the future. The successful database marketing application will combine Pattern Discovery technology with a thorough understanding of business problems and present the results in a way that the user can understand. At that point people who can turn what is known into what can be done will understand the knowledge contained in a database.

REFERENCES

[1.] Jerome H. Friedman. Data Mining and Statistics: What's the Connection? URL:http://stat.stanford.edu/~jhf/dm-stat.ps.Z

[2.] Alex Zekulin, Parsaye

[3.] George H. John. Enhancements to the Data Mining Process. Ph.D. Thesis, Department of Computer Science, Stanford University, March 1997

[4.] Fayyad, Patetsky -Shapiro, Smyth, and Uthurusamy (eds.).Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press, 1995.

[5.] W. Frawley and G. Piatetsky-Shapiro and C. Matheus (Fall 1992). "Knowledge Discovery in Databases: An Overview". AI Magazine: pp. 213-228. ISSN 0738-4602.

[6.] Data Mining and Homeland Security : An Overview

[7.] Kohavi and Provost (1998) Ron Kohavi, Alex Deng, Brian Frasca, Roger Longbotham, Toby Walker, Ya Xu, Trustworthy Online Controlled Experiments: Five Puzzling Outcomes Explained, KDD 2012. Powerpoint slides, DOI.

[8.] Hartigan, J.A. (1975), Clustering Algorithms, New York: John Wiley & Sons, Inc.

[9.] Hartigan, J. A.; Wong, M. A. (1979). "Algorithm AS 136: A K-Means Clustering Algorithm". Journal of the Royal Statistical Society, Series C 28 (1): 100–108. JSTOR 2346830

[10.] C. Győrödi, R. Győrödi. “Mining Association Rules in Large Databases”. Proc. of Oradea EMES’02: 45-50, Oradea, Romania, 2002.

[11.] AGRAWAL, R., IMIELINSKI, T., AND SWAMI, A. N. 1993. Mining association rules between sets of items in large databases. In Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, P. Buneman and S. Jajodia, Eds. Washington, D.C., 207—216.