International Journal of Emerging Technology in Computer Science & Electronics (IJETCSE)
ISSN: 0976-1353 Volume 24 Issue 3 – FEBRUARY 2017 (SPECIAL ISSUE).

SURVEY OF DATA MINING: PROCESS, GOALS, CHALLENGES AND APPLICATIONS

E.PRIYA #1 and G. VINITHA *2

# M.SC COMPUTER SCIENCE, KAMBAN COLLEGE OF ARTS AND SCIENCE FORM WOMEN, THIRUVANNAMALAI, INDIA.

***Abstract*— Generally mining of data is a well-known technique for automatically and intelligently extracting information or knowledge from a large amount of data, however, it can also disclosure sensitive information about individuals compromising the individual's right to privacy. It is a process to extract the implicit information; knowledge which is potentially useful and people do not know in advance, and this extraction is from the mass, incomplete, noisy, fuzzy and random data. Therefore, privacy conserving data mining has becoming an increasingly important field of research. Now a day, data mining is emerging area to extract implicit and useful knowledge and also predictable as an important technology for business internationally and locally. Hence this paper sketches vision of the future work to done in area of data mining. This paper elaborate various topics (starting from the classic definition of "data mining" and its basic terms) included various future challenges and issues in data mining which is important to do further more research in this emerging field.**

***Index Terms*— Data mining, KDD process, algorithms, Data mining goals, Challenging problems, and applications.**

I. Introduction

A. DEFINITION OF DATA MINING:

Data mining (sometimes called data or knowledge discover) is the process of analyzing data from different perspectives and summarizing it into useful information that can be used to increase revenue, cuts costs, or both. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases.

B. Data, Information, and Knowledge

1) Data

Data are any facts, numbers, or text that can be processed by a computer. Today, organizations are accumulating vast and growing amounts of data in different formats and different databases. This includes:

1. Operational or transactional data such as, sales, cost, inventory, payroll, and accounting.
2. Nonoperational data, such as industry sales, forecast data, and macro-economic data.
3. Meta data-data about the data itself, such as logical database design or data dictionary definitions.

2) Information

The patterns, associations, or relationships among all this data can provide information. For example, analysis of retail point of sale transaction data can yield information on which products are selling and when.

3) Knowledge

Information can be converted into knowledge about historical patterns and future trends. For example, summary information on retail supermarket sales can be analyzed in light of promotional struggles to provide knowledge of consumer buying behavior. Thus, a manufacturer or retailer could determine which items are most susceptible to promotional efforts.

C. What is the KDD process?

The term knowledge Discovery in Databases, or KDD for short, refers to the broad process of finding knowledge in data, and emphasizes the "high-level" application of particular data mining methods. It is of interest to researchers in machine learning, pattern recognition, databases, statistics, artificial intelligence, knowledge acquisition for expert systems, and data visualization.

The unifying goal of the KDD process is to extract knowledge from data in the context of large databases.

It does this by using data mining methods(algorithms) to extract(identify) what is deemed knowledge, according to the specifications of measures and thresholds, using a database along with any required preprocessing, subsampling, and transformations of that database.

The KDD(Knowledge Discovery in Database)process comprises of a few steps leading from raw data collections to some form of new knowledge.

a) Data cleaning: It also known as data cleaning, it is a phase in which irrelevant data and noise data are removed from the raw collection of data.

b) Data Integration: In this multiple data sources, often heterogeneous maybe combined in a common source.

c) Data selection: At this step, the data relevant to the analysis is decided on and retrieved from the data collection.

d) Data transformation: It also known as data consolidation, in this process selected data is transferred into forms appropriate for the mining procedure.

e) Pattern evaluation: At this level, strictly interesting patterns representing knowledge are identified based o given measures.

f) Data mining process: Data mining process consists of an iterative sequence of several steps/process: data preprocessing; data management; data mining tasks and algorithms, and post processing.

g) Knowledge representation: It is final phase of KDD process in which discovered knowledge is visually represented to the user. This essential step uses visualization techniques to help interpret the data mining results.

KDD refers to the overall process of discovering useful knowledge from data. It involves the evaluation and possibly interpretation of the patterns to make the decision of what qualifies as knowledge. It also includes the choice of encoding schemes, preprocessing, sampling, and projections of the data prior to the data mining step.

Data mining refers to the application of algorithms for extracting patterns from data without the additional steps of the KDD process.

II. Data Warehouses

Dramatic advances in data capture, processing power, data transmission, and storage capabilities are enabling organizations to integrate their various databases into data warehouses. Data warehousing is defined as a process of centralized data management and retrieval. Data warehousing, like data mining, is a relatively new term although the concept itself has been around for years. Data warehousing represents and ideal vision of maintaining a central repository of all organizational data. Centralization of data is needed to maximize user access and analysis. Dramatic technological advances are making this vision a reality for many companies. And, equally draw the data analysis software is what supports data mining.

A. What can data mining do?

Data mining is primarily used today by companies with a strong consumer focus-retail, financial, communication, and marketing organizations. It enables these companies to determine relationships among “internal” factors such as price, product positioning, or staff skills, and “external” factors such as economic indicators, competition, and customer demographics. And, it enables them to determine the impact on sales, customer satisfaction, and corporate profits. Finally, it enables them to “drill down” into summary information to view detail transactional data.

With data mining, a retailer could use point-of-sake records of customer purchases to send targeted promotions based on an individual’s purchase history. By mining demographic data from comment or warranty cards, the retailer could develop products and promotions to appeal to specific customer segments.

For example, Blockbuster Entertainment mines its video rental history database to recommend rentals to individual customers. American Express can suggest products to its cardholders based on analysis of their monthly expenditures.

B. How does data mining work?

While large-scale information technology has been evolving separate transaction and analytical systems, data mining provides the link between the two. Data mining software analyzes relationships and patterns in stored transaction data based on open-ended user queries. Several types of analytical software are available: statistical, machine learning, and neural networks. Generally, any of four types of relationships are sought:

- **Classes:** Stored data is used to locate data in predetermined groups. For example, a restaurant chain could mine customer purchase data to determine when customers visit and what they typically order. This information could be used to increase traffic by having daily specials.
- **Clusters:** Data items are grouped according to logical relationships or consumer preferences. For example, data can be mined to identify market segments or consumer affinities.
- **Associations:** Data can be mined to identify associations. The beer-diaper example is an example of associative mining.

Sequential patterns: Data is mined to anticipate behavior patterns and trends. For example, an outdoor equipment retailer could predict the likelihood of a backpack being purchased based on a consumer’s purchase of sleeping bags and hiking shoes.

Data mining consists of five major elements:

- Extract, transform, and load transaction data onto the data warehouse system.
- Store and manage the data in a multidimensional database system.
- Provide data access to business analysts and information technology professionals.
- Analyze the data by application software.
- Present the data in a useful format, such as a graph or table.

Different levels of analysis are available:

- **Artificial neural network:** Non-linear predictive models that learn through training and resemble biological neural networks in structure.
- **Genetic algorithms:** Optimization techniques that use processes such as genetic combination, mutation, and natural selection in a design based on the concepts of natural evolution.
- **Decision trees:** Tree-shaped structures that represent sets of decisions. These decisions generate rules form the classification of a dataset. Specific decision tree methods include Classification And Regressing Trees(CART) and Chi Square Automatic Interaction Detection (CHAID). CART and CHAID are decision tree techniques used for classification of a dataset. They provide a set of rules that you can apply to a new (unclassified) dataset to predict which

records will have a given outcome. CART segment a dataset by creating 2-way splits while CHAID segments using chi square tests to create multi-way splits. CART typically requires less data preparation than CHAID.

- **Nearest neighbor method:** A technique that classified each record in a dataset based on a combination of the classes of the k record(s) most similar to it in a historical dataset (where k 1). Sometimes called the k-nearest neighbor technique.
- **Rule induction:** The extraction of useful if-then rules from data based on statistical significance.
- **Data visualization:** The visual interpretation of complex relationships in multidimensional data. Graphics tools are used to illustrate data relationships.

What technological infrastructure is required?

Today, data mining applications are available on all size systems form mainframe, client/server, and PC platforms. System prices range from several thousand dollars for the smallest applications up to $1 million a terabyte for the largest. Enterprise-wide applications generally range in size from 10 gigabytes to over 11 terabytes. NCR has the capacity to deliver applications exceeding 100 terabytes. There are two critical technological drivers:

- **Size of the database:** The more data being processed and maintained, the more powerful the system required.
- **Query complexity:** The more complex the queries and the greater the number of queries being processed, the more powerful the system required.

Relational database storage and management technology is adequate for many data mining applications less than 50 gigabytes. However, this infrastructure needs to be significantly enhanced to support larger applications. Some vendors have added extensive indexing capabilities to improve query performance. Others use new order-of-magnitude improvements in query time. For example, MPP systems from NCR link hundreds of high-speed Pentium processors to achieve performance levels exceeding those of the largest supercomputers.

III. DATA MINING GOALS

In general, data mining is used for a variety of purpose in both private and public sectors. Industries such as insurance, banking, medicine, and retailing commonly use data mining to increase sales, enhance research, and reduce costs. Hence the goals of data mining instead applications also as:

a) **Data processing:** Depending on the goals and requirements of the KDD process, analysts may select, filter, aggregate, sample, clean and/or transform data. Automating some of the most typical data processing tasks and integrating them seamlessly into the overall process may eliminate or at least greatly reduce the need for programming specialized routines and for data export/import, thus improving the analyst's productivity.

b) **Association rule learning:** It searches for relationships between variables for e.g. a supermarket might gather data on customer purchasing habits. Using association rule learning, the supermarket can determine which products are frequently bought together and use this information for marketing purposes. This is sometimes referred to as market basket analysis.

c) **Prediction:** Give a data item and a predictive model, predict the value for a specific attribute of the data item. For example, give a predictive model of credit card transactions, predict the likelihood than a specific transaction is fraudulent. Prediction may also be used to validate a discovered hypothesis.

d) **Regression:** For a given set of data items, regression is the analysis of the dependency of some attribute values upon the values of other attributes in the same item, and the automatic production of a model that can predict these attribute values for new records for e.g. given a data set of credit card transactions, build a model that can predict the likelihood of fraudulence for new transactions.

e) **Classification:** Given a set of predefined categorical classes, determine-which of these classes a specific data item belongs for e.g. given classes of patients that corresponds to medical treatment responses; identify the form of treatment to which a new patient is most likely to respond.

f) **Link analysis/Associations:** In this, items; identify relationships between attributes and items such as the presence of one pattern implies the presence of another pattern. These relations may be associations between attributes within the same data item(out of the shoppers who bought milk, 64% also purchased bread) or associations between different data items (Every time a certain stock drops 5%, raises 13% between 2 and 6 weeks later). The investigation of relationships between items over a period of time is also often referred to as sequential pattern analysis.

g) **Clustering:** In this for a given a set of data items, partition this data items into a set of classes such that items into a items with similar characteristics are grouped together. Clustering is best used for finding groups of items that are similar for e.g. given a data set of customers; identify subgroups of customers that have a similar buying behavior.

h) **Exploratory Data Analysis (EDA):** It is the interactive exploration of a data set without heavy dependence on preconceived assumptions and models, thus attempting to identify interesting patterns. Graphic representations of the data are used very often to exploit the power of the eye and human intuition. While there are dozens of software packets available that were developed exclusively to support data exploration, it might also be desirable to integrate these approaches into an overall KDD environment.

i) **Model Visualization:** Visualization plays an important role in making the discovered knowledge understandable and interpretable by humans. Besides, the human eye-brain system itself still remains the best pattern-recognition device known. Visualization techniques may range from simple scatter plots and histogram plots over parallel coordinates to 3D movies.

j) **Summarization:** In short, it provides a compact description for a subset of data.

k) **Dependency modeling:** It describes significant dependencies among variables.

IV. CHALLENGING PROBLEMS IN DATA MINING

This section discusses various challenges in area of data mining to processing and extracting valuable data from collected data in real world that are discussed by Hristidis etal.(2010), as preprocessing, post processing; data mining tasks and algorithms.

A. *Developing a Unifying theory of data mining*

Several respondents feel that the current stat of the art of the art of data mining research it too “ad-hoc”. Many techniques are designed for individual problems, such as Classification or clustering, but there is no unifying theory. However, a theoretical framework that unifies different data mining tasks including clustering, classification, association rules, etc., as well as different data mining approaches(such as statistics, machine leaning, database systems, etc.), would help the field and provide a basis for future research.

B. *Difficulty Accessing data*

To accessing data the is so typical and have a major challenge for data mining also, for e.g. it is scattered throughout an organization, more commonly accessing data because it does not exist.

Data miners generally agreed that difficulty accessing data is due to the lack of plan or strategy for data i.e.-how it can be obtained, what data is needed, how quality can be assured or improved and how it can be maintained etc. Again data miners suggest working directly with business users to match business problems with data requirements, and to use this as way to begin developing a broader plan for data collection and data accessibility.

C. *Mining Sequence Data and Time series data:*

Sequential and time series data mining remains and important problem. Despite progress in other related fields, how to efficiently cluster, classify and predict the trends of these data is still an important open topic.

A particularly challenging problem is the noise in time series data. It is an important open issue to tackle. Many time series used form predictions are contaminated by noise, making it difficult to do accurate short-term and long-term predictions. Some of the key issues that need to addressed in the design of a practical data miner for noisy time series include:

- **Information/search agents to get information:** Use of wrong, too many, or too little search criteria; possibly inconsistent information from many sources; semantic analysis of (Meta) information; assimilation of information into inputs to predictor agents.
- **Learner/miner to modify information selection criteria:** Apportioning of biases to feedback;

Developing rules for Search Agents to collect information; developing rules for information Agents to assimilate information.

- **Predictor agents to predict trends:** Incorporation of qualitative information; multi objective optimization not in closed form.

V. SECURITY, PRIVACY, AND DATA INTEGRITY

Several researchers considered privacy protection in data mining as an important topic. That is, how to ensure the users privacy while their data are being mined. Related to this topic is data mining for protection of security and privacy. One respondent states that if we do not solve the privacy issue, data mining will become a derogatory term to the general public. Some respondents consider the problem of knowledge integrity assessment to be important. We quote their observations: “Data mining algorithms are frequently applied to data that have been intentionally modified from their original version, in order to misinform the recipients of the data or to counter privacy and security threats. Such modifications can distort, to an unknown extent, the knowledge contained in the original data. As a result, one of the challenges facing researchers is the development of measures not only to evaluate knowledge integrity of a collection of data, but also of measures to evaluate the knowledge integrity of individual patterns. Additionally, the problem of knowledge integrity assessment presents several challenges.”

A. *Dirty Data:*

Here no surprise that dirty data tops the list, because it has been at the top of the list form the past several years in area of data mining as a challenging issue. Many data miners provided input as- How they have tried to overcome the problem and how to provide a clear theme emerges those involving business users. Data miners use descriptive statistics and visualization to assist business in understanding their data and identifying problem areas etc. Helping users understand their data hands on and helps everyone to gain a shared understanding about the quality of the data. This can help manage expectations about providing potential results of a data modeling exercise and also create action plans to improve quality of data.

B. *Data mining in a Network setting*

1) *Community and social networks*

Today’s world is interconnected through many types of links. These links include Web pages, blogs, and emails. Many respondents consider community mining and the mining of social networks as important topics. Community structures are important properties of social networks. The identification problem in itself is a challenging one. First, it’s critical to have the right characterization of the notion of “community” that is to be detected. Second, the entities/nodes involved are distributed in real-life applications, and hence distributed means of identification will be desired, third, a snapshot-based dataset may not be able to capture the real picture; what is most important lies in the local relationships (e.g. the nature and frequency of local interactions) between the entities/nodes. Under these circumstances, our challenge is to understand (1) the network's static structures (e.g. topologies and cluster) and (2) dynamic behavior (such as growth factors, robustness, and functional efficiency). A similar challenge exists in bio-informatics, as we are currently moving our attention to the dynamic studies of regulatory networks.

VI. APPLICATION OF DATA MINING:

A. *Service providers*

The first example of Data Mining and business intelligence comes from service providers in the mobile phone and utilities industries. Mobile phone and utilities companies use Data mining and business intelligence to predict 'churn', the terms they use for when a customer leaves their company to get their phone/gas/broadband from another provider. They collate billing information, customer services interactions, website visits and other metrics to give each customer a probability score, then target offers and incentives to customers whom they perceive to be at a higher risk of churning.

B. *Retail*

Another example of Data mining and business Intelligence comes from the retail sector. Retailers segment customers into 'Recency, Frequency, Monetart' (RFM) groups and target marketing and promotions to those different groups. A customer who spends little but often and last did so recently will be handled differently to a customer who spent big but only once, and also some time ago. The former may receive a loyalty, up sell and cross-sell offers, whereas the latter may be offered a win-back deal, for instance.

C. *E-Commerce*

Perhaps some of the most well-known examples of data mining and analytics come from E-Commerce sites. Many E-Commerce companies use Data mining and business intelligence to offer cross-sells and up-sells though their websites. One of the most famous of these is, of course, Amazon, who use sophisticated mining techniques to drive there, 'People who viewed that product, also liked this functionality.

D. *Supermarkets*

Supermarkets provide another good example of data mining and business intelligence in action. Famously, supermarket loyalty card programmers are usually driven mostly, if not solely, by the desire to gather comprehensive data about customers for use in data mining. One notable recent example of this was with the US retailer target. As part of its data mining program me, the company developed rules to predict if their shoppers were likely to be pregnant. By looking at the contents of their customers shopping baskets, they could spot customers who they thought were likely to be expecting and begin targeting promotions for nappies (diapers), cotton wool and so on. The prediction was so accurate that Target made the news by sending promotional coupons to families who did not yet realize (or who had not yet announced) they were pregnant! You can read the full story here on Forbes.

E. *Crime agencies*

The use of data mining and business intelligence is not solely reserved for corporate applications and this is shown in our final example. Beyond corporate applications, crime prevention agencies use analytics and data mining to spot trends across myriads of data helping with everything from where to deploy police manpower (where is crime most likely to happen and when?), who to search at a border crossing (based on age/type of vehicle, number/age or occupants, border crossing history) and even which intelligence to take seriously in counter terrorism activities.

VII. CONCLUSION

Data mining seeks to extract hidden knowledge from large amount of data. Data mining is the process of extracting and valuable interesting patterns from raw collection of data. Data mining can be used to uncover patterns in the data but it is often carried out only on the samples of data. This mining process will be ineffective if the samples are not a good representation of the larger body of the data. And beside this, today's competition is one of the most important challenges facing by all organizations and industries in data mining issues. That is hard to find in a particular organization or industry which has no rival to him. This paper describes various tasks; goals of data mining. Additionally this paper also discussed about the various valuable problems; future challenges and issues in field of data mining which is important to do further more effective research in this Emerging field.

.

REFERENCES

[1] Clifton, Christopher :Encyclopedia Britannica: Definition of Data Mining", Retrieved 2010-12-09.

[2] Ian H. Witten; Eibe Frank; Mark A. Hall, Data Mining: Practical Machine Learning Tools and Techniques (3rd Ed.), Elsevier, 30 January 2011.

[3] Kantardzic, Mehmed-Data Mining: Concepts, Models, Methods, and Algorithms, John Wiley&Sons, 2003.

[4] Gunnemann, S.; Kremer, H.; Seidl, T.: An extension of the PMML standard to subspace clustering models". Proceedings of the 2011 workshop on Predictive markup language modeling-PMML'11.

[5] Domenico Talia; Paolo Trunfio"How distributed data mining tasks can thrive as knowledge services". Communications of the Acm, 2010.

E. PRIYA, M.SC computer science in Kamban College of Arts and Science for women, Thiruvannamalai.

G. VINITHA, M.SC computer science in Kamban College of Arts and Science for women, Thiruvannamalai.