Data Mining:Review

Paramjit Kaur[#1], Kanwalpreet Singh Attwal[#2]

[#1]*M.Tech Scholar,Department of Computer Engineering,Punjabi University Patiala*
Punjab ,India

[#2]*Astt.Professor,Department of Computer Engineering,Punjabi University Patiala*
Punjab ,India

***Abstract*: Data mining, it is the process of extracting of previously unknown predictive information from large databases. Basically the main purpose of data mining is to find out unsuspected associations in observational data and to sum up the result in form which is clear , understandable, useful to data owner. It is a powerful new technology with great potential to help companies to focus on the most important information in their data warehouses and use it to make proactive and knowledge driven decisions. It presents knowledge in a form which is easily comprehensible to humans. This paper presents an overview of the data mining ,data mining techniques, data mining process, data mining architecture, data mining advantages/disadvantages, applications .**

Keywords: Data Mining ,KDD,Data Warehouses,Data presentation

I. Introduction

Data mining refers to extracting or "mining" knowledge from large amount of data[1].It is analogous to finding gold from the rocks or sand. Data mining is extraction of useful patterns from various data sources, e.g., databases, texts, web, image. It allows us to find the needles hidden in our haystacks of data. It refers to using a variety of techniques to identify information in the bodies of data. Extracted information can be put to use in the areas such as decision support ,prediction, forecasting and estimation. The extracted pattern must be

- valid: hold on new data.
- novel: non-obvious to the system.
- useful: should be possible to act on the them.
- understandable: humans should be able to interpret the pattern.

II. Data mining Process

Some time Data mining is also called *knowledge discovery from data* (KDD) but data mining is essential step in the process of knowledge discovery.

1. Identify goal: The decision maker needs to formulate goals that the data mining process is expected to achieve. One cannot use data mining without the idea of outcomes he/she looking for ,since the techniques to be used and data to be required are different for different goals.
2. Select target data: After understanding the application domains involved and the knowledge that's required one can select the target or relevant data on which mining to be done.
3. Cleaning and integrating data: Cleaning the data means to remove noise ,inconsistent ,irrelevant data and find out the strategies to handle missing data.if data is available on multiple sources they may be combined as per the requirement.
4. Data transformation: During this phase where data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations on data.
5. Data mining: It is vital phase of KDD, here different techniques of data mining are applied to extract data patterns.
6. Pattern evaluation: It means to identify the truly interesting patterns representing knowledge based on some interestingness measures.
7. Knowledge presentation: Where visualization and knowledge representation techniques are used to present the mined knowledge to the user.

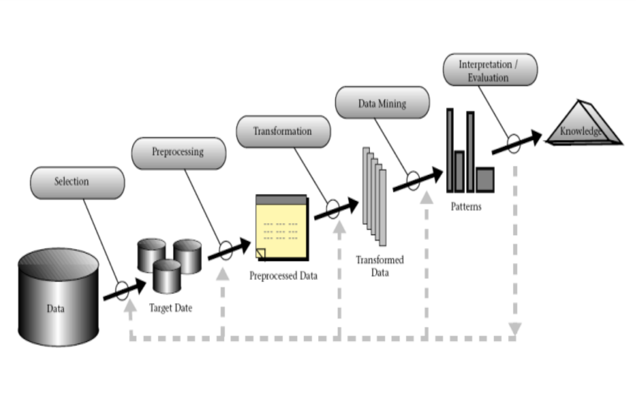

Fig. 1 Knowledge Discovery Process[2]

The steps from 2 to 4 are parts of pre-processing phase. Normally some of the above steps are being combines. For example, data cleaning and data integrating can be performed

together as a pre-processing phase to generate a data warehouse. This is an iterative process. Once the discovered knowledge is presented to the user, the evaluation measures can be enhanced, the mining can be further refined, new data can be selected or further transformed, or new data sources can be integrated, in order to get different, more appropriate results.

III. Need For Data Mining

Data mining tools predict future trends and behaviours, allowing businesses to make proactive, knowledge-driven decisions. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems. Data mining tools can answer business questions that traditionally were too time consuming to resolve. They scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations.

Traditional database system are used to answer "What" type of queries ,but data mining is used to answer "Why" type of questions.

IV. DATA MINING ARCHITECTURE

The following shows data mining architecture[3]:

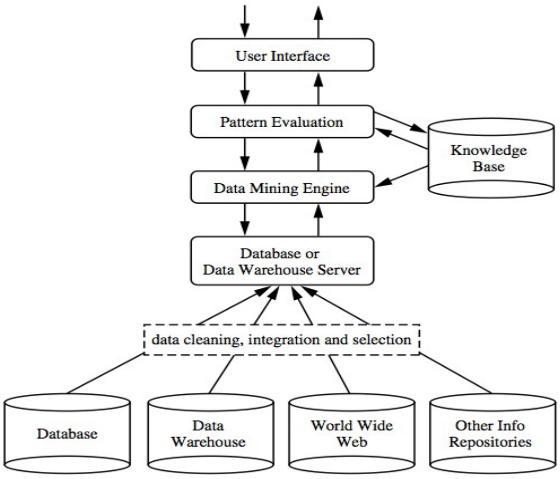

Fig. 2 Data Mining Architecture[3]

- **Data repositories:** Can be Database, data warehouse, World Wide Web, or other kinds of information repositories. Data cleaning and data integration may be performed on the data.
- **Database or data warehouse server:** Responsible for fetching the relevant data, based on the user's data mining request.
- **Knowledge base:** It is the domain knowledge that is used to guide the search or evaluate the interestingness of resulting patterns.
- **Data mining engine:** It is essential to the data mining system and ideally consists of a set of functional modules for tasks such as characterization, association and classification, prediction, cluster analysis and evolution analysis query languages (DMQL) based on mining primitives to access the data.
- **Pattern evaluation module:** Interacts with the data mining modules so as to *focus the search toward interesting patterns.*
- **User interface:** It *communicates between users and the data mining* system.Allows the user to *interact with the system by specifying* a data mining query or task. Allows the user to *browse database and data warehouse* or data structures, evaluate mined patterns, and visualize the patterns in different forms.

V. DATA MINING TECHNIQUES

Data mining techniques are broadly categorized in three types. Following are the some basic data mining techniques[4]:

A. *Predictive technique*

It uses value of some variable to predict or find the future values of other variables. For example we can use the value of climate variable to predict temperature of day.

- **Classification:**Basically classification is used to classify each item in a set of data into one of predefined set of classes or groups. Classification method makes use of mathematical techniques such as decision trees, linear programming, neural network and statistics . For example, we can apply classification in application that "given all records of employees who left the company, predict who will probably leave the company in a future period." In this case, we divide the records of employees into two groups that named "leave" and "stay".

B. *Descriptive*

It characterize the general properties of the data in the database. Example: from shopping data we conclude that children love to buy chocolate and toys than cloths.

- **Association**: Association is one of the best known data mining technique. In association, a pattern is discovered based on a relationship between items in the same transaction. That's is the reason why association technique is also known as relation technique. The association technique is used in market basket analysis to identify a set of products that customers frequently purchase together.

  Retailers are using association technique to research customer's buying habits. Based on historical sale data, retailers might find out that customers always buy crisps when they buy beers, and therefore they can put beers and crisps next to each other to save time for customer and increase sales.
- **Clustering:** Clustering is a data mining technique that makes meaningful or useful cluster of objects which have similar characteristics using automatic technique. The clustering technique defines the classes and puts objects in each class, while in the classification techniques, objects are assigned into predefined classes. E.g. In a library, there is a wide range of books in various topics available. The challenge is how to keep those books in a way that readers can take several books in a particular topic without hassle. By using clustering technique, we can keep books that have some kinds of similarities in one cluster or one shelf and label it with a

meaningful name. If readers want to grab books in that topic, they would only have to go to that shelf instead of looking for entire library.

C. *Sequential pattern analysis*

Sequential patterns analysis is one of data mining technique that seeks to discover or identify similar patterns, regular events or trends in transaction data over a business period.

In sales, with historical transaction data, businesses can identify a set of items that customers buy together a different times in a year. Then businesses can use this information to recommend customers buy it with better deals based on their purchasing frequency in the past.

VI. Data mining Applications

A. *Marketing / Retail*

Data mining helps marketing companies build models based on historical data to predict who will respond to the new marketing campaigns such as direct mail, online marketing campaign etc. Through the results, marketers will have appropriate approach to sell profitable products to targeted customers.

Data mining brings a lot of benefits to retail companies in the same way as marketing. Through market basket analysis, a store can have an appropriate production arrangement in a way that customers can buy frequent buying products together with pleasant. In addition, it also helps the retail companies offer certain discounts for particular products that will attract more customers

B. *Finance / Banking*

Data mining gives financial institutions information about loan information and credit reporting. By building a model from historical customer's data, the bank and financial institution can determine good and bad loans. In addition, data mining helps banks detect fraudulent credit card transactions to protect credit card's owner.

C. *Data Mining Applications in Transportation*

Data mining helps determine the distribution schedules among warehouses and outlets and analyze loading patterns.

D. *Data Mining Applications in Medicine*

Data mining enables to characterize patient activities to see incoming office visits. Data mining helps identify the patterns of successful medical therapies for different illnesses.

E. *Governments*

Data mining helps government agency by digging and analyzing records of financial transaction to build patterns that can detect money laundering or criminal activities.

F. *Scientific analysis*

Data mining speed up the task of data analyzing and thus make the researchers to have more time for other projects.

G. *Data Mining in Insurance*

Data mining enables to forecasts which customers will potentially purchase new policies. Data mining allows insurance companies to detect risky customers' behaviour patterns. Data mining helps detect fraudulent behaviour.

VII. Some issues with data mining[5]

A. *Privacy issues*

Data mining makes it possible to analyze routine business transactions and collect a significant amount of information about individuals buying behaviour and preferences. Such information can be leaked or sold or hacked and this information can be used in unethical way that potentially causing them a lot of troubles.

B. *Data reliability*

Clearly, data analysis can only be as good as the data that is being analyzed. A key implementation challenge is integrating contradictory or redundant data from different sources. For example, a bank may maintain credit cards accounts on several different databases. The addresses(or even the names) of a single cardholder may be different in each. Software must translate data from one system to another and select the address most recently entered.

C. *Cost*

Data mining and data warehousing tend to be self-reinforcing. The more powerful the data mining queries, the greater the utility of the information being gleaned from the data, and the greater the pressure to increase the amount of data being collected and maintained, which increases the pressure for faster, more powerful data mining queries. This increases pressure for larger, more rapid systems, which are more costly.

VIII. Future of data mining

Data mining brings a lot of benefits to businesses, society, governments as well as individual. However privacy, security and misuse of information are the big problems and need to addressed and resolved properly. In the short-term, the results of data mining will be in profitable, if ordinary, business related areas. Micro-marketing campaigns will explore new heights. Advertising will target potential customers with new accuracy. In the medium term, data mining may be as common and easy to use as e-mail. We may use these tools to find the best airfare to Australia, root out a phone number of a long-lost classmate, or find the best prices on lawn mowers. The long-term prospects are truly exciting, Imagine intelligent agents turned loose on medical research data or on sub-atomic particle data. Computers may reveal new treatments for diseases or new insights into the nature of the universe.

Acknowledgment

I express my sincere gratitude to my Guide **Er. Kanwalpreet Singh Attwal(Asst. prof)** who assisted me throughout this work. I thank him for providing me the reinforcement, confidence and most importantly the track for this work whenever I needed it.

REFERENCES

[1] R. Agrawal, T. Imielinski, and A. Swami, “Mining Association Rules Between Sets of Items in Large Databases”, Proceedings of the ACM SIGMOD Conference on Management of data, pp.207-216, May 1993.

[2] [Online Image].Available:http://liris.cnrs.fr/abstract/fayyad1996.png .

[3] Han and Kamber, “Introduction in Data mining concepts and techniques", 2nd ed.,Morgan kaufmann, C A , pp. 5–64.

[4] G.K Gupta, “Data mining and Case Studies", Prentice Hall, India.

[5] Jason Frand, “Data mining”,Anderson Graduate School of Management",[Online]Available: http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/issues.htm.