A Survey Paper of Study about Data Mining

Shruti Upadhyay[1] , Neha Patel[2], Rakesh Patel[3,] Prateek Kumar Singh[4]

Shruti Upadhyay
B.E. Student
Department of CSE -Kirodimal Institute of Technology, Raigarh (C.G.)
Email-u.shruti.su@gmail.com

Neha Patel
B.E. Student
Department of CSE -Kirodimal Institute of Technology, Raigarh (C.G.)
Email-patel.neha0106@gmail.com

Rakesh patel
Assistant Professor
HOD, IT Department, CSVTU University, Chhattisgarh
Department of IT-Kirodimal Institute of Technology, Raigarh (C.G.)
Email-rakeshpatel.kit@gmail.com

Prateek Kumar Singh
Lecturer
Department of IT-Kirodimal Institute of Technology, Raigarh (C.G.)
Email-prateek.kitraigarh@gmail.com

Abstract: ***Data mining is a process which finds useful patterns from large amount of data The process of extracting previously unknown, comprehensible and actionable information from large databases and using it to make crucial business decisions - Simoudis 1996 This data mining definition has business flavor and for business environments. However, data mining is a process that can be applied to any type of data ranging from weather forecasting, electric load prediction, product design, etc. Data mining also can be defined as the computer-aid process that digs and analyzes enormous sets of data and then extracting the knowledge or information out of it. By its simplest definition, data mining automates the detections of relevant patterns in database.***

Keywords: ***Data Mining, Knowledge Discovery Process,*** **Data Mining Life Cycle, Data Mining Techniques, Applications, Scope.**

Introduction

The development of information technology has generated large amount of databases and huge data in various areas. The research in databases and information technology has given rise to an approach to store and manipulate this precious data for further decision making. Data mining is a process of extraction of useful information and patterns from huge data. It is also called as knowledge discovery process, knowledge mining from data, knowledge extraction or data /pattern analysis. In the 21st century the human beings are used in the different technologies to adequate in the society . Each and every day the human beings are using the vast data and these data are in the different fields .It may be in the form of documents, may be graphical formats ,may be the video may be records (varying array) .As the data are available in the different formats so that the proper action to be taken. Not only to analyze these data but also take a good decision and maintain the data .As and when the customer will required the data should be retrieved from the database and make the better decision .

Knowledge Discovery Process

Knowledge discovery is a process that extracts implicit, potentially useful or previously unknown information from the data. This technique is actually we called as a data mining or Knowledge Hub or simply KDD (Knowledge Discovery Process).The important reason that attracted a great deal of attention in information technology the discovery of useful information from large collections of data industry towards field of "Data mining" is due to the perception of ***"we are data rich but information poor"***. There is huge volume of data but we hardly able to turn them in to useful information and knowledge for managerial decision making in business. To generate information it requires massive collection of data. To take complete advantage of data; the data retrieval is simply not enough, it requires a tool for automatic summarization of data, extraction of the essence of information stored, and the discovery of patterns in raw data. Data mining, popularly known as Knowledge Discovery in Database, it is the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases. It is actually the process of finding the hidden information/pattern

of the repositories**.** The knowledge discovery process is described as follows:

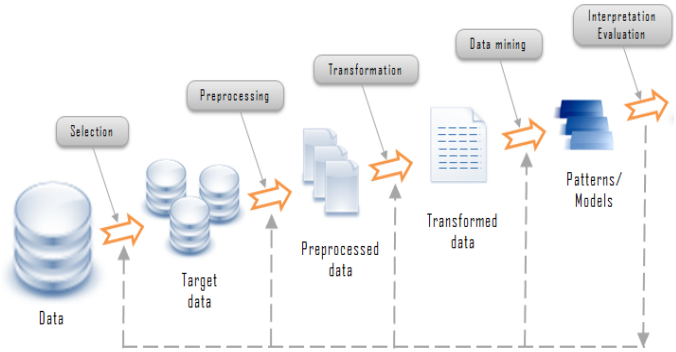

Fig.1:- Knowledge Discovery

Let's examine the knowledge discovery process in the diagram above in details:

- Data comes from variety of sources is integrated into a single data store called target data
- Data then is pre-processed and transformed into standard format.
- The data mining algorithms process the data to the output in form of patterns or rules.
- Then those patterns and rules are interpreted to new or useful knowledge or information.

The ultimate goal of knowledge discovery and data mining process is to find the patterns that are hidden among the huge sets of data and interpret them to useful knowledge and information. As described in process diagram above, data mining is a central part of knowledge discovery process.

Data Mining Life Cycle

The life cycle of a data mining project consists of six phases. The sequence of the phases is not rigid. Moving back and forth between different phases is always required. It depends on the outcome of each phase. The main phases are:

- **Business Understanding:**

This phase focuses on understanding the project objectives and requirements from a business perspective, then converting this knowledge into a data mining problem definition and a preliminary plan designed to achieve the objectives.

- **Data Understanding:**

It starts with an initial data collection, to get familiar with the data, to identify data quality problems, to discover first insights into the data or to detect interesting subsets to form hypotheses for hidden information.

- **Data Preparation:**

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

- **Modeling:**

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

- **Evaluation:**

In this stage the model is thoroughly evaluated and reviewed. The steps executed to construct the model to be certain it properly achieves the business objectives. At the end of this phase, a

decision on the use of the data mining results should be reached.

- **Deployment:**

The purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that the customer can use it. The deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process across the enterprise.

Types of Data Mining System

Data mining systems can be categorized according to various criteria the classification isas follows[3]:

- **Classification of data mining systems according to the type of data source mined:-**

In an organization a huge amount of data's are available where we need to classify these data but these are available most of times in a similar fashion. we need to classify these data according to its type(maybe audio/video ,text format etc)

- **Classification of data mining systems according to the data model:-**

There are so many number of data mining models (Relational data model, Object Model, Object Oriented data Model, Hierarchical data Model/W data model)are available and each and every model we are using the different data .According to these data model the data mining system classify the data in the model.

- **Classification of data mining systems according to the kind of knowledge discovered:-**

This classification based on the kind of knowledge discovered or data mining functionalities, such as characterization, clustering, etc. Some systems tend

to be comprehensive systems offering several data mining functionalities together.

- **Classification of data mining systems according to mining techniques used:-**

  This classification is according to the data analysis approach used such as machine learning, neural network, genetic algorithm, visualization, database oriented or data warehouse-oriented, etc.

The classification can also take into account the degree of user interaction involved in the data mining process such as query-driven systems, interactive exploratory systems, or autonomous systems. A comprehensive system would provide a wide variety of data mining techniques to fit different situations and options, and offer different degrees of user interaction.

Data Mining Techniques

There are several major data mining techniques have been developed and used in data mining projects recently including association, classification, clustering, prediction and sequential patterns etc., are used for knowledge discovery from databases.

Association

Association is one of the best known data mining technique. In association, a pattern is discovered based on a relationship of a particular item on other items in the same transaction. For example, the association technique is used in market basket analysis to identify what products that customers frequently purchase together. Based on this data businesses can have corresponding marketing campaign to sell more products to make more profit. **Applications**: market basket data analysis, cross-marketing, catalog design, loss-leader analysis, etc.

Fig.2:- Association

Types of association rules:

Different types of association rules based on

- Types of values handled
  - Boolean association rules
  - Quantitative association rules
- Levels of abstraction involved
  - Single-level association rules
  - Multilevel association rules
- Dimensions of data involved
  - Single-dimensional association rules
  - Multidimensional association rules

Classification

Goal: Provide an overview of the classification problem and introduce some of the basic algorithms. Classification is a classic data mining technique based on machine learning. Basically classification is used to classify each item in a set of data into one of predefined set of classes or groups. For Example, Teachers classify students' grades as A, B, C, D, or F. Classification method makes use of mathematical techniques such as decision trees, linear programming, neural network and statistics. In classification, we make the software that can learn how to classify the data items into groups. For example, we can apply classification in application that "given all past records of employees who left the company, predict which current employees are probably to leave in the future." In this case, we divide the employee's records into two groups that are "leave" and "stay". And then we can ask our data mining software to classify the employees into each group.

Classification Techniques

- Regression
- Distance
- Decision Trees
- Rules
- Neural Networks

Clustering

Clustering is "the process of organizing objects into groups whose members are similar in some way". A *cluster* is therefore a collection of objects which are "similar" between them and are "dissimilar" to the objects belonging to other clusters.

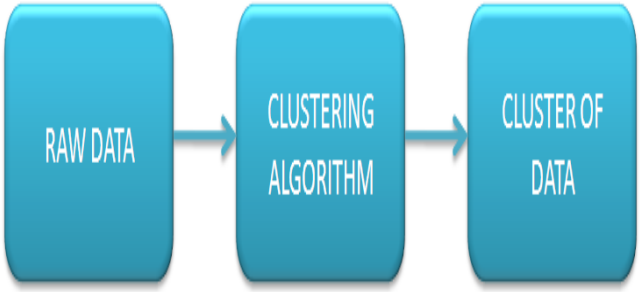

Fig.3:- Clustering

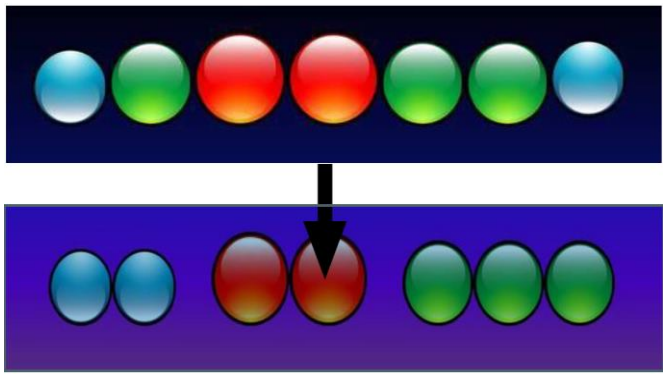

Fig.4:- Clustering

We can take library as an example. In a library, books have a wide range of topics available. The challenge is how to keep those books in a way that readers can take several books in a specific topic without irritate. By using clustering technique, we can keep books that have some kind of similarities in one cluster or one shelf and label it with a meaningful name. If readers want to grab books in a topic, he or she would only go to that shelf instead of looking the whole in the whole library.

Prediction

The prediction as it name implied is one of a data mining techniques that discovers relationship between independent variables and relationship between dependent and independent variables. In data mining independent variables are attributes already known and response variables are what we want to predict unfortunately, many real-world problems are not simply prediction For instance, sales volumes, stock prices, and product failure rates are all very difficult to predict because they may depend on complex interactions of multiple predictor variables. Therefore, more complex techniques (e.g., decision trees) may be necessary to forecast future values. For instance, prediction analysis technique can be used in sale to predict profit for the future if we consider sale is an independent variable, profit could be a dependent variable. Then based on the historical sale and profit data, we can draw a fitted regression curve that is used for profit prediction.

Sequential Patterns

Sequential patterns analysis in one of data mining technique that seeks to discover similar patterns in data transaction over a business period. The uncover patterns are used for further business analysis to recognize relationships among data.

- A sequence is an ordered list of events, denoted < e1e2… eL>.
- Each event **ei** is an unordered set of items.
- Given two sequences α=< a1a2… an> and β=< b1b2… bm>

α is called a subsequence of β, denoted as α⊆β, if there exist integers

1≤ j1< j2<…< jn≤m such that a1⊆bj1, a2⊆bj2,…, an⊆bjn

  - Example: <a(bc)dc> is a *subsequence* of <a(abc)(ac)d(cf)>
- If a sequence contains *l*items, we call it a *l*-sequence
  - Example: <a(bc)dc> is a 5-sequence.
- The support of a sequence α is the number of data sequences that contain α.

Data Mining Applications

In this section, we have focused some of the applications of data mining and its techniques are
analyzed respectively Order.

- **Data Mining Applications in Healthcare**
- **Data mining is used for market basket analysis**
- **Data mining is now used in many different areas in manufacturing engineering**
- **Data Mining Applications can be generic or domain specific**
- **Application of Data Mining techniques in CRM**
- **The Domain Specific Applications**
- **In language research**
- **In Medical Science**
- **Data Mining methods are used in the Web Education**
- **Credit Scoring**
- **The Intrusion Detection in the Network**
- **A malicious Executable is Threat**
- **Sports data Mining**
- **E-commerce is also the most prospective**
- **The Digital Library Retrieves**
- **The prediction in engineering applications**
- **The data mining is used an emerging trends in the education system in the whole world**
- **Data mining is now used in many different areas in manufacturing engineering**

The Scope of Data Mining

Data mining derives its name from the similarities between searching for valuable business information in a large database for example, finding linked products in gigabytes of

store scanner data and mining a mountain for a vein of valuable ore. Both processes require either sifting through an immense amount of material, or intelligently probing it to find exactly where the value resides. Given databases of sufficient size and quality, data mining technology can generate new business opportunities by providing these capabilities:

- **Artificial neural networks**:

  Non-linear predictive models that learn through training and resemble biological neural networks in structure.

- **Decision trees**:

  Tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset. Specific decision tree methods include Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID).

- **Genetic algorithms**:

  Optimization techniques that use process such as genetic combination, mutation, and natural selection in a design based on the concepts of evolution.

- **Nearest neighbor method**:

  A technique that classifies each record in a dataset based on a combination of the classes of the krecord(s) most similar to it in a historical dataset (where k ³ 1). Sometimes called the k-nearest neighbor technique.

- **Rule induction**:

  The extraction of useful if-then rules from data based on statistical significance. Many of these technologies have been in use for more than a decade in specialized analysis tools that work with relatively small volumes of data. These capabilities are now evolving to integrate directly with industry-standard data warehouse and OLAP platforms.

Conclusion

Data mining is a "decision support" process in which we search for patterns of information in data. In other words, Data mining has importance regarding finding the patterns, forecasting, discovery of knowledge etc in different business domains. Data mining techniques such as classification, clustering, prediction, association and sequential patterns etc it helps in finding the patterns to decide upon the future trends in businesses to grow. Data mining has wide application field almost in every industry where the data is generated that's why data mining is considered one of the most important frontiers in database and information systems and one of the most promising interdisciplinary developments in Information Technology also.

References

[1]. Dr. Lokanatha C. Reddy, A Review on Data mining from Past to the Future, *International Journal of Computer Applications (0975 – 8887) Volume 15– No.7, February 2011*
[2]. Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth,From Data Mining to Knowledge Discovery in Databases, AI Magazine Volume 17 Number 3 (1996)
[3]. http://www.slideshare.net/Annie05/sequential-pattern-discovery-presentation
[4]. http://dataminingtools.net/wiki/introduction_to_data_mining.php
[5]. http://www.dataminingtechniques.net
[6]. http://www.slideshare.net/huongcokho/data-mining-concepts **Books:[7]**. Arun K. Pujari, *Data Mining Techniques* [8]. Jiawei Han, Micheline Kamber, Data Mining: Concepts and Techniques
[7] Introduction to Data Mining and Knowledge Discovery, Third Edition ISBN: 1-892095-02-5, Two
Crows Corporation, 10500 Falls Road, Potomac, MD 20854 (U.S.A.), 1999.
[8] Larose, D. T., "Discovering Knowledge in Data: An Introduction to Data Mining", ISBN 0-471-66657-2,
ohn Wiley & Sons, Inc, 2005.
[9] Dunham, M. H., Sridhar S., "Data Mining: Introductory and Advanced Topics", Pearson Education,
New Delhi, ISBN: 81-7758-785-4, 1st Edition, 2006.
[10] Fayyad, U., Piatetsky-Shapiro, G., and Smyth P., "From Data Mining to Knowledge Discovery in Databases," AI Magazine, American Association for Artificial Intelligence, 1996.

Author Profile

Shruti Upadhyay B.E Student
Department of CSE -Kirodimal Institute of Technology, Raigarh (C.G.)
Email-u.shruti.su@gmail.com

Neha Patel B.E Student
Department of CSE -Kirodimal Institute of Technology, Raigarh (C.G.)
Email-patel.neha0106 @gmail.com

Rakesh Patel Working as an HOD & Assistant Professor,
Department of IT-Kirodimal Institute of Technology, Raigarh (C.G.)
Email-rakeshpatel.kit@gmail.com

Prateek Kumar Singh Working as a Lecturer
Department of IT-Kirodimal Institute of Technology, Raigarh (C.G.)
Email-prateek.kitraigarh@gmail.com