ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

An Overview of Knowledge Discovery Database and Data mining Techniques

Priyadharsini.C[1], Dr. Antony Selvadoss Thanamani[2]

M.Phil, Department of Computer Science, NGM College, Pollachi, Coimbatore, Tamilnadu, India[1]

HOD, Department of Computer Science, NGM College, Pollachi, Coimbatore, Tamilnadu, India[2]

ABSTRACT: Data mining is used to find or generate new useful information's from large amount of data base. It is a process of extracting previously unknown and processable information from large databases and using it to make important business decisions. Several emerging applications in information providing services, such as data warehousing and on-line services over the Internet, also call for various data mining and knowledge discovery techniques to understand user behavior better, to improve the service provided, and to increase the business opportunities Of an overview of knowledge discovery database and data mining.

KEY WORDS: KDD–Knowledge Discovery in Data base, data mining process.

I. INTRODUCTION

In real-time information technology has generated and used large amount of databases and stored 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 [1]. Data mining is a process of extracting previously unknown and process able information from large databases and using it to make important business decisions. It is also called as knowledge discovery process, Data mining should be used exclusively for the discovery stage of the KDD process

II. KNOWLEDGE DISCOVERY DATABASE

Data mining is the core part of the knowledge discovery process. In this, process may consist of the following steps Data selection, Data cleaning, Data transformation, pattern searching (data mining), finding presentation, finding interpretation and finding evaluation. The data mining and KDD often used interchangeably because Data mining is the key part of 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 data bases. 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, sub sampling, and transformations of that database.

2.1 The KDD Process

The knowledge discovery process is iterative and interactive, consisting of nine steps [3].Note that the process is iterative at each step, meaning that moving back to previous steps may be required .So it is required to understand the process and the different needs and possibilities in each step. A typical knowledge discovery process is shown in figure 1, and the process is elaborated in each step.

- **Developing an understanding of the application domain**
- **Selecting and creating a data set on which discovery will be performed.**
- **Preprocessing and cleansing.**

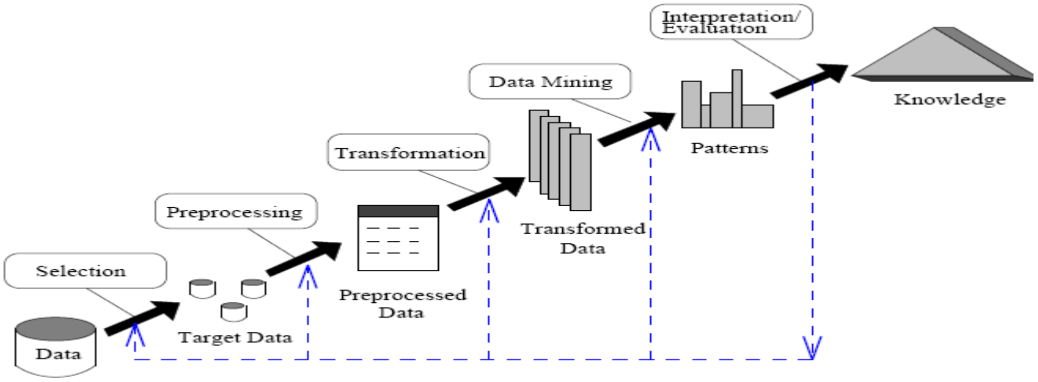

Figure 1: A typical knowledge discovery process

- **Choosing the appropriate Data Mining task.**
- **Choosing the Data Mining algorithm.**
- **Employing the Data Mining algorithm.**
- **Evaluation.**
- **Using the discovered knowledge.**

The terms *knowledge discovery* and *data mining* are distinct. **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.

III. DATA MINING

Data mining is the process of discovering actionable information from large sets of data [4]. Data mining uses mathematical analysis to derive patterns and trends that exist in data. These patterns and trends can be collected and defined as a data mining model. Mining models can be applied to specific scenarios, such as

Forecasting: Estimating sales, predicting server loads or server downtime

Risk and probability: Choosing the best customers for targeted mailings, determining the probable break-even point for risk scenarios, assigning probabilities to diagnoses or other outcomes

Recommendations: Determining which products are likely to be sold together, generating recommendations

Finding sequences: Analyzing customer selections in a shopping cart, predicting next likely events

Grouping: Separating customers or events into cluster of related items, analyzing and predicting affinities.

Building a mining model is part of a larger process that includes everything from asking questions about the data and creating a model to answer those questions, to deploying the model into a working environment [6]. This process can be defined by using the following basic steps:

- Defining the Problem
- Preparing Data
- Exploring Data
- Building Models
- Exploring and Validating Models

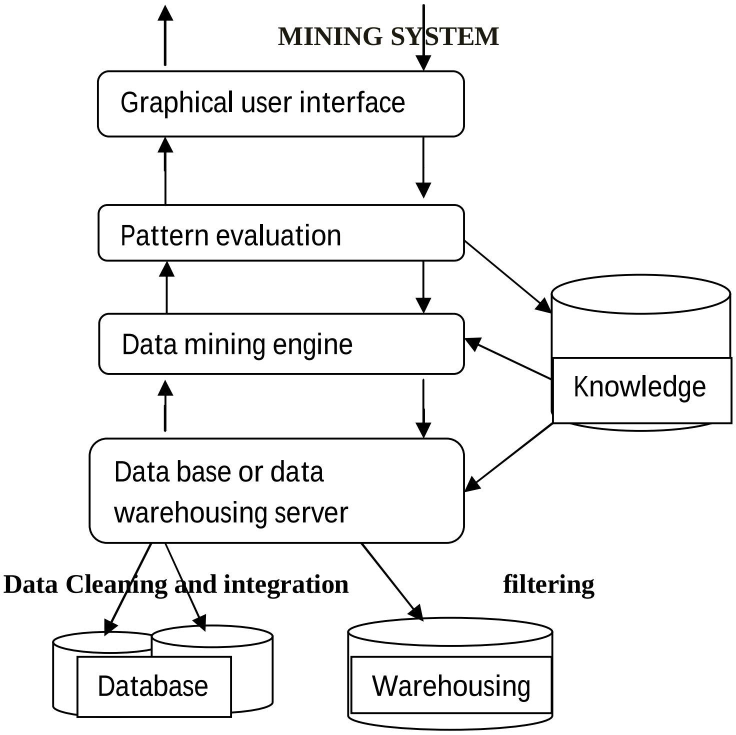

Figure 2: Data mining process

IV. 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[2].

4.1 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.

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

4.2 Classification

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.

Classification Techniques

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

4.3 Clustering

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

4.4 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.

4.5 Sequential Patterns

Sequential patterns analysis in one of data mining technique that seeks to discover similar patterns in data transaction over a business period[5].The uncover patterns are used for further business analysis to recognize relationships among data. A sequence is an ordered list of events, denoted < e_1e_2… eL>.

V. ASPECTS OF DATAMINING

5.1 Data Integration: First of all the data are collected and integrated from all the different sources.

5.2 Data Selection: We may not all the data we have collected in the first step. In this step we select only those data which we think useful for data mining.

5.3 Data Cleaning: The data we have collected are not clean and may contain errors, missing values, noisy or inconsistent data. Therefore we need to apply different techniques to get rid of such anomalies.

5.4 Data Transformation: The data even after cleaning are not ready for mining as we need to transform them into forms appropriate for mining. The techniques used to accomplish this are smoothing, aggregation, normalization etc.

5.5 Data Mining: Now we are ready to apply data mining techniques on the data to discover the interesting patterns. The techniques like clustering and association analysis are among the many different techniques used for data mining.

5.6 Pattern Evaluation and Knowledge Presentation: This step involves visualization, transformation, removing redundant patterns etc from the patterns we generated.

5.7 Decisions / Use of Discovered Knowledge: This step helps user to make use of the knowledge acquired to take better decisions.

There are various steps that are involved in mining data as shown in the picture.

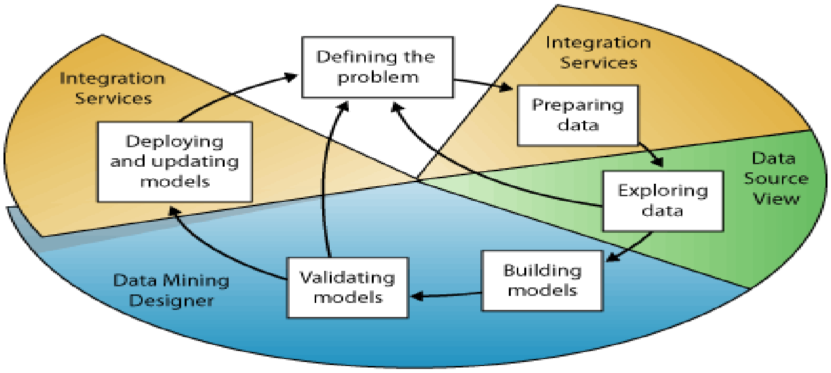

Figure 3: Mining a data

Issues in data mining

A number of issues that need to be addressed by any serious data mining package

- Uncertainty Handling
- Dealing with Missing Values
- Dealing with Noisy data
- Efficiency of algorithms
- Constraining Knowledge Discovered to only Useful or Interesting Knowledge
- Incorporating Domain Knowledge

- Size and Complexity of Data
- Data Selection
- Understandability of Discovered Knowledge: Consistency between Data and Discovered Knowledge

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.

VI. DIFFERENT LEVELS OF ANALYSIS

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

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

6.3 Decision trees: Tree-shaped structures that represent sets of decisions. So these decisions generate rules for the classification of a dataset. The Specific decision tree methods include Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID). The CART and CHAID are decision tree techniques used for classification of a dataset.

6.4 Nearest neighbor method: A technique that classifies 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.

6.5 Rule induction: The extraction of useful if-then rules from data based on statistical significance.

6.6 Data visualization: The visual interpretation of complex relationships in multidimensional data. Graphics tools are used to illustrate data relationships.

VII. DATAMINING ALGORITHMS

7.1 C4.5 and beyond: Systems that construct classifiers are one of the commonly used tools in data mining [8]. Such Systems take as input a collection of cases, each belonging to one of a small number of Classes and described by its values for a fixed set of attributes, and output a classifier that can accurately predict the class to which a new case belongs.

7.2 The *k*-means algorithm: The k-means algorithm is a simple iterative method to partition a given dataset into a user specified Number of clusters, *k*. this algorithm has been discovered by several researchers across different disciplines. The algorithm operates on a set of _*d*-dimensional vectors, $D = \{\mathbf{x}_i | \; i = 1 \ldots N\}$, where $\mathbf{x}_i$□ _*d* denotes the i^{th} data point. The algorithm is initialized by picking *k* points in _*d* as the initial *k* cluster representatives or “centroids”.

Step 1: *Data Assignment*. Each data point is assigned to its *closest* centroid, with ties Broken arbitrarily. This results in a partitioning of the data.

Step 2: *Relocation of "means"*. Each cluster representative is relocated to the center (Mean) of all data points assigned to it.

7.3 Support vector machines: The machine learning applications, Support Vector Machines (SVM) are considered A must try—it offers one of the most robust and accurate methods among all well-known Algorithms. Therefore it has a sound theoretical foundation, requires only a dozen examples for training, and is insensitive to the number of dimensions.

7.4 The Apriori algorithm: One of the most popular data mining approaches is to find frequent item sets from a transaction Dataset and derive association rules. Therefore Finding frequent item sets is not trivial because of its combinatorial explosion.

7.5 The EM algorithm; Finite mixture distributions provide a flexible and mathematical-based approach to the modeling and clustering of data observed on random phenomena. Therefore we focus here on the use of Normal mixture models, which can be used to cluster continuous data and to estimate the Underlying density function.

7.6 PageRank: The most popular Page Ranking algorithm issued by Google search engine. The algorithm assigns ranks for each hyperlink on the web. Based on this algorithm, they built the search engine Google, which has been a huge success. Nowadays every search engine has its own hyperlink based ranking method.

7.7 AdaBoost: *Ensemble learning* deals with methods which employ multiple learners to solve a problem. This generalization ability of an ensemble is usually significantly better than that of a single learner, so ensemble methods are very attractive.

7.8 *k*NN: *k*-nearest neighbor classification: One of the simplest and rather trivial classifiers is the Rote classifier, which memorizes the entire training data and performs classification only if the attributes of the test object match one of the training examples exactly.

VIII. CONCLUSIONS

Data mining has the most important and promising features of interdisciplinary developments in Information technology. This review would help the researchers to focus on the various issues of data mining. An overview of knowledge discovery database and data mining techniques has provided an extensive study on data mining techniques. Data mining is useful for both public and private sectors for finding patterns, forecasting, discovering knowledge in different domains such as finance, marketing, banking, insurance, health care and retailing. Data mining is commonly used in these domains to increase the sales, to reduce the cost and enhance research to reduce costs & enhance research

REFERENCES

[1] Mrs. Bharati M. Ramageri, "Data Mining Techniques and Applications," *Indian Journal of Computer Science and Engineering*, Vol. 1 No. 4, pp. 301-305[2] HemlataSahu, ShaliniShrma and SeemaGondhalakar, "A Brief Overview on Data Mining Survey," *International Journal of Computer Technology and Electronics Engineering (IJCTEE).*, Vol.1, Issue 3, pp.114-121

[2] Kalyani M Raval, "Data Mining Techniques," *International Journal of Advanced Research in Computer Science and Software Engineering*,Vol. 2 Issue 10,pp.439-442[4] SangeetaGoele, NishaChanana, "Data Mining Trend In Past, Current And Future," *International Journal of Computing & Business Research*, in *Proc. I-Society 2012*, 2012.

[3] Mr. S. P. Deshpande and Dr. V. M. Thakare, "Data Mining System and Applications: A Review," International Journal of Distributed and Parallel systems (IJDPS) Vol.1, No.1, September 2010.

[4] Kalyani M Raval, "Data Mining Techniques," *International Journal of Advanced Research in Computer Science and Software Engineering*,Vol. 2 Issue 10,pp.439-442.

[5] SangeetaGoele, NishaChanana, "Data Mining Trend In Past, Current And Future," *International Journal of Computing & Business Research*, in *Proc. I-Society 2012*, 2012

[6] Mr. S. P. Deshpande and Dr. V. M. Thakare, "Data Mining System and Applications: A Review," International Journal of Distributed and Parallel systems (IJDPS) Vol.1, No.1, September 2010, pp.32-44

[7] Y. Ramamohan, K. Vasantharao, C. KalyanaChakravarti, and A.S.K.Ratnam, "A Study of Data Mining Tools in Knowledge Discovery Process," *International Journal of Soft Computing and Engineering (IJSCE)*, Vol. 2, Issue-3, July 2012, pp.191-1994