IJCSN - International Journal of Computer Science and Network, Volume 3, Issue 6, December 2014
ISSN (Online): 2277-5420 **www.IJCSN.org**
Impact Factor: 0.274

Knowledge Discovery in Database

Kiran S.Gaikwad

Department of IT
Government Polytechnic College.

Abstract - The main aim of this paper is to expound about the knowledge discovery and data mining.Data warehouse brings the information from multiple sources so as to provide a consistent database source for decision support queries and off-load decision support applications from the on-line transaction system. Here, data is available but not information and not the right information at the right time. Data mining is extracting interest information or patterns from data in large databases.For processing the data there are many traditional and statistical methods of data analyses and spreadsheets are used to obtained informative reports from data but they can't give the knowledge from data. Closet is an efficient algorithm which is scalable on large databases in order to get the important knowledge hidden inside the data. In this process a set of association rules are discovered at multiple levels of abstraction from the relevant sets of data in a database.

Keywords – *Discovery in database.*

1. Introduction

Many business and government transactions related to activities and decisions generates tremendous amounts of data stored in databases, data warehouses and other information repositories by large and simple transaction i.e. tax returns, telephone calls, business trips, performance tests and product warranty registration which are being handled through computer. So, there is an urgent need for a new generation of computational theories and tools to assist humans in extracting useful information (knowledge) from the rapidly growing volumes of digital data. A data warehouse is a subject-oriented, integrated, time-variant, and nonvolatile collection of data in support of management's decision-making process. The warehouse usually resides on its own server and is separate from the transaction-processing or "run-the-business" systems. We are drowning in data, but starving for knowledge! So, we extract interesting knowledge from data in large databases which is known as data mining.

2. Knowledge Discovery in Databases

Knowledge Discovery in Databases (KDD) is the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data A KDD process includes data cleaning, data integration, data selection, transformation, data mining, pattern evaluation, and knowledge presentation. Knowledge Discovery in Databases (KDD) has made great progress in commercial, industrial, administrative and other applications. The emerging technology KDD having a multi-step process which uses Data Mining Methods (Algorithms)to extract (Identify) what is hidden knowledge in the data according to specifications of measures The basic problem addressed by the KDD process is one of mapping glow level data into other forms that might be more compact, more abstract ,or more useful.

Data mining is a step in the KDD process that consists of applying data analysis and discovering algorithms. Understanding data mining and model induction at this component level clarifies the behavior of any data mining algorithm and makes it easier for the user to understand its overall contribution and applicability to the KDD process.Data mining, also known as knowledge-discovery in databases (KDD), is the practice of automatically searching large stores of data for patterns. To do this, data mining uses computational techniques from statistics and pattern recognition. Problem Analysis is based on manual procedure. The main function is to understanding Application domain and requirements of user related to developing prior knowledge for domain.

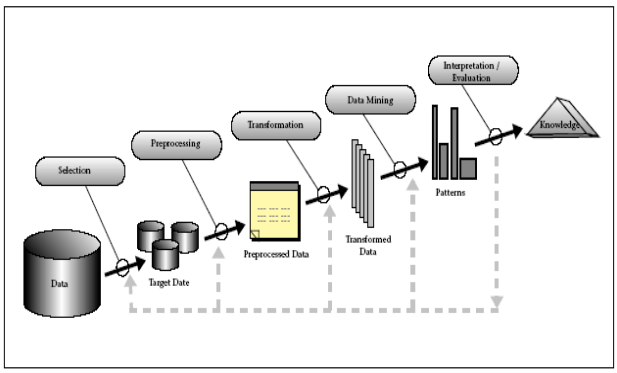

Figure 1. An Overview of the Steps That Compose the KDD Process.

Selection of Target data is for creating target data set and Selecting a data set or its subset on which discovery is to be performed by automatic way. Data Processing which is the third step of KDD process involves removing noise/ handling missing data based on automatic program. Transformation of Data is a procedure which is made manually where data reduction and projection are made and finding useful fields/features/attributes of data according to goal of the problem. Data Mining selects data mining goal and chooses a method according to task and knowledge .It analyzes and verifies knowledge .It is based on automatic manner. Output Analysis and Review evaluates the knowledge and transforms knowledge.

2.1 Data Mining

Data Mining selects data mining goal and chooses a method (algorithms) according to task and knowledge .It analyzes and verifies knowledge.

2.2 Interpretation

Output Analysis and Review evaluates the knowledge and transforms knowledge.

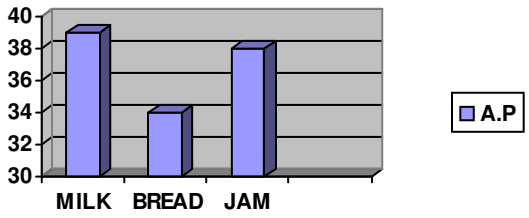

Figure 2 Chart

STEPS THAT COMPOSE THE KDD PROCESS WITH EXAMPLE

Table 1

Sr. no.	State	Milk (Lts)	Bread	Jam (Tins)	Cost (Rs)
1	A.P	5	7	---	400
2	T.N	3	6	2	QUE
3	A.P	----	7	9	1000
4	A.P	7	3	---	500
5	K.A	6	8	4	300
6	M.P	9	---	7	900
7	A.P	8	6	5	700
8	A.P	5	----	9	1500
9	A.P	6	7	4	ABC
10	T.N	----	5	8	500

Selection: In this step selection of target data is for creating target data set and Selecting a data set or its subset on which discovery is to be performed by automatic way.

Table 2

Sr. no.	State	Milk (Lts)	Bread (Packets)	Jam (Tins)	Cost (Rs)
1	A.P	5	7	---	400
3	A.P	----	7	9	1000
4	A.P	7	3	---	500
7	A.P	8	6	5	700
8	A.P	5	----	9	1500
9	A.P	6	7	4	ABC

Preprocessing: Data Processing is the process which involves removing noise/ handling missing data based on automatic program.

Table 3

Sr. no.	State	Milk (Lts)	Bread (Packets)	Jam (Tins)	Cost (Rs)
1	A.P	5	7	9	400
3	A.P	8	7	9	1000
4	A.P	7	3	2	500
7	A.P	8	6	5	700
8	A.P	5	4	9	1500
9	A.P	6	7	4	800

TRANSFORMATION - It is a procedure which is made manually where data reduction and projection are made and finding useful fields/features/attributes of data according to goal of the mining. Here we use Regression technique.

Table 4

Sr. no.	State	Milk (Lts)	Bread (Packets)	Jam (Tins)	Cost (Rs)
1	A.P	5	7	9	400
3	A.P	8	7	9	1000
4	A.P	7	3	2	500
7	A.P	8	6	5	700
8	A.P	5	4	9	1500
9	A.P	6	7	4	800
TOTAL	A.P	39	34	38	5900

Hence Knowledge Is Obtained

Data Mining Techniques: Data Mining has three major components Clustering or Classification, Association Rules and Sequence Analysis.

3. Classification

The clustering techniques analyze a set of data and generate a set of grouping rules that can be used to classify future data. The mining tool automatically identifies the clusters, by studying the pattern in the training data. Once the clusters are generated, classification can be used to identify, to which particular cluster, an input belongs. For example, one may classify diseases and provide the symptoms, which describe each class or subclass.

4. Association

An association rule is a rule that implies certain association relationships among a set of objects in a database. In this process we discover a set of association rules at multiple levels of abstraction from the relevant sets of data in a database. For example, one may discover a set of symptoms often occurring together with certain kinds of diseases and further study the reasons behind them.

5. Sequential Analysis

This deals with data that appear in separate transactions (as opposed to data that appearing the same transaction in the case of association) e.g. if a shopper buys item A in the first week of the month, and then he buys item B in the second week etc.

6. Mining for Association Rules

General Procedure is firstly association rules use apriority to generate frequent item sets of different sizes at each iteration divide each frequent item set X into two parts LHS and RHS. This represents a rule of the form LHS RHS The confidence of such a rule is support(X)/support (LHS). Discard all rules whose confidence is less than minconf.

7. Closet

An Efficient Algorithm for Mining Frequent Closed Item Sets Association mining may often derive an undesirably large set of frequent item sets and association rules which reduces not only efficiency but also the effectiveness of mining since users has to sift through a large number of mining rules to find useful ones. The requirement of mining the complete set of association rules leads to two problems: Firstly there may exist a large number of frequent item sets in a transaction database, especially when the support threshold is low. Secondly when there exist a huge number of association rules. Therefore it is hard for the user to comprehend and manipulate a huge number of rules. Recent studies has proposed an interesting alternative mining frequent closed set items and their corresponding rules which and has the same power as association mining but substantially reduces the number of rules to be present and increase both efficiency and effectiveness of mining. Instead of mining the complete set of frequent item sets and their associations, association mining only need to find frequent closed item sets and their corresponding rules.

A closed item set is frequent if its support passes the given support threshold. Then the closed itemset is called Frequent closed item. Closet is an efficient algorithm on closed items. It is scalable on large databases, and faster than previously proposed methods. Three techniques are developed for this purpose .First is doing the framework of recently developed efficient frequent pattern mining method. Second is strategies are devised to reduce the search space dramatically and identify the frequent closet items quickly .Third is a partion-based projection mechanism is established to make the mining efficient and scalable for large databases. Frequent closed item set mining problem is the way to find the complete set of frequent closed item sets efficiently from larger databases. Mining the frequent closed item sets with projected database can be done as: first finding frequent items then divide the search space and finally find subsets of frequent closed items.

8. Application Areas of Data Mining

Some of the potential areas are i.e. Banking, Finance, and Survey's related to Customer satisfaction, Market, Buying behavior, Customer characteristics, Economic, Direct Marketing.

9. Conclusion

Data mining is an iterative process of extracting interesting knowledge from data in large databases. Where knowledge could be rules, patterns, regularities, relationships, constraints etc. Whereas KDD is the overall process of finding and interpreting knowledge from data it helps the workers in their everyday business activity and improve their productivity and also helps for knowledge workers like executives, managers, analysts to make faster and better decisions .So knowledge should be valid and potentially useful and then the hidden information in the database will be useful.

References

[1] Anahory S, & Dennis M, "Data Warehousing in the Real World" Kimball R, "Data Warehouse Toolkit",John Wiley.

Author: KIRAN S. GAIKWAD,
Working as IT Lecturer,
Government Polytechnic College.
IT Experience: 4years.
B.Tech (Computer Science)