Market basket analysis using apriori algorithm in data mining

SWATI MAHESH JOSHI

Lecturer, Department of Information Technology, V.P.M.'s Polytechnic, Thane, Maharashtra, India.

---***---

Abstract – *The main objective of mining Data is to transform it into knowledge. This paper explores the basics of data mining with its architecture in brief. It also focuses on the patterns which can be mined so as to convert data in to knowledge. The topic elaborates one of the techniques to analyze the data from large amount of data sets. The analysis explained here is market basket analysis which is basically used by marketer to improve the performance of their business. It can be done by analyzing the available data in such way that frequent item set can be found and can be analyzed to define an association rule. One of the algorithm which helps in finding association rule for frequent item set is Apriori algorithm. It is explained at the end with a case study in this paper.*
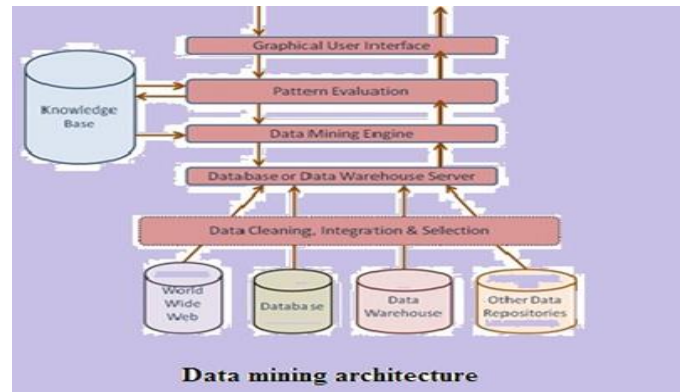
Key Words: KDD, Data mining, Data set, Item Set, Minimum Support, Apriori algorithm, association rule

1. What is Data Mining?

The process of discovering knowledge in data and application of data mining methods is nothing but the term knowledge discovery in databases or KDD.

Data mining is one o0f the steps in KDD process. It applies the appropriate algorithm based on the goal of the KDD process for identifying patterns from data.

Data Mining is a process of identifying patterns and establishing relationships. It is the process of analyzing large datasets stored in data warehouses for useful information which makes use of artificial intelligence techniques, neural networks and advanced statistical tools such as cluster analysis to reveal trends, patterns and relationships, which otherwise may be undetected. Data mining has been used in numerous areas, which include both private and public sectors.

The process of discovering knowledge in data and application of data mining methods is nothing but the term knowledge discovery in databases or KDD.

1.1 Data mining architecture

The major components of any data mining system are data source, data warehouse server, data mining engine, pattern evaluation module, graphical user interface and knowledge base.

Fig -1: Data mining architecture.

1. Database, data warehouse, or other information repository:
These are the actual information repositories.

2. Database or data warehouse server:
Here the data is fetched as per the user's requirement which is need for data mining task.

3. Knowledge base:
The collection of discovered data called as knowledge base.

4. Data mining engine :
It performs the data mining task such as characterization, association, classification, cluster analysis etc.

5. Pattern evaluation module:
It is integrated with the mining module and it helps in searching only the interesting patterns.

6. Graphical user interface:
Through this interface communication between user and the data mining system happens and it allow users to browse the data from database or data warehouse schemas.

1.2 Different kinds of patterns to be mined:

Association: The association rule should be read as "Given that somebody has brought the items in set1 ,they are likely to also buy the items in set2.

The main applications of association rule mining: Market Basket data analysis. It is to analyze the association of purchased items in a single basket or single purchase.

Classification: It maps the data into predefined groups or classes and searches for new patterns. It is a data mining

technique used for systematic placement of group membership of data.

The main application areas are to form decision trees and in neural network.

Regression: It is used to predict for individuals on the basis of information gained from a previous sample of similar individuals.

Clustering: Clustering is a technique used in data mining to place data elements into related groups without advance knowledge of group definitions.

Forecasting: It is nothing but discovering patterns in class that can lead to reasonable predictions about the future.

Outlier analysis: Outlier is the data object that does not comply with the general behavior of the data. It is useful in fraud detection, rare events analysis.

Based on the above functionalities Data mining tasks can be categorized into two categories as 1) Descriptive mining and 2) Predictive mining.

One of the descriptive mining techniques is mining of frequent patterns.

2. Market Basket Analysis

Market Basket analysis is a modeling technique which is also called as affinity analysis, it helps identifying which items are likely to be purchased together.

The market basket problem assumes we have some large number of items. Customer buys the subset of items as per his needs and marketer gets the information that which things customer has taken together. So the marketer uses this information to put the items on different position.

For example: If someone buys a packet of milk also tends to buy bread at the same time. So milk=>bread.

Market basket analysis is used in deciding the location of items inside a store. If a customer buys a packet of bread he is more likely to buy a packet of butter too. Keeping the bread and butter next to each other in a store would result in customers getting tempted to buy one item with other.

Identification of items purchases or events occurring in a sequence can be the area of interest to direct marketers. This approach I termed as predictive market basket analysis.

Identifying the frequent item sets is one of the major processes during market basket analysis. One of the algorithms used to find association rule for frequent item sets.

3. Apriori algorithm for finding frequency item set:

The Apriori algorithm analyses a data set to determine which combinations of items occur together frequently. It is at the core of various algorithms for data mining problems. The best known problem is finding the association rules that hold in a basket-item relation.

Basic idea behind this algorithm is

- An item set can only be a large item set if all its subsets are large item sets.
- The sets of items that have minimum support can be considered.
- Association rules can be generated from frequent item sets.

3.1 Case study:

Consider the following data on which apriori algorithm can be applied to find the association rule for frequent item set.

Table -1: Sample Data D

Transaction ID	Items
100	Bread, Milk,Butter
200	Bread,Butter
300	Bread ,Eggs
400	Milk,Biscuits,Chips

Step 1: Scan data D for count of each candidate

Item set	Support
{Bread}	3
{Milk}	2
{Butter}	2
{Biscuits}	1
{Chips}	1
{Eggs}	1

Step 2: Compare candidate support count with minimum support (50%)

Item set	Support
{Bread}	3
{Milk}	2
{Butter]	2

Step 3: Generate candidate from above table & its support

Item set	Support
{Bread, Milk}	1
{Bread, Butter}	2
{Milk, Butter}	1

Step 4: Compare candidate with minimum support

Item set	Support
{Bread, Butter}	2

Step 5: So the data contains frequent item set {Bread, Butter}
Therefore the association rule can be set as Bread->Butter or Butter->Bread

It can be decided then that these are the frequent items purchased by the customers in a combination and accordingly the marketing strategy can be decided for the store.

4. CONCLUSIONS

Data mining is the major process required in Knowledge discovery database. The databases in general can contain information which is undiscovered as useful information. Market Basket analysis is one of the required analysis and is used in many areas like credit card transactions, for analyzing phone calling patterns, in many of the financial services companies. For this analysis the major process is to find out frequent item sets. Apriori algorithm is the best suitable algorithm to find the frequent item set with heir association rules.

REFERENCES

[1]   Han, Kamber, "Data Mining Concepts and Techniques", Morgan Kaufmann

[2]   Michael Berry and Gordon Linoff "Data Mining Techniques", 2nd Edition Wiley Publications.