International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 3 Issue 8, August - 2014

A Survey on Association Rule Mining Algorithms Preformance Analysis

Suchismita Mishra[1]

ITER, Bhubaneswar

Pranati Mishra[2]

CET, Bhubaneswar

***Abstract*--Association rule mining comes under data mining, which is a phase in knowledge discover in database (KDD). A typical example of association rule mining is market basket analysis. Market basket analysis provides the knowledge about data pattern from itemsets present in the transaction. To extract the knowledge from the dataset the popular Apriori algorithm is used.The knowledge is used in decision making. Many algorithms are developed to infer the co-occurrence of items from an itemset. In this paper we have discussed some algorithms and their performance.**

Keywords- ***Association rule mining, Apriori algorithm, KDD, market basket analysis***

I. INTRODUCTION

The association rules show the relation between items in an itemset. To extract or find out the association from an itemset we need to analyze the itemset. The itemset is nothing but the collection of items during a single purchase. It is also called as market basket data. It contains a set of items to be purchased or already purchased. The market basket data is collected from operational database. The operational database is a collection of transactions made each day.

In the classical model of association rule mining implements the support and confidence measures. But in practice many factors affect a transaction. Also the itemset in each transaction contains different combinations of items. So it is very difficult to find out the association from the data repository. The transactions are stored in operational database as records. The data repository is known as data warehouse , which contains various data in a large amount. It is not possible to process all the data at time. So before processing the data need to be segregated into data marts. Then the processing is started for extracting the association among data and this is called as data mining. Data mining is mainly applied to determine the pattern of data, changes in data pattern from previous records.

The market basket analysis is an typical example of data association rule mining. Association rule mining is one of the data mining techniques. Data mining is one stage of the knowledge discovery process. In the present scenario the collected large amount of data resulted in formation of a data mountain. At the same time it become very difficult to extract only the valuable information from such large data. To resolve such problem different techniques are applied.

The knowledge discovery process consists of an iterative sequence of the following steps:

i) Data Cleansing- to remove noise and inconsistent data.
ii) Data integration- where multiple data sources may be combined.
iii) Data Selection- where data relevant to the analysis task are retrieved from the database.
iv) Data transformation- where data are transformed or consolidated into forms appropriate for mining by performing summery or aggregation operations for instance.
v) Data mining- an essential process where intelligent methods are applied in order to extract data patterns.
vi) Pattern evolution- to identify the truly interesting patterns representing knowledge based on some interestingness measures.
vii) Knowledge presentation- where visualization and knowledge representation techniques are used to present the mined knowledge to the user.

II. ASSOCIATION RULE

Let I={I1,I2,I3,…..In} is defined as the set of items, where I is the itemset and In is the contains of the itemset or the items.
Let D is defined as the set of transactions over the items and each transaction is a set of items. A transaction T contains X, a set of some items in I, if X C T. An association rule is an implication of the form- X ->Y, where XCI and X^Y=Φ.

Association rule X->Y has confidence c if c% of transactions in D that contain X also contain Y. Association rule X->Y has support s,if s% of transactions in D contain X U Y.

The aim of mining association is to generate all association rules that have support and confidence greater than the user-specified minimum support and minimum confidence in a given set of transactions D.

III. LITERATURE SURVEY

The objective of the study relating to association rule mining is to mine the association rules with a more efficient manner i.e. with less time and less memory consumption. And provide this efficient rule to modify the existing marketing strategy [1]. ChintaSomeswaraRao et al. used the retail market database for this study (consumer purchase behavior in computer and related asseceries database). The proposed system is designed in java swing.
ChintaSomeswaraRaoet.al. shows that frequent item sets can be generated from a data set without generation of candidate set. So the volume of invalid data reduced. Only the relevant data set is obtained.

ChintaSomeswaraRao et al. describes the application of Boolean algorithm with association rule.The main advantage is it generates frequent item sets without generating candidate sets from the given data set.

In the study by ChanchalYadav, Shuliang Wang, Manoj Kumar [2] ,the proposed model is discard the pruning of the Apriori because it increases impurities. It uses partitioning in the preprocessing stage. It generates some partitioned values basing on the appropriateness of the attributes. Then mining is performed and the association rules will be generated.
To manage memory size the whole data set is divided into different horizontal partitions. To avoid data duplication in memory at the time of insertion the algorithm checks weather the item exists in memory or not. If present, then increase item count by one. Otherwise item is inserted and count is increased by one. Finally frequent and infrequent items are calculated.

FP-growth algorithm is another efficient algorithm for mining frequent patterns without producing candidate sets. Candidate set generation is time consuming which leads to execution delay. The main advantage of FP-Growth algorithm is it do not produces candidate sets. It represents the transaction information into a tree structure.MBA in retail business refers to research that provides the retailer with information to understand the purchase behavior of a buyer.

Raorane A.A et al. in their study the data set used is collected from a supermarket named Shetkari Bazar in kolhapur city in Maharashtra. Finding the relation among the baskets of customers some interesting association among products come out. For better customer satisfaction and enhance business the storage of products is designed in a better approach.
In their study Raorane A.A et al. analyzes the huge amount of data exploiting the consumer behavior and helps to make the correct decision to lead the competitive market [3].

In their work onA Study on Ant Colony Optimization with Association Rule, Dr. T. Karthikeyan, Mr. J. MohanaSundaram describes the generation of frequent sets from the large data set using the techniques of ant colony optimization (ACO) [4].In the Ant Colony Optimization , A Study on Ant Colony Optimization with Association Rule, is an meta heuristic algorithm .It exhibits the cooperative foraging behavior of ants to find and exploit the food source that is nearest to nest. ACO is based on supportive search paradigm that can be applicable to the solution of combinatorial optimization problem. ACO can also be used for the classification task of data mining. Applying ACO on a large dataset it becomes easy to find the occurrence of the same item set. So it helps in generation of frequent item sets and finding the association of the items [4].

XIE Wen-xiu et al. determined the correlation among the items in a transaction, called basket data is done. The basket data analysis is done using client server architecture. This paper integrates words segmentation technology and association rule mining technology [5].
The client receives a set of items and read the item characteristics from the server to form the association rules and returns a set of items which are strongly associated to the received set. The server generates characteristics automatically for each item by using word segmentation technology. Then mines the items characteristic association rules and stores the set of rules in Database [5].

The author Mohammed M Mazid et al. in their study explores the similarities and dissimilarities among different association rule mining algorithms like Apriori algorithm, Partial Decision Tree algorithm. Apriori algorithm provides more accuracy in training and testing of data comparison, less computational time. But do not provide class attribute rules each time. Whereas Partial Decision Tree (PART) algorithm provides class attribute rules each time [6].

The paper by Mohammed M Mazid et al. based on conceptually introduction for practical applications of association rules in retail marketing. To increase the quality of service, customer satisfaction and analyze product and customer information data mining techniques are implemented in retail sector by Hongwei Liu et.al. [6].

From the transactional database of retail sector a huge amount of data is gathered. From these previous records they can extract some useful knowledge, which can be used to understand the marketing trends and purchasing trends. Here the role of data mining comes into picture. Mining useful knowledge from huge data has significant role in decision making. Association Rule Mining is one of the popular techniques used in data mining. Association rules provide interesting and relevant data from transactional data [6].

Liu Yongmei and Guan Yong [7] explained the market data analysis is applying FP-Growth algorithm. The frequent item sets are generated using a tree structure. But do not require generating candidate sets. The FP-tree provides frequent itemsets from the data sets.

The Application of Association Rules in Retail Marketing Mix by Hongwei Liu, Bin Su, Bixi Zhang [8] made a correlation analysis, business transaction and customer data analysis to exhibit the association among them. In the analysis the authors found some interesting data relating to transaction and customer data and suggested an optimal marketing mix strategy to increase quality of service, profit and customer satisfaction.

Mu-Chen Chen et al.[9] describes data mining adoption to predict customer behavior. Applying data

mining on a large database implicit, previously unknown, and potentially useful information including knowledge rules, constraints and regularities can be obtained. Also this helps to understand the changes in market trend, customer behavior. This information may help to design further marketing strategies.

During mining the large data base, need to ensure the data is accurate and consistent. Sometime from a large amount of raw data some useful information may derive at the time of mining. This information can be used in targeted marketing or in a segmented marketing. At the same time the mining result shows the changes in purchasing patterns, new added patters, deviated patterns etc. Here the dataset used is from FMCG retail sector database. From the study contribution of each customer estimated. Change in customer purchase and expectations also estimated.

The main idea of Apriori algorithm is to find a useful pattern in various sets in dataset. The study by Mu-Chen Chen, Ai-Lun Chiu, Hsu-Hwa Chang [9] suggests improving efficiency and accuracy of the algorithm. The main advantage of this algorithm is it is memory efficient. Here the association rules are used to discover potential relations between the sets of data items.

YücelSaygın, ArnoldReisman and YunTong Wang [10], designed mathematical model on conflict and cooperation between intelligent relational decision makers. It is also called interactive decision theory. The main concept of game theory is zero-sum game. That means distribution of equal amount of the total amount among all the participants without loss.

The author has explained using this zero-sum game in data mining the market competitors can get benefit. By exchanging information mutually both the business competitors can be beneficial. For this purpose retailers implement bundling. Bundling is useful when customers have heterogeneous demands and they cannot be classified or price discriminated.

With the use of data mining market basket analysis the information can easily obtained from the huge data without disclosing the privacy of the company's database. Because here, only the pattern of data are the matters of concern.

In Mining Fuzzy Association Rules in a Bank-Account Database by Wai-Ho Au and Keith C. C. Chan[11] proposed a technique to discover some hidden data patterns from banking database. The aim is to improve related banking services to its customers. Wai-Ho Au and Keith C. C. Chan represented fuzzy value of the linguistic terms to determine data patterns. In their method conjunction and disjunction also used to get the fuzzy value of the data. The process is named as fuzzy association rule. From the study Wai-Ho Au and Keith C. C. Chan found some new interesting hidden patterns of the bank account data which are originally not contained in the database [11].

IV. COMPARISON STUDY

In our study we have analyzed some algorithms which are well known for finding the association rules. The most popular Apriori algorithm is well known to find out the interestingness from a data set. But the main problem with this algorithm is time complexity and space complexity. This algorithm executes the whole data set again and again to find out the association, which takes a lot of time. After Apriori another algorithm is developed named as Frequent Pattern Growth algorithm. It represents the associations among items in nodes and stored for further use. So the time complexity and space complexity is reduced. Also some modifications are done with Apriri algorithm to improve its performance. Zero-sum game theory is also implemented to exhibit the advantages of Apriori algorithm. Fuzzy technique is also implemented to extract the association rules. Boolean value is also used with Apriori to exhibit the improved performance of the algorithm

V. RESEARCH FINDINGS

Basically the data mining deals with pattern extraction or information extraction. The process of data mining consists of three stages. They are:

1. Initial exploration
2. Model building or pattern identification with validation and verification
3. Deployment of the application of the model to new data in order to generate predictions

The aim of data mining is rule mining or in simple words we can say pattern extraction. For this purpose many algorithms are developed and also the development process is going on. The most popular algorithm is Apriori. Apriori is gone through many modifications to improve its performance. FP-Growth algorithm is also based on the same concept, finding the co-occurrences of item(s).

From the study of the existing algorithms and methods we find that the existing data mining algorithms can extract the pattern form large data set successfully. But the data volume is increasing day by day. So the methods need to be very efficient to accommodate the increasing data volume. The method should be compatible with different types of databases that is it should be robust.

VI. CONCLUSION AND FUTURE SCOPE

In this paper the objective of our study is to analyze the performance of the existing data mining algorithms. Our discussions focused on Apriori algorithm, FP-Growth algorithm and modified Apriori algorithm. We have discussed the performance and execution of the algorithms. In future a lot of research needs to be done to improve the efficiency of the existing data mining algorithms.

REFERENCES

1. Mining Association Rules Based on Boolean Algorithm – a Study in Large Databases, ChintaSomeswaraRao, D. Ravi Babu, R. Shiva Shankar, V. Pradeep Kumar, J. Rajanikanth, and Ch. Chandra Sekhar, International Journal of Machine Learning and Computing, Vol. 3, No. 4, August 2013
2. An Approach to Improve Apriori Algorithm Based On Association rule Mining, ChanchalYadav, Shuliang Wang, Manoj Kumar, 4th International Conference on Computing Communication and Networking Techonologies - 2013 - 2013 July 4-6, 2013, Tiruchengode, India.
3. Association Rule – Extracting Knowledge Using Market Basket Analysis,Research Journal of Recent Sciences, Vol. 1(2), 19-27, Feb. (2012)-Raorane A.A., Kulkarni R.V. and Jitkar B.D.
4. A Study on Ant Colony Optimization with Association Rule,Dr. T. Karthikeyan, Mr. J. MohanaSundaram, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 5, May 2012.
5. Market basket analysis based on text segmentation and association rule mining, XIE Wen-xiu, QiHeng-nian, Huang Mei-li,2010 First International Conference on Networking and Distributed Computing.
6. A Comparison Between Rule Based and Association Rule Mining Algorithms, Mohammed M Mazid, A B M Shawkat Ali, Kevin S Tickle, 2009 Third International Conference on Network and System Security.
7. Application in Market Basket Analysis Based on FP-growth Algorithm , Liu Yongmei, Guan Yong, 2009 World Congress on Computer Science and Information Engineering
8. The Application of Association Rules in Retail Marketing Mix,Hongwei Liu, Bin Su, Bixi Zhang, Proceedings of the IEEE International Conference on Automation and Logistics August 18 - 21, 2007, Jinan, China.
9. Mining changes in customer behavior in retail marketing, Mu-Chen Chen, Ai-Lun Chiu, Hsu-HwaChang, Expert Systems with Applications 28 (2005) 773–781
10. Value of Information Gained From Data Mining in the Context of Information Sharing, YücelSaygın, Arnold Reisman, and YunTong Wang, IEEE TRANSACTIONS ON ENGINEERING MANAGEMENT, VOL. 51, NO. 4, NOVEMBER 2004,pp. 441-450.
11. Mining Fuzzy Association Rules in a Bank-Account Database, Wai-Ho Au and Keith C. C. Chan, IEEE TRANSACTIONS ON FUZZY SYSTEMS, VOL. 11, NO. 2, APRIL 2003