ISSN 2320 - 2602
Volume 5 No.12, December 2016
International Journal of Advances in Computer Science and Technology
Available Online at http://www.warse.org/IJACST/static/pdf/file/ijacst025122016.pdf

A Survey on Improved Association Rule Mining for market based analysis

Girish Kumar Ameta [1], Dr. Vibhakar Pathak [2]
[1]M.Tech Scholar, Arya College of engineering & I.T.,Jaipur ,India
Email:kumargirish360@gmail.com
[2]Professor, Arya College of engineering & I.T.,Jaipur ,India [3]Affiliation, Country
Email: vibhakarp@rediffmail.com

ABSTRACT

Association rule mining is a very important process of data mining. It is used to generate all significant association rules between the item sets in a large repository of market data. The basic algorithm to find association rules is apriori algorithm which is frequently used by researchers to improvise the better results in terms of fast processing and meaning full mining over the period of time. The association rule mining is frequently used in market basket analysis. Market basket analysis provides the knowledge about data pattern from itemsets present in the transaction. 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*

1. INTRODUCTION

The association rules find 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 . 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 Selection- where data relevant to the analysis task are retrieved from the database.

iii) Data transformation- where data are transformed or consolidated into forms appropriate for mining by performing summery or aggregation operations for instance.

iv) Data mining- an essential process where intelligent methods are applied in order to extract data patterns.

v) Pattern evolution- to identify the truly interesting patterns representing knowledge based on some interestingness measures.

vi) Knowledge presentation- where visualization and knowledge representation techniques are used to present the mined knowledge to the user.

2. ASSOCIATION RULE MINING

Association rule mining is to find out association rules that satisfy the predefined minimum support and confidence from a given database. The problem is usually decomposed into two sub problems. One is to find those item sets whose occurrences exceed a predefined threshold in the database; those item sets are called frequent or large item sets. The second problem is to generate association rules from those large item sets with the constraints of minimal confidence. Suppose one of the large item sets is Lk, Lk = {I1, I2… Ik},

association rules with this item sets are generated in the following way: the first rule is {I1, I2… Ik-1} ⇒ {Ik}, by checking the confidence this rule can be determined as interesting or not. Then other rules are generated by deleting the last items in the antecedent and inserting it to the consequent, further the confidences of the new rules are checked to determine the interestingness of them. [1]

Those processes iterated until the antecedent becomes empty. Since the second sub problem is quite straight forward, most of the researches focus on the first sub problem. The first subproblem can be further divided into two sub-problems: candidate large item sets generation process and frequent item sets generation process. We call those item sets whose support exceed the support threshold as large or frequent itemsets, those item sets that are expected or have the hope to be large or frequent are called candidate item sets. In many cases, the algorithms generate an extremely large number of association rules, often in thousands or even millions. Further, the association rules are sometimes very large[3]. It is nearly impossible for the end users to comprehend or validate such large number of complex association rules, thereby limiting the usefulness of the data mining results. Several strategies have been proposed to reduce the number of association rules, such as generating only "interesting" rules, generating only "no redundant" rules, or generating only those rules satisfying certain other criteria such as coverage, leverage, lift or strength.[1]

3. 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 [2]. Chinta Someswara Rao[2] 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.[2] 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. Chinta Someswara Rao et al.[2] 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 Chanchal Yadav, 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 on A Study on Ant Colony Optimization with Association Rule, Dr. T. Karthikeyan, Mr. J. Mohana Sundaram 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 , 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.[7]

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. 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.[9]

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 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. [9]

4. PERFORMANCE SURVEY OF SOME ALGORITHMS

Table 4.1
Brief Comparison Between Apriori and FP-Growth Algorithm

Factor	Apriori Algorithm	FP-Growth Algorithm
Data Structure	Array Based	Tree based
Techniques	Join and prune Method	it constructs conditional frequent pattern tree and conditional pattern base from database which satisfy the minimum support
Memory Utilization	Requires large memory space	Less memory
Number of Scans	Multiple scan for generating candidate sets	Twice only
Execution Time	More	Less than apriori
Database	Sparse database and dense database	Large and medium database
Accuracy	less	More Accuate
Applications	Best for closed itemsets.	Large itemsets.

Thus in our study we have analyzed some algorithms which are popular 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

REFERENCES

1. Sotiris Kotsiantis and Dimitris Kanellopoulos, “Association Rules Mining: A Recent Overview”, GESTS International Transactions on Computer Science and Engineering, Vol.32 (1), 2006, pp. 71-82.
2. ChintaSomeswaraRao, D. Ravi Babu, R. Shiva Shankar, V. Pradeep Kumar, J. Rajanikanth, and Ch. Chandra Sekhar “Mining Association Rules Based on Boolean Algorithm – a Study in Large Databases”, International

Journal of Machine Learning and Computing, Vol. 3, No. 4, August 2013, pp 231-238.

3. Chanchal Yadav, Shuliang Wang and Manoj Kumar, " An Approach to Improve Apriori Algorithm Based On Association rule Mining", 4th International Conference on Computing Communication and Networking Techonologies - 2013 July 4-6, 2013,Tiruchengode, India.
4. Dr. T. Karthikeyan and Mr. J. MohanaSundaram, "A Study on Ant Colony Optimization with Association Rule", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 5, May 2012.
5. XIE Wen-xiu, QiHeng-nian and Huang Mei-li, "Market basket analysis based on text segmentation and association rule mining" , 2010 First International Conference on Networking and Distributed Computing.
6. Mohammed M Mazid, A B M Shawkat Ali, and Kevin S Tickle, "A Comparison Between Rule Based and Association Rule Mining Algorithms", 2009 Third International Conference on Network and System Security. IEEE, 2009. p. 452-455 4 pages Refereed 9780769538389 (online).
7. L. Yongmei and G. Yong, "Application in Market Basket Research Based on FP-Growth Algorithm," *2009 WRI World Congress on Computer Science and Information Engineering*, Los Angeles, CA, 2009, pp. 112-115.doi: 10.1109/CSIE.2009.1073.
8. Hongwei Liu, Bin Su and Bixi Zhang, "The Application of Association Rules in Retail Marketing Mix", International Conference on Automation and Logistics August 18 - 21, 2007, Jinan, China.
9. Mu-Chen Chen, Ai-Lun Chiu and Hsu-Hwa Chang, "Mining changes in customer behavior in retail marketing, Expert Systems with Applications" , Expert Systems with Applications 28 (2005) 773–781.