An Efficient Technique for mining Association rules using Enhanced Apriori Algorithm A Literature survey

Mr.Gaurav P.Wankhade
ME 3rd Sem Information Technology
Prof. Ram Meghe Institute of Technology & Research
Badnera-Amravati, India
gaurav.wankhade487@gmail.com

Prof. Ms. V.M.Deshmukh
Associate Professor & Head Information Technology
Prof. Ram Meghe Institute of Technology & Research
Badnera-Amravati, India
msvmdeshmukh@rediffmail.com

***Abstract*—** At present Data mining has a lot of e-Commerce applications. The key problem in this is how to find useful hidden patterns for better business applications in the retail sector. For the solution of those problems, The Apriori algorithm is the most popular data mining approach for finding frequent item sets from a transaction dataset and derives association rules. Association Rules are the discovered knowledge from the data base. Finding frequent item set (item sets with frequency larger than or equal to a user specified minimum support) is not trivial because of its combinatorial explosion. Once item sets are obtained, it is straightforward approach to generate association rules with confidence value larger than or equal to a user specified minimum confidence value.

Keywords- *Data mining, e-Commerce , apriori algorithm , association rules , retail sector.*

*****

I. INTRODUCTION

Data mining is nothing but, the extract of knowledge from databases. Data generally by organized in tables that hold a set of complete database environment. A table can be looked like as a matrix. Each matrix row symbolizes an instance that is coupled with a test case to analyze. An example of an instance may be a teacher, while all the instances may be a population of college teachers to analyze. Each matrix column symbolizes all the values associated with the domain name.

Association rule mining is one of the best discussed models for data mining. *Definition-* The discovery of association rules from a transaction database DB, Y={y1, y2 ….. yn} be array of n different element called item sets in DB, each transaction T in DB is a set of item (i.e. item sets) .The support of an association pattern is the percentage of task-relevant data transactions for which the pattern is true. Confidence is defined as the measure of certainty or trustworthiness associated with each discovered pattern.

1.1 Apriori Algorithm

Following the original definition by Agrawal [1] the problem of association rule mining is defined as: Let I = {i1, i2, ..., in} be a set of n binary attributes called items. Let D = {t1, t2... tn} be a set of transactions called the database. Each transaction in a set has a unique transaction ID and contains a subset of the items in binary attribute. A rule is defined as an implication of the form $X \rightarrow Y$ where $X, Y \subseteq I$ and $X \cap Y = \emptyset$. The sets of items (for short item sets) X and Y are called antecedent (left-hand-side or LHS) and consequent (right-hand-side or RHS) of the rule respectively. To illustrate the concepts, we use a small example from the supermarket domain. The set of items is I = {milk, bread, butter, beer} and a small database containing the items (1 codes presence and 0 absence of an item in a transaction) is shown in the table below. An example association rule for the supermarket could be {milk, bread} => {butter} meaning that if milk and bread is bought, customers also bought butter.

Apriori uses bottom up strategy. It is the most famous and classical algorithm for mining frequent patterns. Apriori algorithm works on categorical attributes. Apriori uses breadth first search

II. LITERATURE REVIEW/SURVEY

Huan Wu et al. [2] (2009) proposed an improved algorithm IAA based on the Apriori algorithm. It adopts a new count based method to prune candidate item sets and uses generation record to reduce total data scan amount. Experiment demonstrates that their algorithm outperforms the original Apriori and some other existing ARM methods.

Goswami D.N. et al [3] (2010) considered three different frequent pattern mining approaches (Record filter, intersection and Proposed Algorithm) based on classical Apriori algorithm. The algorithm of association rule discovery proceeds in two and more steps but in their approach discovery of all frequent items take the same steps but it takes lesser time as compared to the conventional algorithm. They concluded that in this new approach, they have various key ideas of reducing time.

K. Vanitha and R. Santhi [4] (2011) described an implementation of Hash based Apriori algorithm. They analyzed, theoretically and experimentally, the principal data structure of their solution. It is the main factor in the efficiency of their implementation. They proposed a very effective hash

based algorithm for the candidate set generation. The number of candidate 2 –item sets generated by the proposed algorithm is, in order of magnitude, smaller than that by previous methods, thus resolved the performance bottleneck.

Sunil Kumar et al [5] (2012) proposed a new algorithm which takes less number of scans to mining the frequent item sets from the large database which leads to mine the association rule between the database. They addressed the importance of knowledge mining from a large data set and overview of existing algorithm and its flaws and innovative solution with a new algorithm for data mining from the large datasets. It was seen in the observation and analysis that the new proposed algorithm performs better than the existing Apriori algorithm. The relative performance of the proposed algorithm using probability and matrix under different minsupport specify its excellence and features. That improved algorithm can be used in many areas such as medical, image processing and ERP etc. with a reduced time and space complexity requirements.

Rehab H. Alwa and Anasuya V. Patil In 2013 [6] described a novel approach to improve the Apriori algorithm through the creation of Matrix – File. For this purpose the MATLAB is used, where the database transactions are stred. Thus repeated scanning is avoided and particular rows & columns are extracted and perform a function on that rather than scanning entire database.

Mohammed Al-Maolegi and Bassam Arkok 2013[7] indicated the limitation of the original Apriori algorithm of wasting time for scanning the whole database searching on the frequent item sets, and presented an improvement on Apriori by reducing that wasted time depending on scanning only some transactions.

Rina Raval et al 2013[8] performed a survey on few good improved approaches of Apriori algorithm such as Record Filter and intersection approach, Improvements based on set size frequency and trade list, by reducing candidate set and memory utilization, improvement based on frequency of item . They found that mostly improved Apriori algorithms aims to generate less candidate sets and yet get all frequent items. They concludes that many improvements are needed basically on pruning in Apriori to improve efficiency of algorithm.

In 2013, Jugendra Dongre , Gend Lal Prajapati and S. V. Tokekar [9] described an approach for mining association rules using apriori algorithm by calculating various support and confidence values for each transaction to find frequent item sets. But it takes more time for that process and it is not efficient for large database upto 2000 transactions. . They concluded that many improvements are needed basically on pruning in Apriori to improve efficiency of algorithm.

In 2013 Jaishree Singh et al [10] proposed an Improved Apriori algorithm which reduces the scanning time by cutting down not required transaction records as well as reduce the redundant generation of sub items during pruning the candidate item sets, which can form directly the set of frequent item sets and eliminate candidate having a subset that is not frequent. The improved algorithm not only optimize the algorithm by reducing the size of the candidate set of k-item sets, Ck, but it also reduces the amount of I / O spending by cutting down transaction records in the database. The performance of algorithm is optimized so that we can mine association information from massive data faster and better.

In 2013 Jose L. Balcazar [11] proposed a measure to the confidence boost of a rule. Acting as balance to confidence and support, it helps to gain small and crisp sets of mined association rules and solves the well known problem such as rules of negative correlation may pass the confidence bound. i.e analyzed the properties of two versions of the notion of confidence boost, one being the natural generalization of the other. They proposed algorithms to filter rules according to their confidence boost, and compared the given concepts to some similar notions in the literature, and described the results of experiments employing the new notions on standard benchmark datasets. i.e also described an open source association mining tool that embodies one of our variants of confidence boost in such a way that the data mining process does not require the user to select any value for any parameter.

In 2014, Sallam Osman Fageeri, Rohiza Ahmad, Baharum B. Baharudi [12] proposed a semi Apriori Algorithm for mining association rules using binary-based data structure that is used to discover the frequent item sets as well as association rules. The proposed technique avoids the cost of generating large number of candidate sets through using the binary-based data structure; hence, minimize the execution time.

III. PROPOSED WORK

In Apriori algorithm the time taken is directly proportional to the number of transactions where as in the Enhanced apriori algorithm the time taken is directly proportional to important transactions only. In proposed system important transaction extract on the basis of support and confidence value which will predetermine the exact need of customer, which can be perform high operations in less time as compared to existing apriori Algorithm, this is a feature over apriori algorithm, so for better utilization of data mining technique an enhanced apriori algorithm is preferred. In this case the comparison between Apriori and enhanced Apriori algorithm is takes place.

CONCLUSION

Data mining as a tool used to find the hidden pattern of the frequently used item-sets. An enhanced Apriori Algorithm may play an important role for finding these patterns from large databases so that various sectors can make better business decisions especially in the retail sector. Enhanced Apriori algorithm may find the tendency of a customer on the

basis of frequently purchased item-sets. There are wide range of industries have deployed successful applications of data mining. It in retail industry can be deployed for market campaigns, to target profitable customers using reward based points. The retail industry will have gain, sustain and will be more successful in this competitive market if adopted data mining technology for market campaigns.

REFERENCES

[1] Agrawal.R and Srikant.R. "Fast algorithms for mining association rules". In Proc.Int'l Conf. Very Large Data Bases (VLDB), Sept. 1994, pages 487–499.

[2] Huan Wu, Zhigang Lu, Lin Pan, Rongsheng Xu and Wenbao Jiang,"An improved Apriori-Based Algorithm for Association Rules Mining", Sixth international Conference on Fuzzy Systems and Knowledge Discovery,2009.

[3] Goswami D. N., Chaturvedi Anshu and Raghuvanshi C. S., "An Algorithm for Frequent Pattern Mining Based on Apriori", international Journal on Computer Science and Engineering, Vol.02, No. 04, 942-947, 2010.

[4] K. Vanitha and R.Santh,"Using Hash Based Apriori Algorithm to Reduce the Candidate 2-itemsets for Mining Association Rule", Journal of Global Research in Computer Science, ISSN: 2229-371X, Volume 2, No. 5, April 2011.

[5] Sunil Kumar S, Shyam Karanth S, Akshay K C, Ananth Prabhu, Bharatraj Kumar S, "Improved Apriori Algorithm Based on Bottom Up Approach Using Probability and Matrix" ,International Journal of Computer Science issues, Vol. 9, issue 2, No. 3,242-246, March 2012.

[6] A. Rehab H. Alwa and B. Anasuya V Patil,"New Matrix Approach to improve Apriori Algorithm" in international Journal of Computer Science and Network Solutions, Vol. 1 No 4 December 2013.

[7] Mohammed Al - Maolegi and Bassam Arkok, "An improved Apriori Algorithm for Association Rules"in international Journal on Natural Language Computing Vol. 3 No. 1, February 2014

[8] Rina Rawal, Prof indrjeet Rajput, Prof. Vinit kumar Gupta "Survey on several improved apriori algorithms"in IOSR Journal of Computer Engineering,Volume 9, April 2013..

[9] Jugendra Dongre , Gend Lal Prajapati and S. V. Tokekar , "The role of Apriori algorithm for finding Association rules in data mining" in International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT) 2014.

[10] Jiao Yabing, "Research of an Improved Apriori Algorithm in Data Mining Association Rules", in internatinal Journal of Computer and Communication Engineering, Vol. 2,No.1, January 2013.

[11] Jose L. Balcazar, " Formal and Computational Properties of the Confidence Boost of Association Rules" in ACM Transactions on Kmowledge Discovery from Data, Vol 7, No. 4,November 2013.

[12] Sallam osman fageeri, Rohiza ahmad, Baharum b. baharudi, "an enhanced semi apriori algorithm for mining association rules" .In Journal of Theoretical and Applied Information Technology Vol. 63 No.2 ,may 2014.