An Efficient Technique for mining Association rules using Enhanced Apriori Algorithm

Mr. Gaurav P. Wankhade
ME 4th 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*—** There are Various mining algorithms of association rules. One of the most popular algorithm is Apriori that extracts frequent itemset from large database and getting the association rule for discovering the knowledge. this paper pitfalls the limitation of the original Apriori algorithm for wasting time for scanning the whole database searching on to the frequent itemsets, and presents an technique on Apriori by reducing that wasted time depending on scanning only some transactions whose support value is bigger than 25% of minimum Support is taken as frequent item set and is added to the frequent item sets and then rules are formed. An enhanced Apriori algorithm may find the tendency of a customer on the basis of frequently purchased item-sets The proposed algorithm is useful as a frequent item sets predictor with lower number of scans.

***Keywords*—** *Association rules, Apriori, Enhanced Apriori, Frequent item set, Support.*

*****

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 looked like as a matrix. Each matrix row symbolizes an instance that is coupled with a testcase 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, every transaction T in DB is a set of item (i.e. item sets) .The support of an 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 = {r1, r2... rn} 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 . 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, customer 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 Goswami D.N. et al [3] (2010) proposed three different frequent pattern mining approaches (Record filter, intersection and Proposed Algorithm) based on classical Apriori algorithm. K. Vanitha and R. Santhi [4] described an implementation of Hash based Apriori algorithm. They analyzed, theoretically and experimentally, the principal data structure of their solution. 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. 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.

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. 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 items, which can form directly the set of frequent item sets and eliminate candidate having a subset that is infrequent. 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.

III. ASSOCIATION ALGORITHMS

The Apriori Algorithm

Is the most famous of all association rule mining algorithms, They have decomposed the problem of mining association rules into two parts:

a. Get all combinations of items that have transaction support above minimum support Until item-set size reaches maximum.. Call these combinations frequent item sets.

b. Use that frequent item sets to generate the desired rules.

Let's illustrate the process of apriori with an example,the database which is showing the number of item sets purchased by the consumers from a Stationary shop.Let Minimum support is 0.3.Single item like Eraser, Cutter etc. in the given database every item occurs three or more time than the minimum support or minimum support threshold is 0.3. Now focus on interestingness of the one element items, so database has many items like Sharpener, cutter, Colors etc.So interestingness 1- element item-sets{Eraser}, {Cutter}, {Sharpener}, {inkpot}, {Scale} etc.

Similarly We can find two or more element item Sets by using Apriori Algorithm and by using that items we find Strong Rules by using following Procedure:

- At each iteration divide every frequent item-set X into two parts antecedent (LHS) and consequent (RHS) this represents a rule of the form LHS->RHS.
- The confidence of a rule is support(X) /support(LHS)
- Delete rules whose confidence is less than minimum confidence.

Limitations Of Existing System:

Apriori algorithm has some weakness in spite of being clear and simple. The main limitation is costly wasting of time to hold a vast number of candidate sets with much frequent itemsets, low minimum support and large itemsets. For example, if there are 100 from frequent 1-itemsets, it needs to generate more than 107 candidates into 2-length which in turn they will be tested and accumulate. Furthermore, to detect frequent pattern in size 100 (e.g.) v1, v2… v100, it has to generate 2100 candidate item sets that yield on costly and wasting of time of candidate generation. So, it will scan database many times repeatedly for finding candidate item sets. It will be very low and inefficiency when memory capacity is limited with large number of transactions [7].

Enhanced Apriori Algorithm

The algorithm starts by taking the item sets and the database transactions; the next step is generating the item sets candidates from size 1 while supporting each element in the item sets. Each item whose support is bigger than 25% of minimum Support is taken as frequent item set and is added to the frequent item sets. After this step all frequent item sets with size 1 become the same as the original Apriori algorithm, then the algorithm generates the frequent itemsets from size 2 to the size of the stop condition by combining the itemsets. The generated frequent itemsets is a prediction of the frequent itemsets that should be generated by the original Apriori algorithm. After this step, the algorithm calculates the support values of the generated frequent itemsets and they are compared to the support value of the original Apriori algorithm and the results are verified.

. Let's illustrate the process of proposed model with an example, the database which is showing the number of item sets purchased by the consumers from a Stationary shop.

Let Minimum support is 0.3 then the minimum Support value becomes minsup(0.3+0.3*0.25)=0.375.Single item like Eraser, Cutter etc. in the given database every item occurs three or more time than the minimum support or minimum support threshold is 0.375. Now focus on interestingness of the one element items, so database contains many items like Sharpener, cutter, Colors etc.So interestingness single- element item-sets

{Eraser}, {Cutter} etc.

Similarly We can find two or more element item Sets by using Apriori Algorithm and by using that items we find Strong Rules by using following Procedure:

- At each iteration divide every frequent item-set X into two rows antecedent (LHS) and consequent (RHS) this represents a rule of the form LHS->RHS.
- The confidence of a rule is support(X) /support(LHS)
- Delete rules whose confidence is less than minimum confidence.

IV. EXPERIMENTAL RESULTS

[13]The performance of the proposed algorithm has been evaluated by comparing its execution time and Scans to the execution time and Scans of the original Apriori algorithm. In order to test the performance of the two algorithms, an extensive analysis has been done. The procedures of the experiment are as follows:

1. Preparing the dataset: transaction dataset from a Stationary Shop that contains 100 transactions with 9 items. The items are aggregated to the department level in the shop,for example, the customer's shopping cart contains at least one product from that department.
2. An implementation plan for the original Apriori and the improved algorithm has been developed.
3. The original Apriori algorithm has been tested on the supermarket dataset several times with different support and confidence values and the stop condition generates frequent item sets with size 3.
4. The proposed algorithm will be tested on the same dataset and the stop condition generates frequent item sets with size 3.

The proposed algorithm has compared to the original Apriori algorithm regarding the following two points:

1. Number of database scans to find the frequent items
1. The time required for finding frequent item sets by both algorithms.

The time consumed to generate frequent item sets in the proposed algorithm is less than the time consumed in the original Apriori; the proposed algorithm reduces the time consumed by 40%. The proposed algorithm is useful as a frequent item sets predictor with lower number of scans.

V. CONCLUSION

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 work important transaction extract on the basis of support and confidence value which will predetermine the exact need of customer, which can perform high operations in less time as compared to existing apriori Algorithm. So for better utilization of data mining technique an enhanced apriori algorithm is preferred. By performing various experiments conclude that,The time consumed to generate frequent item sets in the proposed algorithm is less than the time consumed in the original Apriori; the proposed algorithm reduces the time consumed by 40%. The proposed algorithm is useful as a frequent item sets predictor with lower number of scans.

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.

[13] Dr. Ayman E. Khedr, Fahad Kamal Alsheref,” Improved Apriori Algorithm VIA Frequent Itemsets Prediction”.In International Journal of Advanced Research in Computer

ISSN: 2321-8169
3663 - 3666

Science and Software Engineering Volume 4, Issue 4, April 2014 ISSN: 2277 128X.

[14] Mr. Gaurav P. Wankhade, Prof. Ms. V. M. Deshmukh, "An Efficient Technique for mining Association rules using Enhanced Apriori Algorithm A Literature survey", December 14 Volume 2 Issue 12 , International Journal on Recent and Innovation Trends in Computing and Communication (IJRITCC), ISSN: 2321-8169, PP: 3856 - 3858