"An Effective Analysis of Apriori Algorithm with Vectorization Approach"

Ms.Arpita Lodha[1], Vishal Shrivastava[2]

[1]M, Tech Scholar, Deptt. of CS/IT, ACEIT, Arya Group of Colleges, Jaipur, Rajasthan, India

[2] Associate Professor, Deptt. of CS/IT, ACEIT, Arya Group of Colleges, Jaipur, Rajasthan, India

ABSTRACT

Data mining is young and fast growing field of research. It also known as knowledge Discovery from data (KDD).Data mining is the process of discovering interesting patterns and knowledge from huge amount of data. Apriori algorithm is the traditional algorithm of generating association rules, which derives all of the maximum frequent item sets. When this algorithm encountered dense data due to the large number of long patterns emerge, this algorithm's performance declined dramatically. In order to find more valuable rules, this paper proposes a modified algorithm of association rules, the Vectorised Apriori (VA) algorithm. Finally, the improved algorithm is verified, the results show that the improved algorithm is reasonable and effective, can extract more value information.

Keywords:*ModifiedApriori, Frequentitemsets,threshold,confidence,vectorization*_

1INTRODUCTION

Data mining also known as the computer-aid process that digs and analyzes enormous sets of data and then extracting the knowledge or information out of it. By its simplest definition, data mining automates the detections of relevant patterns in the database [1]. Discovery of frequent item set is done by association rules. Discovery of frequent item set is done by association rules mining. It consists of procedure [2] first, finding frequent itemsets in the database using a threshold value and constructing the association rule from the frequent itemsets with specified confidence. It relates to the association of items wherein for every occurrence of A, there exists an occurrence of B.

Discovery of frequent item set is done by association rules. Retail store also use the concept of association rule for managing marketing, advertising, and errors that are presented in the telecommunication network.

Apriori algorithm is the most widely used association rule mining algorithm [3]. However, several limitations have been discovered in this method [4] such as:

- Several iterations of data are needed for mining data
- Usually generates items which are irrelevant
- Difficulties in finding unusual events

With these limitation several works have been noted to improve the efficiency of Apriori algorithm.

2LITERATURE SURVEY

In this section author has discussed some research papers which had been previously undertaken in the field of association rule mining.

X. Luo and W. Wang[5]. In this paper an improved apriori is to make a Matrix library. The matrix library contains a binary representation where 1 indicated item present in transaction and 0 indicated it is absent. Assume that in the event Matrix library of database D, the matrix is Amxn, then the corresponding BOOL data item set of item I_j(1<= j <= n)in P in Matrix Amxn is the mat of I_j, Mati is items in the mat. Now by counting the number of 1's in the matrix we can easily find the occurrence of that item. For 2-itemset we can just multiply the binary representation of the items to get the occurrence to that items together. To find how many times item $I_j$and I_k are appearing together we have to multiply the MAT(Ij) and MAT(Ik). (i.e) MAT(Ij,Ik)=MAT(Ij) * MAT(Ik).

T.Junfang[6]. In this paper Improved Apriori algorithm works by compressing transaction database, by using an attribute named count the efficiency of the algorithm is improved. The transaction database creates lots of same records after a certain amount of time. So clustering can be done for these kinds of databases. Only one entry is made in the database and whenever the same item in transaction occurs as the previous one it is discarded. To show the frequency of repeated records an attribute is added named count. The next steps are similar to apriori algorithm like candidate set generation and pruning.

Goswami D.N., ChaturvediAnshu. Raghuvanshi C.S [7].In this paper they presented a different approach in Apriori algorithm to count the support of candidate item set. In this when we count the support of candidate set of length k, we also check its occurrence in transaction whose length may be greater than, less than or equal to the k. But in the new approach we count the support of candidate set only in the transaction record whose length is greater than or equal to the length of candidate set, because candidate set of length k, cannot exist in the transaction record of length k-1, it may exist only in the transaction of length greater than or equal to k. This approach has taken very less time as compared to classical Apriori.

Goswami D.N., ChaturvediAnshu. Raghuvanshi C.S [7] In the previous section they have described the Record filter approach based on Apriori, now they suggested one another changes in Apriori which gives the better result as compare to the Record Filter approach. The Intersection Algorithm is designed to improve the efficiency, memory management and remove the complexity of Apriori. Here they proposed a different approach in Apriori algorithm to count the support of candidate item set. Basically this approach is more appropriate for vertical data layout, since Apriori basically works on horizontal data layout. In this new approach, they used the set theory concept of intersection. In Classical Apriori algorithm, to count the support of candidate set each record is scanned one by one and check the existence of each candidate, if candidate exists then we increase the support by one. This process takes a lot of time, requires iterative scan of whole database for each candidate set, which is equal to the max length of candidate item set. In modified approach, to calculate the support we count the common transaction that contains in each elements of candidate set, by using the intersect query of SQL. This approach requires very less time as compared to classical Apriori

D.N., ChaturvediAnshu. Raghuvanshi C.S [7] in this new approach they have determined changes that are going to serve the best in the field of frequent pattern mining. In this new approach, they proposed an algorithm that uses the concept of both algorithm i.e. Record filter approach and Intersection approach in Apriori algorithm .To count the support of candidate item set ,we have considered both above mentioned approach. In this new approach, they used the set theory concept of intersection with the record filter approach. In proposed algorithm, to calculate the support, they count the common transaction that contains in each elements of candidate set, with the help of the intersect query of SQL. In this approach, they have applied a constraints that it will consider only those transaction that contain at least k items, not less than k in process of support counting for candidate set of k length. This approach requires very less time as compared to all other approaches.

3APRIORI ALGORITHM

Apriori is an algorithm proposed by R. Agrawal and R Srikant in 1994 [8] for mining frequent item sets for Boolean association rule. The name of algorithm is based on the fact that the algorithm uses prior knowledge of frequent item set properties. Apriori employs an iterative approach known as level-wise search, where k item set are used to explore (k+1) item sets.

There are two steps in each mining association rules between sets of items in large databases. The first step generates a set of candidate item sets. Then, in the second step we count the occurrence of each candidate set in database and prunes all disqualified candidates (i.e. all infrequent item sets). Apriori uses two pruning technique, first on the bases of support count (should be greater than user specified support threshold) and second for an item set to be frequent , all its subset should be in last frequent item set The iterations begin with size 2 item sets and the size is incremented after each iteration. The algorithm is based on the closure property of frequent item sets: if a set of items is frequent, then all its proper subsets are also frequent.

Disadvantage of Apriori Algorithm

- Requires too many database scans.
- Consumes large amount of time.
- Generates redundant item-sets

4MODIFIED VECTORIZED APRIORI

According to proposed work main focus to enhance the performance in term of execution time and number of database passes. In this method we applied vectorization on datasets to reduce the number of iterations.

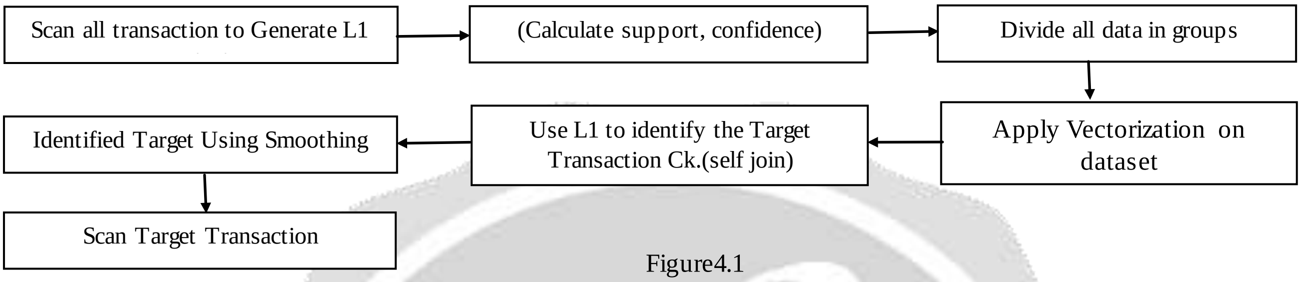
Scan all transaction to Generate L1
(Calculate support, confidence)
Divide all data in groups
Apply Vectorization on dataset
Use L1 to identify the Target Transaction Ck.(self join)
Identified Target Using Smoothing
Scan Target Transaction

Figure4.1

4.1VECTORIZATION [9]

Vectorization refers to a powerful way to speed up our algorithms with a minimum of effort. The idea of vectorization is that we would like to express our learning algorithms in terms of highly optimized operations.Vectorization (mathematics), a linear transformation which converts a matrix into a column vector. The process of converting a scalar implementation, which processes a single pair of operands at a time, to a vector implementation, which processes one operation on multiple pairs of operands at once, is called vectorization.

MATLAB is optimized for operations involving matrices and vectors. The process of revising loop-based, scalar-oriented code to use MATLAB matrix and vector operations is called vectorization. Vectorizing the code is worthwhile for several reasons:

- Appearance: Vectorized mathematical code appears more like the mathematical expressions found in textbooks, making the code easier to understand.
- Less Error Prone: Without loops, vectorized code is often shorter. Fewer lines of code mean fewer opportunities to introduce programming errors.
- Performance: Vectorized code often runs much faster than the corresponding code containing loops.

5 EXPERIMENTAL RESULTS

The modified algorithm was implemented using MATLAB by 3 random test cases used to evaluate the performance of the algorithms.

Sr.No	Association Rules	Support	Confidence
1	client33 -> client34	29.4118%	83.3333%
2	client34 -> client33	29.4118%,	58.8235%
3	client1 -> client2	20.5882%	43.75%
4	client2 -> client1	20.5882%	77.7778%
5	client3 -> client34	17.6471%	60.00%
6	client3 -> client1	14.7059%	50.00%
7	client4 -> client1	14.7059%	83.3333%
8	client2 -> client3	11.7647%	44.4444%

9	client2 -> client4	11.7647%	44.4444%
10	client4 -> client2	11.7647%	66.6667%
11	client4 -> client3	11.7647%	66.6667%
12	client4 -> client3	11.7647%	66.6667%
13	client14 -> client8	11.7647%	80.00%

Table5.1

Sl. No.	No.of transactions	VA(Execution Time)in sec
1	34	0.64761
2	244	0.28176
3	516	0.22944

Table5.2

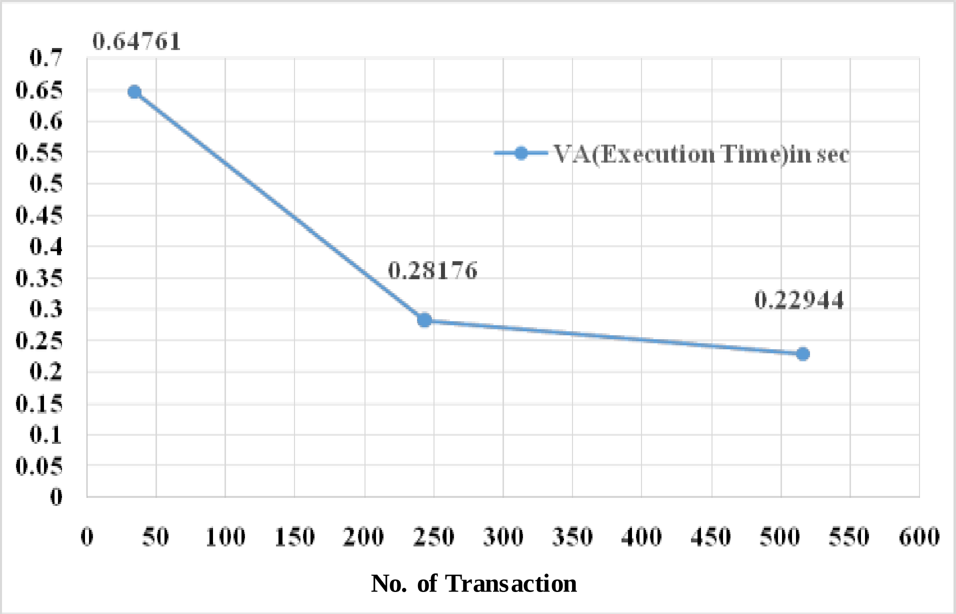

Figure5.1

VI CONCLUSION

In this paper, the execution time of 3 random test cases is analysed with vectorization approach. This study is focused on how to solve the efficient problems of Apriori algorithm and raise another association rules mining algorithm. This has certain reference value to research and solve the issues of data expiation and information lacking. It hopes to dig out more useful information.

REFERNCES

[1] "Data Mining", http://www.zentut.com/Data Mining.

[2] R. Agrawal, T. Imielinski and A. Swami, ―Mining association rules between sets of items in large databases.‖ SIGMOD'93, 207-216, Washington, D.C.

[3]MamtaDhanda, ―An Approach to Extract Efficient Frequent Patterns from transactional database,International Journal of Engineering Science and Technology (IJEST), Vol.3 No.7, July 2011, pp.

5652-5658

[4] R. Agrawal and R. Srikant, ―Fast algorithms for mining association rules,in Proceedings of the 20th VLDB Conference, 1994, pp. 487-499

[5] X. Luo and W. Wang, "Improved Algorithms Research for Association Rule Based on Matrix," 2010 International Conference on Intelligent Computing and Cognitive Informatics, pp. 415–419,Jun. 2010.

[6]T. Junfang, "An Improved Algorithm of Apriori Based on Transaction Compression," vol. 00, pp. 356–358, 2011.

[7] Goswami D.N. et. al. "An Algorithm for Frequent Pattern Mining Based On Apriori" (IJCSE) International Journal on Computer Science and Engineering Vol. 02, No. 04, 2010, 942-947

[8]Agrawal, R., Imielinski, T., and Swami, A. N. 1993. Mining association rules between sets of items in large databases. In Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, 207-216

[9]http://in.mathworks.com/help/matlab/matlab_prog/vectorization.html?requestedDomain=www.mathworks.co m