ISSN(Online): 2320 - 9801
ISSN (Print) : 2320 - 9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 6, June 2016

A Modified Apriori Algorithm for Mining Frequent Pattern and Deriving Association Rules using Greedy and Vectorization Method

Arpita Lodha[1], Vishal Shrivastava[2]

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

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

ABSTRACT: Data mining came into the existence in response to technological advances in many diverse disciplines. In other words, all the data in the world are of no value without mechanisms to efficiently and effectively extract information and knowledge from them. In comparison to other data mining fields, frequent pattern mining is a relatively recent development. This paper presents a novel approach through which the Apriori algorithm can be improved. The modified algorithm introduces factors time consumed in transactions scanning for candidate itemsets and the numbers of rules generated are also reduced.

KEYWORDS: Apriori, Frequent - itemsets, Minimum Support, Confidence, Greedy Method, Vectorization.

I. INTRODUCTION

Association rules are one of the major techniques of data mining. Association rule mining finding frequent patterns, associations, correlations, or causal structures among sets of items or objects in transaction databases, relational databases, and other information repositories.

The techniques for discovering association rules from the data have traditionally focused on identifying relationships between items telling some aspect of human behaviour, usually buying behaviour for determining items that customers buy together. All rules of this type describe a particular local pattern. The group of association rules can be easily interpreted and communicated.

Apriori Algorithm is used to extract frequent itemsets from large database and getting the association rule for discovering the knowledge.

II. LITERATURE SURVEY

In this section author has discussed some research papers which had been previously undertaken in the field of association rule mining, greedy method and FP tree representation.

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 A mxn , then the corresponding BOOL data item set of item Ij(1<= j <= n)in P in Matrix Amxn is the mat of I_j, Mat_i 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 $MAT(I_j)$ and $MAT(I_k)$. i.e. $MAT(I_j,I_k)=MAT(I_j) * MAT(I_k)$.

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.

DOI: 10.15680/IJIRCCE.2016. 0406073

Goswami D.N., ChaturvediAnshu. RaghuvanshiC.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 ,can not 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 element's of candidate set, by using the intersect query of SQL. This approach requires very less time as compared to classical Apriori

Goswami D.N., ChaturvediAnshuRaghuvanshi 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 element's of candidate set, with the help of the intersect query of SQL. In this approach, they have applied a constraint 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.

III. APRIORI ALGORITHM

Apriori is a algorithm proposed by R. Aagarwal 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 prune 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.

IV. GREEDY ALGORITHM

A Greedy Algorithm is an algorithm that follows the problem solvingheuristic of making the locally optimal choice at each stage with the hope of finding a global optimum.Greedy algorithms work by recursively constructing a set of objects from the smallest possible constituent parts.

- **Orthogonal Greedy Algorithm**

Orthogonal matching pursuit (OMP) algorithm has received much attention in recent years. OMP algorithm is an iterative greedy algorithm that selects at each step the column[9]. Orthogonal matching pursuit (OMP) constructs an approximation by going through an iteration process. At each iteration the locally optimum solution is calculated. This is done by finding the column vector in A which most closely resembles a residual vector r. The residual vector starts out being equal to the vector that is required to be approximated i.e. r = b and is adjusted at each iteration to take into account the vector previously chosen. It is the hope that this sequence of locally optimum solutions will lead to the global optimum solution. As usual this is not the case in general although there are conditions under which the result will be the optimum solution. OMP is based on a variation of an earlier algorithm called Matching Pursuit (MP). MP simply removes the selected column vector from the residual vector at each iteration.

$$r_t = r_{t-1} - \langle aOP , r_{t-1} \rangle r_{t-1}$$

Where aOP is the column vector in A which most closely resembles r_{r-1}.OMP uses a least-squares step at each iteration to update the residual vector in order to improve the approximation. The OMP is a stepwise forward selection algorithm and is easy to implement.

V. VECTORIZATION

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 [10].

VI. PROCESS OF PROPOSED WORK

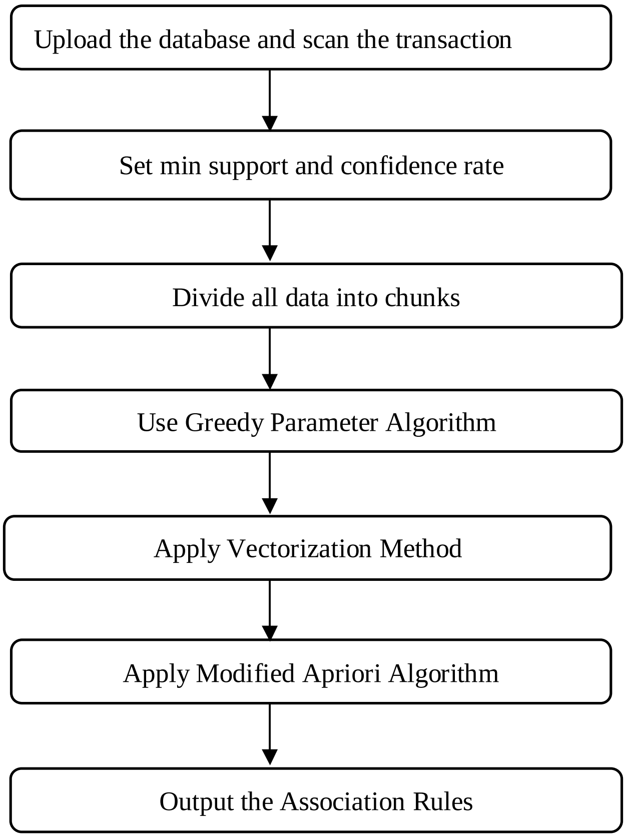
Upload the database and scan the transaction
Set min support and confidence rate
Divide all data into chunks
Use Greedy Parameter Algorithm
Apply Vectorization Method
Apply Modified Apriori Algorithm
Output the Association Rules

Figure 6.1 Proposed Modified Algorithm Process

The presentation of database is an efficient consideration in almost all algorithms. The most commonly database layout is the horizontal and vertical layout. In both layouts, the size of database is very large the data transformation is an essential process in data pre-processing step which can reduce the size of database. By reducing of the size of database can enhance performance of mining algorithms. Proposed method uses firstly a greedy-based data transformation technique to reduce the size of transaction database and then apply vectorization method to speed up algorithm.

VII. EXPERIMENTAL RESULTS

The modified algorithm was implemented using MATLAB by 4 random test cases used to evaluate the performance of the algorithm.

Time Comparison:

Table 7.1 shows the performance of execution time for different size of transactions per iteration for executing proposed modified Apriori algorithm and traditional Apriori.

Sr. No.	No of Transactions	Execution Time in Apriori (in seconds)	Execution Time in Modified Apriori (in seconds)	Percentage Improvement
1	100	1.067334	0.2152	80%
2	200	2.006108	0.4234	79%
3	500	5.172584	1.1544	78%
4	1000	10.157086	2.1014	79%

Table No 7.1 Transactions v/s Execution Time

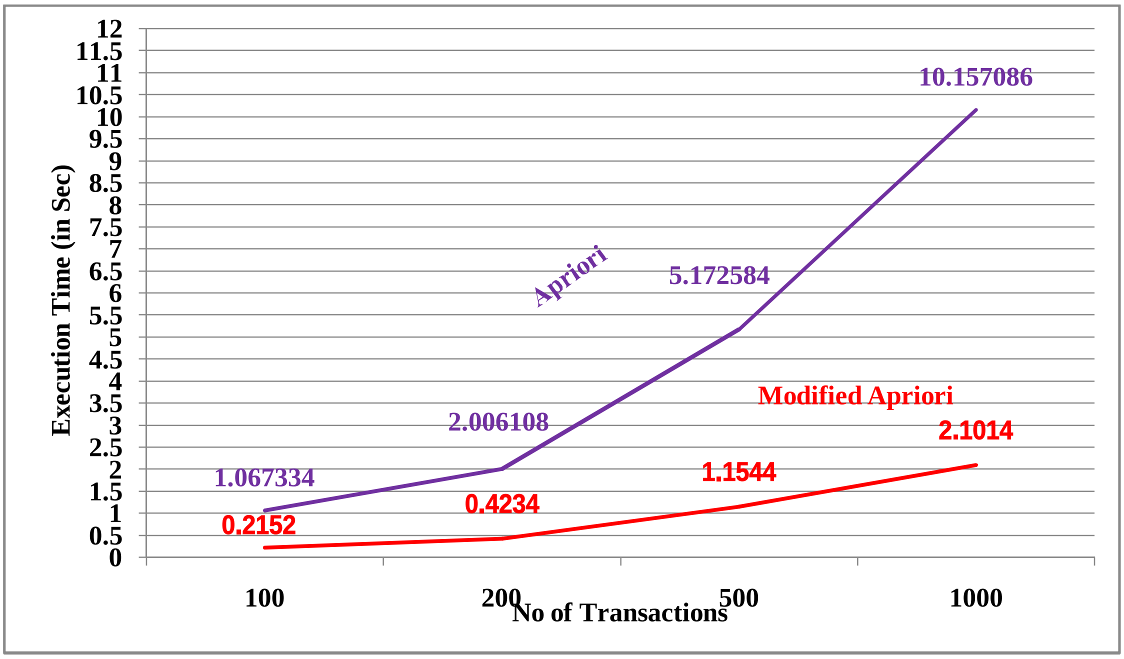

Figure: 7.1 No. of Transactions v/s Execution Time

VIII. CONCLUSION

A Modified Apriori is proposed by reducing the time consumed in transactions scanning for candidate itemsets and also by reducing the number of transactions to be scanned. Further, the numbers of rules generated are also reduced. The time consumed to generate candidate support count in modified Apriori is less than the time consumed in the traditional Apriori. Modified Apriori reduces the time consumed by 79 percentage. This factor is optimized by the greedy and vectorization approach for finding the frequent itemsets. Hence, this approach is far more efficient than the original Apriori algorithm.

REFERNCES

[1] Jiawei Han, MichelineKamber, "Data Mining, Concepts and Techniques", ISBN 978-81-312-0535-8, Elsevier India Private Limited, 2006.

[2] R. Agarwal. 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 Conferenceon 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] T. Tony Cai and Lie Wang,"Orthogonal Matching Pursuit for Sparse Signal Recovery With Noise", IEEE Transactions on Information Theory, Vol. 57, No. 7, July 2011.

[10] https://en.wikipedia.org/wiki/Vectorization.