Improved Apriori Algorithm and Association Mining By Using Market Basket Analysis

1st Md. Nour Noby
Dept. of CSE
Varendra University
Rajshahi, Bangladesh
nournoby6@gmail.com

2nd Md. Meraj Ali
Dept. of CSE,
Varendra University
Rajshahi, Bangladesh
meraj09034@gmail.com

3rd Hadisur Rahman
Dept. of CSE,
Varendra University
Rajshahi, Bangladesh
hudacse6@gmail.com

Abstract — **There are many mining algorithms of association rules. One in every of the foremost common algorithms is Apriori that's want to extract frequent itemsets from massive information and obtaining the association rule for locating the information. Supporting this algorithmic program, market basket analysis is allotted and build suggestion to the consumers to shop for an acceptable product or interested product. This paper indicates the limitation of the initial Apriori algorithmic program of dalliance for scanning the full information looking out on the frequent itemsets. Moreover, the initial Apriori algorithmic program removes some frequent itemsets which may be additional valuable in rule mining or market basket analysis. This presents AN improvement in Association Mining by reducing that wasted time counting on scanning just some transactions and additionally together with the information that is eliminated in victimization original Apriori algorithmic program in market basket analysis.**

Keywords— Data mining, apriori algorithm, association mining, market basket analysis.

I. INTRODUCTION

Data mining is the discovery of the hidden data found in databases and may be viewed as plenty of information discovery method. Data processing perform includes bunch, classification, prediction, and associations [1].

Apriori formula is one among the foremost vital algorithms that are employed to extract frequent itemsets from massive information, and obtain the association rule for locating the information. It essentially needs two vital things: minimum support, and minimum confidence. First, check whether or not the things area unit bigger than or capable the minimum support and realize the frequent itemsets severally. Secondly, the minimum confidence constraint is employed to make association rules.

Frequent pattern mining, association mining, correlation mining, association rule mining these area unit all connected, however distinct, ideas that are used for terribly a very long time to explain a facet information of knowledge mining that several would argue is that the very essence of the term data mining: taking a group of information and applying applied math ways to search out fascinating, and previously-unknown patterns among aforementioned set of information [2]. Association rule mining takes on 2 main steps. The primary step is to search out all itemsets with adequate supports, and also the second step is to get association rules by combining these frequent or massive item-sets [3, 4, 5]. Within the ancient association rules mining [6, 7] minimum support threshold, and minimum confidence threshold values area unit assumed to be offered for mining frequent itemsets, that is tough to be set while not specific knowledge; users have difficulties in setting the support threshold to get their needed results. Setting the support threshold overlarge, would turn out solely a tiny low range of rules or maybe no rules to conclude. Therein case, a smaller threshold price ought to be guessed (imposed) to try and do the mining once more, which can or might not provide a higher result, as by setting the edge too tiny, too several results would be made for the users, too several results would need not solely terribly very long time for computation however conjointly for screening these rules.

To use association rule for mining while not support threshold [8, 9, 10, 11] another constraint like similarity or confidence pruning is sometimes introduced. However, the synchronous itemset downside had not been directly thought-about by any of those researches. There is an area unit some researches that area unit relevant to the synchronous itemset downside, and projected a further life [12] so as to enhance the support-confidence framework.

II. METHODOLOGY AND ALGORITHM

A. Market Basket Analysis

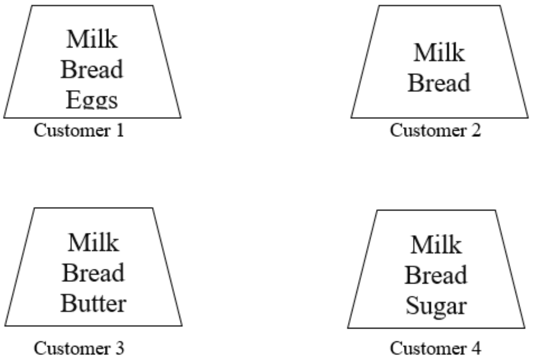


Fig. 1. Market Basket Analysis

Market Basket Analysis in Associate in the Nursing application of Frequent pattern mining. Frequent pattern mining is most simply explained by introducing market basket analysis (or affinity analysis), a typical usage that it's well-known. Market basket analysis tries to spot associations, or patterns, between the assorted things that are chosen by a selected shopper and placed in their market basket, be it real or virtual and assigns support and confidence measures for comparison. The worth of this lies in cross-marketing and client behavior analysis.

The generalization of market basket analysis is frequent pattern mining and is really so quite the same as classification except that any attribute, or combination of attributes (and not simply the class), will be foreseen in the association. An association doesn't need the pre-labeling of categories, it's a type of unattended learning.

B. Drawback of Apriori Algorithm

1. Calculating support is so much expensive because it has to go through the entire database again and again.
2. Sometimes, it may need to find a large number of candidate rules which can be computationally expensive.
3. Repeatedly scans the database for a large set of candidates by pattern matching, hence the original algorithm consumes more time and space.
4. Frequent item sets generation steps take more time for making all the possible combination of the item sets.

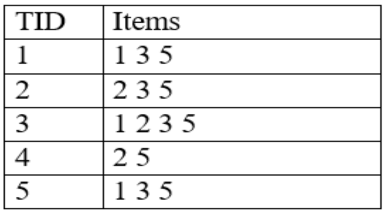
C. Improved Frequency Counting Process

Frequency counting process can be improved by getting the transaction id (TIDs) of the items and list them into a Table. We can easily get the frequency or number of transaction id (TIDs) from the table. It also reduces the database scanning process.

D. Frequency Counting of Item sets

At first, we need to scan the original transaction database and reorganized the storage structure of the database, that is reconstructed the former level structure of original number of transactions Tid and Item to the Item-Tid structure, then store it. A table [Table I.] is shown below. For simplicity a very small table is taken which contains only 5 transaction and the items can be within (1-5) which is considered as product id.

Table I. Transaction Table

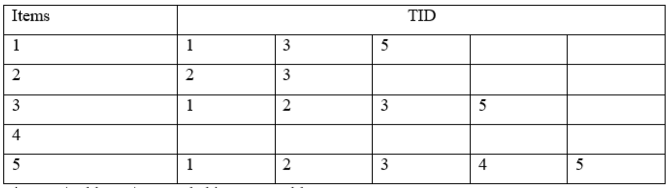


TID	Items
1	1 3 5
2	2 3 5
3	1 2 3 5
4	2 5
5	1 3 5

Then we scan transaction [Table I.] save the data to the recombined Item-Tid database, as shown in Table II, record the number of transactions per item in the scan process.

We have taken all the items and find out the TID where does they belong to. In this way we can make a table by scanning the database only once. From this step we will get a table as the result which contains all the transaction id (TIDs) associated to each item.

Table II. Recorded item sets table



Items	TID				
1	1	3	5		
2	2	3			
3	1	2	3	5	
4					
5	1	2	3	4	5

The number of TIDs represent the frequency of each items. By counting the number of transactions for each element we will get the frequency of each element.

For example,

Freq (1) = Count (Tid, 1) = 3
Freq (2) = Count (Tid, 2) = 2

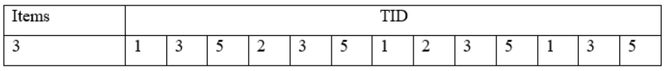
we also need frequency of a pair of elements such as freq. (X, Y). To calculate this frequency, we need to follow the further instructions.

E. Support Counting

We need to insert all the elements of transactions (TIDs) for each element which are associated to that items in a vector. Then we sort the list in any order either ascending or descending and we can easily get the frequency of the items related to the specific item.

For example, let consider item 3. Item 5 was in all the transaction except 4. (TIDs 1, 2, 3, 5). If we insert all the elements of TIDs (1, 2, 3, 5) we will get a list similar to the following one.

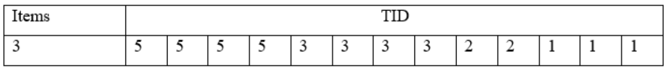
Table III. associated item sets table



Items	TID												
3	1	3	5	2	3	5	1	2	3	5	1	3	5

After sorting in descending the list will look similar to the following one.

Table IV. sorted item sets table for counting frequency



Items	TID												
3	5	5	5	5	3	3	3	3	2	2	1	1	1

From these data, we can calculate the support and confidence by applying the equation.

$$\text{Supp}(XY) = \text{freq}(X, Y) / N = \text{freq}(3, 5) / 5 = 4 / 5$$

In the similar way we can calculate all the support such as supp (3,2), supp (3,1).

F. Confidence Counting

Confidence can be calculating in the following way from the vector and previously counted frequency of the single elements.

$$\text{Conf}(X \rightarrow Y) = \text{freq}(X, Y) / \text{freq}(X)$$

$$\text{Conf}(3 \rightarrow 5) = \text{freq}(3,5) / \text{freq}(3)$$

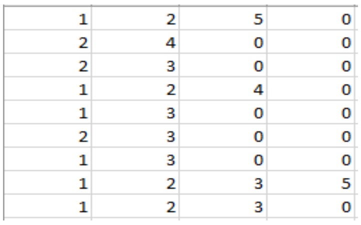
$$\text{Conf}(3 \rightarrow 5) = 4 / 4$$

Since we can calculate the support and the confidence of items within single step at the same time according to the table [Table IV.] we can also generate the rules more specifically association rules. There we require minimum support and minimum confidence which are user specified. We will compare with the threshold value of minimum support and minimum confidence and generate the association rules.

G. Experimental Result

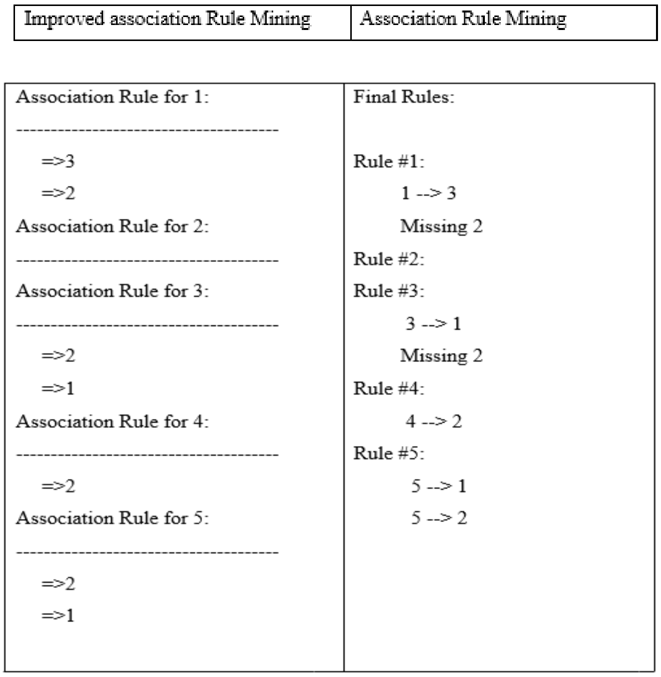
In first case we have taken a table of 9 transaction. Each transaction contains minimum of 5 data or items. Minimum support has been taken 0.22 and minimum confidence has been taken 0.6.

H. Sample data and Result



1	2	5	0
2	4	0	0
2	3	0	0
1	2	4	0
1	3	0	0
2	3	0	0
1	3	0	0
1	2	3	5
1	2	3	0

Fig. 2. Transaction table



Improved association Rule Mining	Association Rule Mining
Association Rule for 1: -------------------------------------- =>3 =>2 Association Rule for 2: -------------------------------------- Association Rule for 3: -------------------------------------- =>2 =>1 Association Rule for 4: -------------------------------------- =>2 Association Rule for 5: -------------------------------------- =>2 =>1	Final Rules: Rule #1: 1 --> 3 Missing 2 Rule #2: Rule #3: 3 --> 1 Missing 2 Rule #4: 4 --> 2 Rule #5: 5 --> 1 5 --> 2

Fig. 3. Result from transaction data

III. DISCUSION

Each time we perform the experiment, we have gotten the better result than previous algorithm. This algorithm produces better result and better data. A lot of new data have been found from this process since no valuable data have been during performing the steps of the algorithm.

Consider the following case of the Apriori algorithm steps of generating largest list of frequent itemset.

Item Bought	Support
Milk, Tea	5
Eggs, Tea	2
Eggs, Cold Drink	3
Tea, Cold Drink	2

Fig. 4. Frequent Itemset Table

Item Bought	Support
Eggs, Tea, Cold Drink	2

Fig. 5. Frequent Itemset Table

To generate the largest frequent itemset we've to create the mixture of the itemsets and calculate support. Then we'd like to check with the minimum support and eliminate the sets that square measure but the edge minimum support. However, some smaller itemsets that have such a lot higher support worth than the minimum support worth is lost within the method of Apriori rule.

In the on top of example, we will see, once we attempt to generate the largest list of frequent itemset, some itemset is lost which may be additional valuable in the data processing. Here in [Fig. 4.] has support five. However, within the next step of generating larger candidate set and scheming the frequency and comparison with minimum frequency, the smaller set has lost. Additional, frequent itemsets are found in Improved Association Rule Mining. No valuable frequent knowledge is going to be lost that we've seen in apriori rule and association rule mining.

IV. CONCLUSION

The most vital tasks of frequent pattern mining approaches are itemset mining, successive pattern mining, successive rule mining, and association rule mining. An honest variety of the economic data processing algorithms exist within the literature for mining frequent patterns. With over a decade of in-depth analysis, an honest variety of analysis publications, development and application activities during this domain are planned. This paper presents the in depth of study of Association Rule Mining algorithmic program in data processing that is widely used, and extremely abundant required in market basket analysis. The manner within which this algorithmic program works is sort of simple; it computes all of the foundations that meet minimum support and confidence values. The quantity of attainable potential rules will increase exponentially with the number of things within the itemsets.

Existing algorithms for mining association rules developed up to now are supported by Breadth-First Search (BFS) approaches. In the BFS, the frequent itemsets are generated level by level. The information is scanned at every level to work out the support price for every generated candidate itemsets. This may increase execution time and memory consumption. However, the planned algorithmic program performs higher. This may work fine on smaller datasets compared to larger datasets. The performance of such algorithms is additionally smart at larger support values.

The main focus of this analysis work is to enhance the prediction rate of the Inter-transaction association rules, scale back the execution time for locating the association rules, scale back the search house and generate stronger, useful, abstract and purposeful association rules.

To obtain more reduction in execution time and memory house consumption, the conception of compressed steps association rule mining within the planned approach. There are two steps during this approach, within the commencement, it scans the information once and finds the group action IDs. Within the second step, it calculates frequency, support, confidence and generates association rule by comparison with minimum support and minimum confidence threshold values.

VIII. REFERENCES

[1] Jiawei Han and Micheline Kamber, Data Mining Concepts and Techniques, Morgan Kaufmann Publishers, 2001.

[2] https://www.kdnuggets.com/2016/10/association-rule-learning-concise-technical-overview.html

[3] H. Mahgoub, “Mining association rules from unstructured documents" in Proc. 3rd Int. Conf. on Knowledge Mining, ICKM, Prague, Czech Republic, Aug. 25-27, 2006, pp. 167-172.

[4] S. Kannan, and R. Bhaskaran "Association rule pruning based on interestingness meas ures with clustering". International Journal of Com puter Science Issues, IJCSI, 6(1), 2009, pp. 35-43.

[5] M. Ashrafi, D. Taniar, and K. Smith "A New Approach of Eliminating Redundant Association Rules". Lecture Notes in Computer Science, Volume 31S0, 2004, pp. 465 -474.

[6] P. Tang, M. Turkia "Para llelizing frequent itemset mining with FP-trees". Technical Report titus.compsci.ualr.edu/-ptang/papers/par-fi.pdf, Department of Computer Science, University of Arkansas at Little Rock, 2005.

[7] M. Ashrafi, D. Taniar, and K. Smith "Redundant Association Rules Reduction Techniques". Lecture Notes in Computer Science, Volume 3S09, 2005, pp. 254-263.

[8] M. Dimitrijevic, and Z. Bosnjak "Discovering interesting association rules in the web log usage data". Interdisciplinary Journal of Information, Knowledge, and Management, 5, 20I 0, pp.191-207

[9] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, and A. I. Verkamo.: Fast discovery of association rules- In Advances in Knowledge Discovery and Data Mining (1996).

[10] Z. HONG-ZHEN, C. DIAN-HUI, and, Z. DE-CHEN "Association Rule Algorithm Based on Bitmap and Granular Computing". AIML Journal, Volume (5), Issue (3), September, 2005.

[11] K. Yun Sing "Mining Non-coincidental Rules without a User Defined Support Threshold". 2009.

[12] C. Yin-Ling and F. Ada Wai-Chee "Mining Frequent Itemsets without Support Thres hold: With and without Item Constraints". 2004.