A Review on Association Rule Mining and Improved Apriori Algorithms

Ms. Arti Rathod' Mr. Ajaysingh Dhabariya and Mr. Chintan Thacker

Abstract— Association rule mining is the most important and well researched techniques of data mining. It aims to extract interesting correlations, rules, frequent patterns and associations among sets of items in the transactional databases. Decision making and understanding the behavior of the customer has become challenging problem for organizations, so one of the data mining analysis technique is introduced which is called Market Basket Analysis. Apriori is the classical algorithm for learning association rules. This algorithm finds the frequent pattern based on support and confidence measures. Support and Confidence are two measures which limit the generated levels. It's a simple algorithm but having many drawbacks .Many researchers have been done improvement on this algorithm. This paper shows a Survey on improved approaches of Apriori algorithm.

Keywords— **Association, Apriori, Market Basket Analysis, Support, Confidence.**

I. INTRODUCTION

The main Purpose of data mining is to disclose the hidden information from the database[1].Due to the growth of data volume in an organization sectors like banking, marketing, telecommunication, manufacturing, transportation etc, a different technique for deletion of repetitive data and conversion of data to more usable forms has been proposed under data mining[19]. Data mining also known as knowledge discovery is used to discover useful patterns from the database. Many techniques have been developed in data mining amongst which association rule mining is very important. Apriori is one of the best algorithms for the association rule mining. The Apriori algorithm Discover the frequent patterns from database whose support and confidence must satisfy the minimum support and confidence.

II. ASSOCIATION RULE MINING

In Data Mining Association rule learning is a method for discovering interesting relations between variables in large database. Association rule discovers interesting association/correlation among a large set of data items. [3] the sales of Super market would indicate that if a customer buys onions and potatoes together, he or she is likely to also buy burger. This information will help business to know the behavior of the customers. Shopping centers also uses the association rule mining to place the items next to each other so that user buy more items .Another application of Association mining is the goggle auto complete, where we type a word it searches frequently associated words that user type after that particular word.

III. CONCEPTS OF ASSOCIATION RULE MINING

Support: how many transactions have such itemsets that match both sides of the implication in the association rule? (If x and y are two items in database then both comes together). [2]

Support(X, Y) = n (XUY)/N

N=Total no .of transactions.

Confidence: Ttransactions that contain *X* also contain *Y. [2]*

Confidence(X, Y) = support (XUY)/support(X)

Item: It is a field of transactional database.

Consider the following Transactional database Table-I:

TABLE I
TRANSACTIONAL DATABASE

Transaction Id	Milk	Bread	Butter
1	1	1	0
2	0	0	1
3	0	0	0
4	1	1	1
5	0	1	0

In Table I, 1 represent the presence of item and 0 represent the absence of items. Now let's count the support and confidence.

Consider X=milk and Bread, Y = Butter.

Support {milk, Bread}→{ Butter} =Support(X→Y)
=1/5
=0.2(20%)

Confidence{Milk,Bread} → {Butter} =Confidence(X→Y)
=0.2/0.4
=0.5(50%)

Support says that milk butter and bread all purchased together while confidence says that whenever milk and bread purchased there is also possibility of butter.

IV. MARKET BASKET ANALYSIS

Market Basket Analysis is the best example of association rule mining. This analysis identifies the buying behaviour of the customer among various items that customer places in their shopping baskets [4]. The identifications of such customer's behaviour can assist retailers to gain the marketing strategies to gain the profit into business. Market Basket analysis is the technique to derive associations between datasets [4]. Let's take Example to derive association rules.

Ms. Arti Rathod is with Department of Computer Science, Shrinathji Institute Of technical Education, Rajasthan Technical University, Nathdwara, India, artirathod@ymail.com

Let us call the items currently seen by the customer as X and Y is the item associated with the current item(X).If we have 2 items namely P and Q then the possible association rules are only two [4]: P→Q and Q→P. If we have 3 items P, Q and R as follow (Table II).Then we will have 12 possible association rules (Table-III).

TABLE II
TRANSACTIONAL DATABASE

Transaction Id	P	Q	R
1	1	1	0
2	0	0	1
3	0	0	0
4	1	1	1
5	0	1	0

Based on table-2 we can derive the association rule using market basket analysis as: 1) Generate all possible association rules 2) Computer the support and confidence of all possible association rules. 3) Apply two threshold criteria: Minimum support and Minimum confidence. [6]

TABLE III
COMBINATION OF PURCHASED ITEMS

	X	Y	Support	Confidence
1	P	Q	0.4(40%)	0.4(40%)
2	P	R	0.2(20%)	0.5(50%)
3	P	Q,R	0.2(20%)	0.5(50%)
4	Q	P	0.4(40%)	0.66(66%)
5	Q	R	0.2(20%)	0.33(33%)
6	Q	P,R	0.2(20%)	0.33(33%)
7	R	P	0.2(20%)	0.5(50%)
8	R	Q	0.2(20%)	0.5(50%)
9	R	P,Q	0.2(20%)	0.5(50%)
10	P,Q	R	0.2(20%)	0.5(50%)
11	P,R	Q	0.2(20%)	1(100%)
12	Q,R	P	0.2(20%)	1(100%)

From the table-III we can give the threshold value to the support and confidence for getting association rule. so let's give the minimum value to support 30% means the frequency of the item X and Y Buy together and minimum value to the confidence 60% means the frequency of the transaction when customer buy item X also buy item Y.

From table-III the only transaction no-4 (Table IV) has the higher then the threshold value of support and confidence. Means the product in the transaction-4 are certainly purchased by the customers.

TABLE IV
OBTAINED TRANSACTIONS AS A ASSOCIATION RULE

	X	Y	Support	Confidence
1	Q	P	0.4(40%)	0.66(66%)

X is the Combination of items up to n-1 where n is the number of items. Y is the combination of the set difference between all items and items listed on the Y.

V. Apriori Algorithm

Apriori is one of the association rule mining algorithm which is used to discover all frequent itemsets from transactional database [6]. This algorithm uses prior information of frequent item set properties that is why it is known as Apriori algorithm [5]. To understand the Apriori algorithm we need to understand the definition of following terms:

Itemset: It's a collection of items in a database.

Transaction: It's a database entry which contains s a collection of items.

Support: Interesting association rule can be measured with the help of support criteria. Support is nothing but how many transactions have such itemsets that match both sides of implications in the association rule.[8]

Support (i) =Count (i)/total transaction

Candidate Itemset (L_i): Items which are only used for the processing. Candidate itemsets are all possible combination of itemsets. [9]

Minimum Support: It's a condition which helps to eliminate the non frequent item s from database. [9]

Frequent Itemset (Large Itemset (L_i)): The itemsets which satisfies the minimum support criteria are known as frequent itemsets. [9]

Apriori uses iterative approach known as breadth first search (level-wise search), where k-1 itemsets are used to generate k itemsets. Apriori states that "All nonempty subsets of a frequent item sets must be frequent". Apriori algorithm works on two concepts: [7]

(i) Joining and
(ii) Pruning

Apriori Algorithm Steps [9]:

(i) First, the set of frequent 1-itemsets is found. (Known as C_1).
(ii) Then support is calculated by counting the occurrence of the item in transactional database.
(iii) After that we will prune the C1 using minimum support Criteria. The item which satisfies the minimum support criteria is taken into consideration for the next process and which is known as L1.
(iv) Then again candidate set generation is carried out and the 2-itemset which is generated known as C_2.
(v) Again we will calculate the support of the 2-Itemset.and we will prune C2 using Minimum support and generate L2.
(vi) This Process Continues till there is no Candidate set and frequent itemsets can be generated.

Let's consider one example to understand the concept of Apriori algorithm: Table v shows transactional database having 4 transactions.

TABLE V

Transaction	Items
1	P,R,S
2	Q,R,T
3	P,Q,R,T
4	Q,T

Performing first step by scanning database to identify the number of occurrence of specific item. After that we will get C_1 as shown in Table VI below:

TABLE VI

Itemset	Support
P	2
Q	3
R	3
S	1
T	3

The next step is pruning in which we will consider the minimum support criteria=2. The items which does not have minimum support Criteria will be eliminated. And we will get L_1.Table VII shows the pruning step.

TABLE VII

Itemset	Support
P	2
Q	3
R	3
T	3

Now the candidate generation step is carried out and 2-itemset candidates are generated this is denoted as C_2. (Table VIII).

TABLE VIII

Itemset	Support
P,Q	1
P,R	2
P,T	1
Q,R	2
Q,T	3
R,T	2

Now pruning has to be done by considering minimum support criteria=2 and then we will get L_2. (Table IX)

TABLE IX

Itemset	Support
P,R	2
Q,R	2
Q,T	3
R,T	2

Again we will generate candidate set $C_{3.}$ (Table X)

TABLE X

Itemset	Support
P,Q,R	1
P,Q,T	1
Q,R,T	2

Now pruning (minimum Support criteria=2) has been done to get L_3 As in (Table XI)

TABLE XI

Itemset	Support
Q,R,T	2

As we can see in Table XI the frequent items are Q, R, and T.

Pseudo Code [9]:

```
$C_k$ : Candidate itemsets of size k
$L_k$: Frequent itemsets of size k
$L_1$ = {Frequent items};
For (k=1; $L_k$! =null; k++) do begin
    Increment the count of all candidates in $C_{k+1}$
$L_{k+1} =$ Candidates in $C_{k+1}$ with minimum support
End
```

Apriori algorithm is the classical and simplest algorithm to implement the concept of association rule mining. But there are some disadvantages as follow:

Drawbacks of Apriori Algorithm [10]:

1. It takes too much time to scan the database.
2. It generates large number of in-frequent itemsets which Increase the space complexity.
3. Generates large amount of frequent itemsets which are Not Efficient.
4. It needs several iterations for mining data.
5. Treats all the items in database equally by considering Only the presence and absence of an item within the Transaction. It does not take into account the Significance of item.

Many researchers introduced several improved Apriori algorithm to remove the limitations of Apriori algorithms as follow:

VI. Improved Apriori Algorithms

A. *Intersection Approach:*

Intersection algorithm is designed to improve the efficiency, memory management and remove the complexity of Apriori .in this approach , to calculate the support ,count the common transaction that contain in each element's of candidate set.[11][12][13]

B. *Record Filter Approach:*

In the classical Apriori algorithm ,we check the occurrence of candidate item in each transaction of any length .In this approach 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 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.[11][12][13][14].

C. *Set size frequency approach:*

In this approach set size is the number of items per transaction and set size frequency which is the number of transactions that have at least set size items. Initially database is given with set size and second database is of set size frequency of the initial database .Removes items with frequency less than the minimum support value initially and determine initial set size to get the highest set size whose frequency is greater than or equal to minimum support of the set size. Set size which are not greater than or equal to min set size support are eliminated [11][12][15].

D. *Utilization of attribute approach*

Association rule mining treats all the items in database equally by considering only the presence and absence of an item within the transaction .it does not take into account the significance of item to user or business. This limitation can be removed by using attributes like profit, weight and quantity [11][12][16].

E. *Apriori based on Matrix approach:*

Matrix approach contains a binary representation in which 1 represent the presence of item and 0 represent the absence of an item in database. So by counting the number of 1's in the matrix we can easily find the occurrence of that item. For 2-itemset we can multiply the binary representation of the item to get the occurrence of that item together. This approach needs to scan the database only once and does not require finding the candidate set when searching for frequent item set [9].

F. *Apriori based on Interest Itemset:*

In this approach the candidate set can be reduced and also the speed of the algorithm is accelerated .this algorithm is based on interest measures. There are some constraints on the selection of interest Itemset like selection of interest items is done by users. This algorithm is very efficient in terms of database scans [9].

G. *Reducing candidate set and memory utilization Approach:*

This algorithm [18] introduces a more efficient way to achieve the pruning operation. The algorithm only needs to search Lk-1 one time to complete the deletion and the remaining of each element X in Ck. The idea of the algorithm is as follows. Ik is a k-dimensional itemsets. If the number of (k-1)-dimensional subsets of all (k-1)-dimensional frequent itemsets Lk-1, which contains Ik, is less than k, then Ik is not a k-dimensional frequent itemsets..So the improved algorithm only needs to match up the count of each element of Lk-1with the count of each element (X) of Ck (each element X has a count). If the count of the element X equals to k, then keep X. Otherwise X must be deleted.

I/O speed can be deduced by cutting down unnecessary transaction records. The item that not appears in Lk-1 will no longer appear in Lk. So we can revise these items to null in the transaction database. Then we can pay no attention to these data information in any search work to D. At the same time, delete the transaction records (T) of which the number of valid data is less than k so as to deduce the database [13]. Then the candidate set Ck will be generated by latest D. The deletion of D will greatly reduce the number of transaction records which will effectively increase the speed of the implementation of the algorithm. Ultimately this will increase efficiency and I/O speed of algorithm.

H. *Apriori based on frequency of items approach:*

In this approach, first frequent patterns are discovered from the transactional database using the Apriori algorithm. From the frequent patterns mined, this approach extracts novel interesting association patterns with emphasis on significance, quantity, profit and confidence. To overcome the weakness of the traditional association rules mining, weighted association rule mining has been proposed. Weighted association rule mining considers both the frequency and significance of itemsets. It is helpful in identifying the most precious and high selling items which contribute more to the company's profit.

This approach proposes an efficient idea based on mainly weight factor and utility for mining of high utility patterns. Initially, the proposed approach makes use of the classical Apriori algorithm to generate a set of association rules from a database [17].

Firstly it uses attributes to get frequent itemsets. These attributes are like profit ratio calculation using Q-factor.

$$\text{Q-Factor}=P/\,\Sigma P_i \qquad (1)$$

Than it gives transactional database where each item's frequency is counted in each transaction. From that pruning is done with minimum support and confidence .Finally calculation of Weighing-factor is done based on frequency of itemsets and Q-factor.

$$\text{PW0} = \sum_{i=1}^{100} \text{Frequency} * \text{Q-Factor} \qquad (2)$$

Finally efficient frequent pattern is selected based on minimum PW-factor.

VII. Conclusion And FUTURE WORK

Association rule mining is used to discover useful patterns from transactional database, and Apriori algorithm is used to implement the association rule mining. but this classical algorithm has several limitations like scanning time, memory optimization, candidate generation which can be solved by several improved Apriori approaches like record filter approach, intersection approach ,matrix based approach, set size frequency approach, interest item approach .

This classical Apriori treats all the items in database equally by considering only the presence and absence of an item within the transaction .it does not take into account the significance of item to user or business. So, Apriori algorithm efficiency can be improved by using quantity, profit attributes and support count which will give the valuable information to customer as well as business.

ACKNOWLEDGEMENTS

I am thankful to my advisor and mentor Mr. Ajaysingh Dhabariya, Asst. Prof, S.I.T.E, Nathdhwara, and my External Guide Mr. Chintan Thacker, Asst. Prof, HJD Institute of Technical Education and Research, for inspiring and guiding me throughout my project work in Association rule mining with Apriori algorithm.

REFERENCES

[1] "The Survey on Data Mining Algorithm for market basket analysis." , Dr. Dhanabhakyam, Dr.M.Punithavalli, Dr.SNS college of Arts and Science, Global Journal of Computer Science and Technology (IJCSEIT), Vol.11,Issue 11 Version 1.0, July 2012,ISSN:0975-4172.

[2] "Association Model for market basket analysis, customer behaviour analysis and business intelligent solution embedded with Apriori concept",J.M Lakshmi ,Mahesh,SCMS school of technology and management Muttom, cochin, Kerala, International journal of research in finance & marketing ,vol 2,Issue 1, January 2012 ISSN:2231-5985.

[3] "Association Rule Mining as a Data Mining Technique",Irena Tudor, Universitatea Petrol-Gaze din ploeiesti, Bd.Bucuresti 39, ploeiesti,Catedra de Informatica, Vol-LX,No.1,2008.

[4] "Association Rule- Extracting Knowledge Using Market Basket Analysis", Raorane A.A ,Kulkarni R.V and Jitkar B.D,Dept. of Computer Science,Vivekanand College, Tarabai Parkkolhapur, Research Journal of Recent Sciences, Vo11(2)19-27,Feb. 2012.

[5] "Mining Efficient Association rules through Apriori algorithm using attributes and comparative analysis of various Asssociation rule algorithm", Ms Shweta, Dept. of Computer Science and application, Kurukshetra University, Kurukshetra, India, International Journal of Advance Research in Computer Science and software engineering, Vol-3,Issue-6, June
2013.

[6] "Mining Sequqntial Patterns", P.S.Yu and A.S.P Chen, Agrawal ,R. And Srikant R. 1995. Eds. In IEEE Computer Society Press, Ta ipei, Taiwan, 3{14}

[7] "Ranking and Suggesting Popular Itemsets In Mobile Stores Using Modified Apriori Algorithm", P V Vara Prasad, Sayempu Susmitha, Badduri Divya, Gogineni Riharika, Guntur Vijya Raghu Ram., International Journal of Modern Engineering Research(IJMER),Vol 2,Issue 1,Jan-Feb2012,pp-431-435.

[8] "An improved Apriori based algorithm for association rule mining" ,Haun Wu Zhigang Lu,Lin Pan,Rongsheng Xu,Computer center,Institute of High energy physics,chinese academy of sciences,Beijing100049,China, 2009-Sixth international conference on Fuzzy Systems and knowledge Discovery IEEE Exlopre.

[9] "Improved Apriori Algorithms-A Survey",Pranay Bhandari, K.Rajeswari,Swati Tonge, MahadevShindalkar, Dept. of Computer Engineering , Pimpri Chinchwad College of Engineering Pune, Maharastra India, International Journal of Advance Computational Engineering and Networking,ISSN:2320-2106,Vol-1,Issue-2-2013.

[10] "Drawbacks and solutions of applying association rule mining in learning management system", Enrique Garcia, Cristobal Romero, Sebastian Ventura, Toon Calders , Cordoba University, campus Universitario de Rabanales,14071,Cordoba,spain,Eindhoven university of Technology (TU/e),Netherlands, International workshop on Applying datamining in E-learning 2007.

[11] "A review Approach on various form of Apriori with association rule mining", Ms.Pooja Agrawal,Mr.Suresh Kashyap,Mr.Vikas Chandra Pandey,Mr. Suraj Prasad Keshri,International Journal on Recent and Innovation Trenda In Computing and Communication.volume-1,Issue-5,May2013.

[12] "Survey on several Improved Apriori algorithm", Ms.Rina Raval, Prof Indr jeet Rajput, Prof Vinit kumar Gupta, Dept of Computer Engineering, H.G.C.E, vahelal, Ahmedabad, Gujarat, India, IOSR Journal of Computer Engineerring, ISSN:2278-0661 Vol-9,Issue-4,Mar-Apr.2013.

[13] "An Algorithm For Frequent Pattern Mining Based On Apriori", D.N Goswami, Anshu Chaturvedi, and C.S Raghuvanshi, International Journal of Computer science and Engineering,Vol-02,No.-4,2010,942-947,ISSN:0975-3397.

[14] " Frequent Pattern Mining using Record Filter Approach", D.N Goswami, Anshu Chaturvedi, and C.S Raghuvanshi, International Journal of Computer science issues Vol-7,Issue-4,No 7,July 2010.

[15] "Association Rule Mining based on Apriori algorithm in minimizing candidate generation", Sheila A.Abaya, International Journal of Scientific & engineering Research, Vo1-3,Issue-7,July 2012,ISSN 2229-5518.

[16] "Mining Efficient Association rules through Apriori algorithm using attributes" ,Mamta Dhanda, Sonali Guglani, Gaurav Gupta ,Dept. of Computer Science ,RIMT-IET, Mandi-Govindgard, Punjab,India, , International Journal of Advance Computer Science and Technology,Vol-2,Issue-3,September 2011.

[17] "An approach to extract efficient frequent patterns from transactional database", Mamta Dhanda ,Dept. of Computer Science ,RIMT-IET, Mandi-Govindgard, Punjab Technical University Jalandhar, Punjab, India ,International Journal of Engineering Science and Technology,Vol-3,No 7,ISSN:0975-5462, July 2011.

[18] "An improved algorithm for Apriori" ,Zhang Changsheng,Li Zhongyue, Zheng Dongsong, In;IEEE,first international workshop on education technology and computer science, March 07-March 08,ISBN: 978-0-7695-3557-9

[19] "Data Mining Concepts and Techniques", Jiawei Han and Michelin Kamber-book second edition.