A Survey on Apriori Algorithm

Ankita M Parmar
PG Student, Computer Department
VVP Engineering College
Rajkot , India
Ankita.parmar2188@gmail.com

Kamal K Sutaria
Assistant Professor
VVP Engineering College

Krutarth V. Joshi
Systems Engineer, BI COE
TATA Consultancy Services (TCS)
Pune , India
joshi.krutarth@gmail.com

***Abstract*— Apriori algorithm is a one of the simple and famous data mining technique used for pulling out hidden patterns from data. In this paper we talk about the practical problems which are embroiled in the different fields in the real world entities. We had discussions on the number of articles which including various problem during the process.**

***Index Terms*— Data mining , Association rule , Apriori Algorithm , frequent itemsets , minimum support.**

I. INTRODUCTION

Data mining is the process of analyzing the data from different perspectives and going over the useful information – information that can be used to increase revenue, cuts costs, or both[10]. Precisely data mining is the process of finding correlation or patterns among dozens of fields in large relational database. One of the most important data mining applications is that of mining association rules. Data mining has many virtues and vices which involves in many fields. Some of the examples are (1) Bank – to identify patterns that help be used to decide result for loan application to the customer, (2) Satellite research – to identify potential undetected natural resources or to identify disaster situations , (3) Medical fields – to protect the patients from infectious diseases , (4) Market strategy – to predict the profit and loss in purchase. Data mining functions include clustering, classification, prediction, and link analysis (associations). One of the popular techniques used for mining data is knowledge discover database for pattern discovery is the association rule. It implies certain association relationships among a set of objects. An algorithm for association rule induction is the Apriori algorithm, proven to be one of the popular data mining techniques used to extract association between various item set among large amount of data. Many algorithms come under association rule mining but Apriori algorithm is one of the typical algorithms. The rules produced by Apriori algorithm makes it easier for the user to understand and further apply the result. It was introduced by Agarwal in 1993; it is a strong algorithm which helps in finding association between itemsets. A basic property of apriori algorithm is “every subset of a frequent item sets is still frequent item set, and every superset of a non-frequent item set is not a frequent item set”. This property is used in apriori algorithm to discover all the frequent item sets. Further in the paper we will see more about the Apriori algorithm steps in detail.

II. ASSOCIATION RULES

Association rules are used to unearth relationships between apparently unrelated data in a relational database[13].It is having two important things support and confidence. Support is the number of transactions in which the association rule holds[14]. It is the percentage of transactions that demonstrate the rule. Suppose the support of an item is 0.4%, it means only 0.4 percent of the transaction contain purchasing of this item.

Support(AB) = Support count of(A ⋃B)/Total number of transactions in database

Confidence is the conditional probability that ,given A present in transaction, B will also be present.

Confidence(AB) = Support count of (A ⋃ B) / Support(A)

A. *Positive Association Rules*

The normal convention in discovering the association rules is by means of any frequent item sets that are present in the given transactional database. The rules that are normally obtained by means of using minimum support threshold and minimum confidence threshold are generally referred as the positive association rules and the rule is of the form ¬A ¬B. That means that they are capable of associating one element to the other element in a given set of transactional records.

B. *Negative Association Rules*

Contrary to the positive association rules described above, negative association rules are defmed as the rule that involves the absence of item sets. For example, consider A => ¬B, here , “¬”, indicates the absence of an item set B in aset of given transactional records. The rules of the forms (A□ ¬B, ¬A□ B and ¬A□¬B) are negative association rules

C. *Constraints based association rule mining*

In an interactive mining environment, it becomes a necessity to enable the user to express his interests through constraints on the discovered rules, and to change these interests interactively. The most famous constraints are item constraints, which are those that impose restrictions on the presence or absence of items in a rule. These constraints can be in the form of conjunction or a disjunction. Such constraints have been introduced first in where a new method, for incorporating the constraints into the candidate generation phase of the Apriori algorithm, was proposed. In this way, candidates are assured to obey the Item constraints besides the original support and confidence constraints.

Item constraints are not only possible type of rule constraints. Ng et al. [15] presented a wide range of constraints on rules that extends from relational operators on values of the items to constraints on the value of some aggregate functions calculated on the rule items. They defined what is called Constrained Frequent Queries (CAQs) (later named Constrained Frequent-set Queries in [16]) and presented an excellent classification of constraints constructs that can be exploited in them by introducing the notions of succinct and anti-monotone constraints. The CAP (Constrained APriori) waspresented for efficient discovery of constrained association rules.

The aim of association rule is to discover all association problems having support and confidence not less than the given value of threshold. If the support and confidence of item set of database is less than minimum support and confidence than that item set is not frequent item set[7].

In Association Rule mining find rules that will predict the occurrence of an item based on the occurrence of the other items in the transaction. Table shows Market-Basket Transactions

Table I: Market – Basket Transactions

TID	**Items**
1	Bread , Milk
2	Bread , Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk , Diaper, Beer
5	Bread, Milk, Diaper, Coke

Example of Association Rules:
{Diaper} → {Beer},
{Bread, Milk} → {Egg, Coke},
{Bread, Beer} → {Milk},

Implication means co-occurrence, not causality.

Association rule is an implication expression of the form X →Y, where X and Y are itemsets.
Example: {Milk, Diaper} → {Beer}

Rule Evaluation:

- *Support (S):* Fraction of transactions that contain both X and Y.
- *Confidence (C):* Measures how often items in Y appear in transactions that contain X.
  Example:
  {Milk, Diaper} → {Beer}
  $S = \sigma(\{Milk, Diaper, Beer\}) / [T]$
  $S = 2/5$ $S = 0.4$
  $C = \sigma(\{Milk, Diaper, Beer\} / \sigma(\{Milk, Diaper\}$
  $S = 2/3$ $S = 0.67$
- *Itemset:* A collection of one or more items.
  Example : {Milk, Diaper, Beer}. k-itemset that contains k-items.
- *Frequent Itemset:* An itemset whose support is greater than or equal to a min_sup threshold. In association rule mining task from a set of transactions T, the goal of association rule mining is to find all rules having Support >= min_sup threshold and Confidence>= min_conf threshold.

There are two phases in the problem of data mining association rules.

1) Find all frequent itemsets: i.e. all itemsets that have support s above a predetermined minimum threshold.
2) Generate strong association rules from the frequent itemsets: these association rules must have confidence c above a predetermined minimum threshold.

After the large item sets are identified, the corresponding association rules can be derived in a relatively straightforward manner. Thus the overall Performance of mining association rules is determined primarily by the first step. Efficient counting of large itemsets is thus the focus of most association rules mining algorithms.

III. ISSUES IN FINDING ASSOCIATION RULES

A. *Minimum Support Threshold*

Many algorithms such as apriori, FPtree etc. use this minimum support threshold in finding the frequent item sets. This threshold value is pre-set by the users. This value is set by user only. When user set high threshold value any infrequent item sets will lost. And if it is set low, many infrequent item sets will come into consideration. Due to this problem an optimized decision cannot be taken. So threshold should be set very precisely.

B. *Multiple Scans across the Transactional Database*

While finding any frequent item sets, we have to scan whole database many times. This multiple scan will lead to following problems:

1. Wastage of time, because searching entire database for any item takes lot of time.
2. Wastage of space, because lot of memory is needed.

C. *Performance on Scaling*

If the no of transactions increased, performance is not scaled with increasing transactions. Scalability is an important factor which is difficult to implement with algorithms of association rule mining.

IV. Apriori Algorithm

Apriori is very much basic algorithm of Association rule mining. It was initially proposed by R. Agrawal and R Srikant[8] for mining frequent item sets. This algorithm uses prior knowledge of frequent itemset properties that is why it is named as Apriori algorithm. Apriori makes use of an iterative approach known as breath-first search, where k-1 item set are used to search k item sets. There are two main steps in Apriori. 1) Join - The candidates are generated by joining among the frequent item sets level-wise. 2) Prune-Discard items set if support is less than minimum threshold value and discard the item set if its subset is not frequent[9].

A. *Apriori Algorithm*

```
L1 = find_frequent_1-itemsets(D);
for(k=2; Lk-1≠ ф ; k++)
{
        Ck = apriori_gen(Lk-1 , mn_sup);
        for each transaction t ∈ D
        {
                Ct = subset (Ck, t);
                for each candidate c ∈ Ct
                c.count++;
        }
        Lk = { c ∈ Ck | c.count ≥ min_sup}
    }
    Answer = UkLk;

    Procedure priori_gen (Lk-1 : frequent(k-1) – itemsets)
    for each itemset l1 ∈ Lk-1
    {
        for each itemset l2 ∈ Lk-1
        {
                if(l1[1] = l2[1] ∧ (l1[2] = l2[2] ∧ . . . ∧ (l1[k-2] = l2[k-2]) ∧ l1[k-1] < l2[k-1]) then
                {
                        C = l1 ⋈ l2 ;
                        if  infrequent_subset (c, Lk-1) then
                        delete c ;
                        else
                        add c to Ck;
                }
        }
    }
    return Ck;

    Procedure infrequent_subset( c : candidate k- itemset; Lk-1 : frequen(k-1) – itemset)
    for each (k-1) – subset s of c
    {
        if s ∉ Lk-1 then
        return true;
    }
    return false;
```

where D=database,minsup=user defined minimum support

The basic steps to mine the frequent elements are as follows:

Generate and test:
In this first find the 1-itemset frequent elements L by scanning the database and
removing all those elements from C which cannot satisfy the minimum support criteria.

Join step:
To attain the next level elements Ck join the previous frequent elements by self join i.e. Lk-1*Lk-1 known as Cartesian product of Lk-1 . I.e. This step generates new candidate k-itemsets based on joining Lk-1 with itself which is found in the previous iteration. Let Ck denote candidate k-itemset and L be the frequent k-itemset.

Prune step:
Ck is the superset of Lk so members of Ck may or may not be frequent but all K ' 1 frequent itemsets are included in Ck thus prunes the Ck to find K frequent itemsets with the help of Apriori property. I.e. This step eliminates some of the candidate k-itemsets using the Apriori property A scan of the database to determine the count of each candidate in Ck would result in the determination of Lk (i.e., all candidates having a count no less than the minimum support count are frequent by definition, and therefore belong to Lk). Ck, however, can be huge, and so this could involve grave computation. To shrink the size of Ck, the Apriori property is used as follows. Any (k-1)-itemset that is not frequent cannot be a subset of a frequent k-itemset. Hence, if any (k-1)-subset of candidate k-itemset is not in Lk-1 then the candidate cannot be frequent either and so can be removed from Ck. Step 2 and 3 is repeated until no new candidate set is generated.

It is no doubt that Apriori algorithm successfully finds the frequent elements from the database. But as the dimensionality of the database increase with the number of items then:

- More search space is needed and I/O cost will increase.
- Number of database scan is increased thus candidate generation will increase results in increase in computational cost.

B. *Advantage*

1. Ii is very easy and simple algorithm.
2. Its implementation is easy.

C. *Disadvantage*

1. It does multiple scan over the database to generate candidate set.
2. The number of database passes are equal to the max length of frequent item set.
3. For candidate generation process it takes more memory, space and time

V. REVIEW ON VARIOUS IMPROVEMENTS OF APRIORI ALGORITHM

In *An improved Apriori Algorithm for Association Rules of Mining* [1] the basic concepts of association rule mining and the classical Apriori algorithm is discussed. The idea to improve the algorithm is also discussed. The new algorithm is made that works on the following technique, firstly, separate every acquired data according to discretization of data items and count the data while scan the database, secondly, prune the acquired item sets. After analysis, the improved algorithm reduces the system resources occupied and improves the efficiency and quality.

Using distributed apriori association rule and classical apriori mining algorithms for grid based knowledge discovery [2] the paper presents the implementation of an association rules discovery data mining task using Grid technologies. A result of implementation with a comparison of classic apriori and distributed apriori is also discussed. Distributed data mining systems provide an efficient use of multiple processors and databases to speed up the execution of data mining and enable data distribution. The main aim of grid computing is to give organizations and application developers the ability to create distributed computing environments that can utilize computing resources on demand. Therefore, it can help increase efficiencies and reduce the cost of computing networks by decreasing data processing time and optimizing resources and distributing workloads, thereby allowing users to achieve much faster results on large operations and at lower costs. In this paper distributed apriori association rule on grid based environment is mined and the knowledge obtained is interpreted.

Optimization of association rule mining and apriori algorithm Using Ant colony optimization [3].This paper is on Apriori algorithm and association rule mining to improved algorithm based on the Ant colony optimization algorithm. ACO was introduced by dorigo and has evolved significantly in the last few years. Many organizations have collected massive amount data. This data set is usually stored on storage database systems. Two major problems arise in the analysis of the information system. One is reducing unnecessary objects and attributes so as to get the minimum subset of attributes ensuring a good approximation of classes and an acceptable quality of classification. Another one is representing the information system as a decision table which shows dependencies between the minimum subset of attributes and particular class numbers without redundancy. In Apriori algorithm, is working process explained in steps. Two step processes is used to find the frequent item set to join and prune. ACO algorithm was inspired from natural behavior of ant colonies. ACO is used to solve to numerous hard optimizations including the travelling salesman problem. ACO system contains two rules .One is local pheromone update rule, which is applied in constructing solution. Another one is global pheromone update rule which is applied in ant construction.ACO algorithm includes two more mechanisms, namely trail evaporation and optionally deamonactions.ACO algorithm is used for the specific problem of minimizing the number of association rules. Apriori algorithm uses transaction data set and uses a user interested support and confidence value then produces the association rule set. These association rule set is discrete and continues. Hence weak rule set are required to prune.

An Improved Apriori Algorithm Based on Pruning Optimization and Transaction Reduction [4] elaborates the basic ideas and the shortcomings of Apriori algorithm, studies the current major improvement strategies of it. The improved Apriori algorithm based on pruning optimization and transaction reduction is proposed. According to the performance comparison in the simulation experiment, the number of frequent item sets is much less and the running time is significantly reduced as well as the performance is enhanced then finally the algorithm is enhanced.

An Improved Apriori Algorithm [5] called APRIORI-IMPROVE is proposed based on the limitations of Apriori. APRIORI-IMPROVE algorithm presents optimizations on 2-items generation, transactions compression and uses hash structure to generate L2, uses an efficient.

Optimization of Association Rule Mining through Genetic Algorithm [6] explains the Strong rule generation is an important area of data mining. In this paper authors design a novel method for generation of strong rule. In which a general Apriori algorithm is used to generate the rules after that authors use the optimization techniques. Genetic algorithm is one of the best ways to optimize the rules. In this direction for the optimization of the rule set they design a new fitness function that uses the concept of supervised learning then the GA will be able to generate the stronger rule set.

The Research of Improved Association Rules Mining Apriori Algorithm [11] points out the bottleneck of classical Apriori's algorithm, presents an improved association rule mining algorithm. The new algorithm is based on reducing the times of scanning candidate sets and using hash tree to store candidate item sets. According to the running result of the algorithm, the processing time of mining is decreased and the efficiency of algorithm has improved.

An Improved Apriori-based Algorithm for Association Rules Mining [12] elaborates that because of the rapid growth in worldwide information, efficiency of association rules mining (ARM) has been concerned for several years. In this paper, based on the original Apriori algorithm, an improved algorithm IAA is proposed. IAA adopts a new count-based method to prune candidate itemsets and uses generation record to reduce total data scan amount. Experiments demonstrate that our algorithm outperforms the original Apriori and some other existing ARM methods. In this paper, an improved Apriori-based algorithm IAA is proposed. Through pruning candidate itemsets by a new count-based method and decreasing the mount of scan data by candidate generation record, this algorithm can reduce the redundant operation while generating frequent itemsets and association rules in the database. Validated by the experiments, the improvement is notable. This work is part of our Distributed Network Behavior Analysis System, though we have considered C-R problem in our algorithm, for specific dataset, more work is still needed. We also need further research to implement this algorithm in our distributed system.

VI. CONCLUSION

Association rule mining is an interesting topic of research in the field of data mining. We have presented a survey of most recent research work. However association rule mining is still in a stage of exploration and development. There are still some essential issues that need to be studied for identifying useful association rules. We hope that data mining researchers can solve these problems as soon as possible.

1. To make frequent pattern mining an essential task in data mining, much research is needed.
2. Most approaches are based on some strict assumptions. They should be generalized so that they can be more widely used.
3. More efficient and scalable methods for Association Rule mining should be developed.
4. Single scan and online mining methods should be developed.
5. Database-independent measurements should be established.
6. Deep-level association rules should be identified.
7. Techniques for mining association rules in multi-databases should be explored.
8. Effective techniques for Web Usage Mining should be developed.
9. New applications of association rule mining should be explored.

REFERENCES

[1] WEI Yong-qing, YANG Ren-hua, LIU Pei-yu, ”An Improved Apriori Algorithm for Association Rules of Mining” IEEE(2009).

[2] Mrs. R. Sumithra, Dr (Mrs). Sujni Paul, “Using distributed apriori association rule and classical apriori mining algorithms for grid based knowledge discovery”, 2010 Second International conference on Computing, Communication and Networking Technologies, IEEE.

[3] Badri patel ,Vijay K Chaudahri,Rajneesh K Karan,YK Rana, “Optimization of association rule mining apriori algorithm using Ant Colony optimization” International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-1, Issue-1, March 2011.

[4] Zhuang Chen, Shibang CAI, Qiulin Song and Chonglai Zhu, “An Improved Apriori Algorithm Based on Pruning Optimization and Transaction Reduction”, IEEE 2011.

[5] Rui Chang, Zhiyi Liu, “An Improved Apriori Algorithm”, 2011 International Conference on Electronics and Optoelectronics (ICEOE 2011).

[6] Rupali Haldulakar, Prof. Jitendra Agrawal,“ Optimization of Association Rule Mining through Genetic Algorithm”, International Journal on Computer Science and Engineering (IJCSE), Vol. 3, Issue. 3, Mar 2011

[7] Charanjeet Kaur, International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 2, Issue 6, June 2013 “Association Rule Mining using Apriori Algorithm: A Survey”

[8] Agrawal, R. and Srikant, R. 1995.” Mining sequential patterns”, P. S. Yu and A. S. P. Chen, Eds.In:IEEE Computer Society Pres Taipei, Taiwan, 3{14}.

[9] “Andrew Kusiak, Association Rules-The Apriori algorithm [Online],Available: http://www.engineering. uiowa. edu/~comp /Public /Apriori.pdf.

[10] M.Sathish “A Literature Survey on association rule using apriori algorithm in various fields”

[11] Huiying Wang, Xiangwei Liu, “The Research of Improved Association Rules Mining Apriori Algorithm” 2011 Eighth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD).

[12] Huan Wu, Zhigang Lu, Lin Pan, Rongsheng Xu, Wenbao Jiang,“ An Improved Apriori-based Algorithm for Association Rules Mining”, Sixth International Conference on Fuzzy Systems and Knowledge Discovery, IEEE Society community, 2009.

[13] Karl Aberer, (2007-2008),Data mining-A short introduction[Online],Available:http://lsirwww.epfl.ch/courses/dis/2003ws/lecturenotes/week13-Datamining print.pdf

[14] R.Divya , S.Vinod kumar ,”Survey on AIS,Apriori and FP-Tree algorithms”,In: International Journal of Computer Science and Management Research Vol 1 Issue 2 September 2012, ISSN 2278-733X

[15] Kiran R. U., and Reddy P. K.: An Improved Multiple Minimum Support Based Approach to Mine Rare Association Rules. http://www.iiit.net/techreports/2009_24.pdf

[16] Agrawal R., and Srikant R.: Fast Algorithms for Mining Association Rules. Proc. Very Large Database International Conference, Santiago, pp. 487–498, 1994.