IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 9, Issue 4 (Mar. - Apr. 2013), PP 57-61
www.iosrjournals.org

Survey on several improved Apriori algorithms

Ms. Rina Raval[1], Prof. Indr Jeet Rajput[2] , Prof. Vinitkumar Gupta[3]
[1](Department of Computer Engineering ,H.G.C.E., Vahelal, Ahmedabad, Gujarat, India)
[2](Department of Computer Engineering, H.G.C.E., Vahelal, Ahmedabad, Gujarat, India)
[3](Department of Computer Engineering, H.G.C.E., Vahelal, Ahmedabad, Gujarat, India)

Abstract : *Apriori algorithm has been vital algorithm in association rule mining. Main idea of this algorithm is to find useful patterns between different set of data. It is a simple algorithm yet having many drawbacks. Many researches have been done for the improvement of this algorithm. This paper does a survey on few good improved approaches of Apriori algorithm. This will be really very helpful for the upcoming researchers to find some new ideas from these approaches.*

Keywords *- Apriori, Association, PW-factor, Q-factor, Record filter, Trade list*

I. Introduction

As Information Technology is growing, databases created by the organizations are becoming huge. These organization sectors include banking, marketing, telecommunications, manufacturing, transportation etc. To extract valuable data, it necessary to explore the databases completely and efficiently. Data mining which helps to identify valuable information in such huge databases. Data Mining is an analytic process designed to explore data in search of consistent patterns and/or systematic relationships between variables, and then to validate the findings by applying the detected patterns to new subsets of data. Data mining has been very interesting topic for the researchers as it leads to automatic discovery of useful patterns from the database. This is also known as 'Knowledge Discovery'. Many techniques have been developed in data mining amongst which primarily Association rule mining is very important which results in association rules. These rules are applied on market based analysis, medical applications, science and engineering, music data mining ,banking etc for decision making.

1.1 Motivation

Databases are increasing in the fields like business, medical, sports, education, transportation, IT etc .With the increase of database for a particular user who is looking for a pattern not for complete data in the database, it is better to read wanted data than unwanted data .This wanted data and other advantage by data mining technique is that only required pattern will be drawn from database with in short time. As there are huge databases in every field, the need of finding efficient pattern from database is increased .

I. Association Rules

Association rules are used to unearth relationships between apparently unrelated data in a relational database[1].It is having two important things support and confidence. Support is the number of transactions in which the association rule holds[3]. 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)

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.

II. Apriori Algorithm

Apriori is very much basic algorithm of Association rule mining. It was initially proposed by R. Agrawal and R Srikant[2] for mining frequent item sets. This algorithm uses prior knowledge of frequent item set 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].

3.1 Apriori Algorithm

```
L1=find_frequent_1-itemsets(D);
 for(k=2; Lk-1≠Φ; k++)
{
 Ck=apriori_gen(Lk-1, min_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 apriori_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:frequent(k-1)-itemsets)
 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

3.2. Advantage

1) Ii is very easy and simple algorithm.
2) Its implementation is easy[11].

3.3 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

III. Review On Various Improvements Of Apriori Algorithm

Several improved algorithms have been proposed to conquer drawbacks of Apriori algorithm in several ways. Here presents six different approaches that face the common drawback.

4.1 Intersection and Record filter approach

4.1.1 Enlightenment

To present proposed algorithm, Goswami D.N., Chaturvedi Anshu and Raghuvanshi C.S.[4] has given Record filter and Intersection approach. In Record filter approach, 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. In Intersection approach, to calculate the support , count the common transaction that contains in each element's of candidate set. This approach requires very less time as compared to classical Apriori. In Proposed Algorithm, set theory concept of intersection is used with the record filter approach. In proposed algorithm, to calculate the support, count the common transaction that contains in each element's of candidate set. In this approach, constraints are applied that will consider only those transaction that contain at least k items.

4.1.2 Disadvantage

Memory optimization is done but still it needs much optimization.

4.2 Improvement based on set size frequency

4.2.1 Enlightenment

To eradicate non significant candidate keys the modified algorithm introduces issues such as set size and set size frequency .These issues can diminish candidate keys in a more efficient way. The improved algorithm for Apriori[5] takes for the set size which 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. Remove 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 set size. Set size which are not greater than or equal to min set size support are eliminated.

4.2.2 Disadvantage

Ideal starting size of combination size for pruning candidate keys is not given.

4.3 Improvement by reducing candidate set and memory utilization

4.3.1 Enlightenment

This algorithm[6] 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 itemset. If the number
of (k-1)-dimensional subsets of all (k-1)-dimensional frequent itemset Lk-1, which contains Ik, is less than k, then Ik is not a k-dimensional frequent itemset..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.[4] 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.

4.4 Algorithm based on Trade list

4.4.1 Enlightenment

This algorithm scans the database at the start only once and then makes the undirected item set graph.[7] From this graph by considering minimum support it finds the frequent item set and by considering the minimum confidence it generates the association rule. If database and minimum support is changed, the new algorithm finds the new frequent items by scanning undirected item set graph. That is why it's executing efficiency is improved conspicuously compared to classical algorithm.
It makes each item as a node(V) and at the same time it makes the supporting trade list for each node. Supporting trade list is a binary group T= {Tid, Itemset} (where Tid is transaction id and Itemset is trade item set). So the side between nodes can be accomplished by corresponding trade list operation. The algorithm does the intersection of two nodes with supporting trade list.

For search strategy select a node Vi from node set V. If the number of times Vi appears in the database is not less than the minimum support minsup, then {Vi} will belong to the frequent 1-item set. If count of node

Vi adjacent to node Vj's side is not less than support S, then { Vi, Vj } will belong to the item in frequent 2-iterm set. When there are three nodes in undirected item set graph and count of each side of the node is not less than minimum support minsup, these three nodes { Vk,Vm,Vn} will belong to frequent 3-item set. When there more than three nodes in undirected item sets graph then count of each side of the node should not be less than minimum support minsup and all the subset of these n nodes should be frequent. Subsequently nodes are added to k-item set. Main advantage of this approach is scanning of database is done once and after that the graph to find frequent item set.

4.5 Algorithm based on frequency of items

4.5.1 Enlightenment

In this paper, Mamta dhanda suggests a ground-breaking and attentive approach for the mining of interesting association patterns from transaction database.[8] 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[6] have 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.

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

$$Q - Factor = P / \Sigma Pi \quad (1)$$

Than it gives Transactional database where each item's frequency is counted in each transaction. From that pruning is done with minsup and confedence. Finally calculation of Weighting –factor is done based on frequency of itemset and Q-factor.

$$PW = \sum_{i=1}^{n} \text{frequency} * Q - Factor \quad (2)$$

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

4.5.2 Disadvantage

Initially classical algorithm is used. To improve efficiency some improvement can be done on pruning for faster execution.

4.6 Utilization of Attributes

4.6.1. Enlightenment

In this approach[10] using Tanagra Tool frequent item set is found by applying Apriori algorithm on database. Main problem of finding all association rules that satisfy minimum support and confedence thresholds given by users. Work illustrates that Association rule mining has several problems that it only tells whether item is present in database or absent , it treats all present or absent items equally ,it does not consider importance of item to user/business perspective and it fails to associate output i.e. frequent items with user and business objectives. These disadvantages can be removed by using attributes like profit, quantity, frequency of items which will give important information to user and business.

4.6.2 Disadvantage

Various attributes like frequency, weight can be associated with frequent item set which can provide more information for business and user point of view ,which is not done here.

IV. Conclusion And Future Work

After doing survey of above algorithms conclusion can be given that mostly in improved Apriori algorithms ,aim is to generate less candidate sets and yet get all frequent items. In the approach of Intersection and Record filter, intersection is used with the record filter approach where to calculate the support, count the common transaction that contains in each element's of candidate set. In this approach, only those transactions are considered that contain at least k items. In other approach set size and set size frequency are considered. Set

size which are not greater than or equal to min set size support are eliminated. Improvement by reducing candidate set and memory utilization only needs to compare the count of each element of Lk-1with the count of each element (X) of Ck. If the count of the element X equals to k, then only keep X. Also the item that not appears in Lk-1 will no longer appear in Lk so it is deleted. Trade list approach uses undirected item set graph. From this graph by considering minimum support it finds the frequent item set and by considering the minimum confidence it generates the association rule. Second last approach considers frequency and profit of items and generates association rules. Last approach suggests utilization of attributes like weight ,profit to associate with frequent item set for better information gain for user and business standpoint.

Survey concludes that many improvements are needed basically on pruning in Apriori to improve efficiency of algorithm. Considering frequency of items in database is also a good area to work on

Acknowledgement

Rina Raval wishes to thank Asst. Professor Indr Jeet Rajput for his guidance for paper and Asst Prof.Vinitkumar Gupta also for his suggestions in paper.

REFERENCES

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

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

[3] 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*

[4] Goswami D.N., Chaturvedi Anshu.,Raghuvanshi C.S.," An Algorithm for Frequent Pattern Mining Based On Apriori", In: *Goswami D.N. et. al. / (IJCSE) International Journal on Computer Science and Engineering „Vol. 02, No. 04, 2010, 942-947, ISSN : 0975-3397*

[5] Sheila A. Abaya, "Association Rule Mining based on Apriori Algorithm in Minimizing Candidate Generation",In:*International Journal of Scientific & Engineering Research Volume 3, Issue 7, July-2012*

[6] Zhang Changsheng, Li Zhongyue, Zheng Dongsong," An Improved Algorithm for Apriori",In: *IEEE,First International Workshop on Education Technology and Computer Science,2009*

[7] Ms. Sanober Shaikh, Ms. Madhuri Rao,Dr. S. S. Mantha," A New Association Rule Mining Based On Frequent Item Set",In: *CS & IT-CSCP 2011*

[8] Mamta Dhanda," An Approach To Extract Efficient Frequent Patterns From Transactional Database",In: *International Journal of Engineering Science and Technology (IJEST), Vol.3 No.7 July 2011, ISSN:0975-5462*

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

[10] Mamta Dhanda, Sonali Guglani , Gaurav Gupta, "Mining Efficient Association Rules Through Apriori Algorithm Using Attributes", In: *International Journal of Computer Science and Technology Vol 2,Issue 3,September 2011,ISSN:0976-8491*

[11] Hilderman R. J., Hamilton H. J.,"Knowledge Discovery and Interest Measures",In: *Kluwer Academic Publishers, Boston, 2002*