Novel Method of Apriori Algorithm using Top Down Approach

Shikha Maheshwari Pooja Jain, Asst.Prof
Department of Computer Science & Engineering,
Shri Vaishnav Institute of Technology & Sciences
& Sciences Indore(M.P)

ABSTRACT

Association Rule mining is one of the important and most popular data mining techniques. It extracts interesting correlations, frequent patterns and associations among sets of items in the transaction databases or other data repositories. Apriori algorithm is an influential algorithm for mining frequent itemsets for Boolean association rules. Firstly, the concept of association rules is introduced and the classic algorithms of association rule are analyzed. In Apriori algorithm, most time is consumed for scanning the database repeatedly. Therefore, the methods are presented about improving the Apriori algorithm efficiency, which reduces a lot of time of scanning database and shortens the computation time of the algorithm.

Key words*:* Data mining, Association rule, Apriori algorithm, frequent itemset

1. INTRODUCTION

Data mining is a kind of process of decision support. It gets the potential and useful information and acknowledges from practical application data which is large, incomplete, noisy ambiguous and random[1].Association rule mining finds interesting association or correlation relationships among a large set of data items[2,3]. Association rule mining has been well studied in data mining, especially for basket transaction data analysis. Association rule also used in various areas such as telecommunication networks, market, risk management and inventory control etc. Aside from being applicable for e-commerce, business intelligence and marketing applications, it helps web designers to restructure their web site. Apriori utilizes a complete bottom up search with a horizontal layout and enumerate all frequent item sets [4]. The proposed improved method of Apriori algorithm utilizes top down approach, where the rules are generated by avoiding generation of un-necessary patterns. The major advantage of this method is, the number of database scans is greatly reduced.

2. RELATED WORK

One of the most well known and popular data mining techniques is the Association rules or frequent item sets mining algorithm. The algorithm was originally proposed by Agrawal et al. [1] [2] for market basket analysis. Because of its significant applicability, many revised algorithms have been introduced since then, and Association rule mining is still a widely researched area. Many variations done on the frequent pattern mining algorithm of Apriori is discussed in this section. Association rule generation is used to relate pages that are most often referenced together in a single server sessions . In the context of web usage mining, association rules refer to sets of pages that are accessed together with a support value exceeding some specified threshold. Agrawal et al. presented an AIS algorithm in [1] which generates candidate item sets on-the-fly during each pass of the database scan. Large item sets from previous pass are checked if they are present in the current transaction. Thus new item sets are formed by extending existing item sets. This algorithm turns out to be ineffective because it generates too many candidate item sets. It requires more space and at the same time this algorithm requires too many passes over the whole database and also it generates rules with one consequent item.

Agrawal et. al. [2] developed various versions of Apriori algorithm such as Apriori, AprioriTid, and AprioriHybrid. Apriori and AprioriTid generate item sets using the large item sets found in the previous pass, without considering the transactions. AprioriTid improves Apriori by using the database at the first pass. Counting in subsequent passes is done using encodings created in the first pass, which is much smaller than the database. This leads to a dramatic performance improvement of three times faster than AIS. A further improvement, called AprioriHybrid, is achieved when Apriori is used in the initial passes and switches to AprioriTid in the later passes if the candidate k-itemset is expected to fit into the main memory. Even though different versions of Apriori are available, the problem with Apriori is that it generates too many 2-item sets that are not frequent. A Direct Hashing and Pruning (DHP) algorithm is developed in that reduce the size of candidate set by filtering any k-item set out of the hash table, if the hash entry does not have minimum support. This powerful filtering capability allows DHP to complete execution when Apriori is still at its second pass and hence shows improvement in execution time and utilization of space. Scalability is another important area of data mining because of its huge size. Hence, algorithms must be able to "scale up" to handle large amount of data. Eui-Hong et. al [4] tried to make data distribution and candidate distribution scalable by Intelligent Data Distribution (IDD) algorithm and Hybrid Distribution (HD) algorithm respectively. IDD addresses the issues of communication overhead and redundant computation by using aggregate memory to partition candidates and move data efficiently. HD improves over IDD by dynamically partitioning the candidate set to maintain good load balance. Another scalability study of data mining is reported by introducing a light-weight data structure called Segment. Support Map (SSM) that reduces the number of candidate item sets needed for counting . SSM contains the support count for the 1-item set. The individual support counts are added together as the upper bound for k-item sets. Applying this to Apriori, the effort to generate 1-item set is saved by

simply inspecting those SSM support counts that exceed the support threshold. Furthermore, those 1-item sets that do not meet the threshold will be discarded to reduce the number of higher level item sets to be counted. Evolutionary Algorithms (EA) are widely adopted in many scientific areas. EA borrows mechanisms of biological evolution and applies them in problem-solving, especially suitable for searching and optimization problems. Hence, the problem of mining with Association rules is a natural fit. Besides Association rule mining Evolutionary algorithms are also reported that can generate association rules . It allows overlapping intervals in different item sets[8].

The quality of the association rule discovered is measured in terms of confidence. The rules with confidence above a certain level (threshold value) are considered as interesting and deserve attention. Most algorithms define interestingness in terms of user-supply thresholds for support and confidence. The problem is that these algorithms rely on the users to set suitable values. Another algorithm called APACS2 is proposed in , that makes use of an objective interestingness measure called adjusted difference. It also discovers both positive and negative association rules. APACS2 uses adjusted difference as an objective interestingness measure. Adjusted difference is defined in terms of standardized difference and maximum likelihood estimate. A survey on different methods and algorithms used to find frequent patterns is presented in . Analysis of algorithms and descriptions for AprioriTid, AprioriHybrid, Continuous Association Rule Mining Algorithm (CARMA), Eclat algorithm, and Direct hashing and Pruning (DHP) algorithm is explained in detail. Conclusions are drawn as, for dense databases Éclat algorithm is better, for sparse databases the Hybrid algorithm is the best choice and as long as the database fits in main memory the Hybrid algorithm (combination of optimized version of Apriori and Eclat) is most efficient one. An improved version of original Apriori- All algorithm is developed for sequence mining in [15]. It adds the property of the userID during every step of producing the candidate set and every step of scanning the database to decide about whether an item in the candidate set should be used to produce next candidate set. The algorithm reduces the size of candidate set in order to reduce the number of database scanning. Based on the temporal association rule [3] [5], retailers make better promotion strategies. The time dimension exists in all transaction, and is included in finding large item sets, especially when not all items exist throughout the entire data gathering period. The temporal concept introduced in [9] addition to the normal support and confidence. The temporal support is the minimum interval width. Thus, a rule is considered as long as there is enough support or temporal support.

Different works are reported in the literature to modify the Apriori logic so as to improve the efficiency of generating rules. Enhanced version of Apriori algorithm is presented in where, the efficiency is improved by scanning the database in forward and backward directions. Xiang-wei Liu et.al presented an improved association rule mining algorithm that reduces scanning time of candidate sets using hash tree. Another version of Apriori is reported in as an algorithm called Apriori algorithm, which optimizes the join procedure of frequent item sets generated to reduce the size of the candidate item sets. The algorithm presented in scans the database only once to generate a frequent item sets, thereby saving time and increasing efficiency. These methods even though focused on reducing time and space, in real time still needs improvement. Another way to improve Apriori is to use most suitable data structure such as frequent pattern tree. Han et. al., in [7] introduced an algorithm known as FP-Tree algorithm for frequent pattern mining. It is another milestone in the development of association rule mining and avoids the candidate generation process with less passes over the database. FP-Tree algorithm breaks the bottlenecks of Apriori series algorithms but suffers with limitations. It is difficult to use in an environment that users may change the support threshold with regard to the mining results, and once the support threshold changed, the old FP-Tree cannot be used anymore, hence additional effort is needed to re-construct the corresponding FP-Tree. It is not suitable for incremental mining, since as time goes on databases keep changing, new datasets may be inserted into the database or old datasets be deleted, and hence these changes lead to a re-construction of the FP-Tree[6]. Even though fast algorithms are reported for Association mining it still inherits the drawback of scanning the whole data base many times. The survey reveals that more attention is required to address the issues related to reduce the number of database scan, and also to reduce memory space with less execution speed. This results in a large number of disk reads and placing a huge burden on the I/O subsystem. These limitations and other related issues motivated us to continue the research work in this area.

3. ASSOCIATION RULE MINING

Mining association rule is one of main content of data mining research at present, and emphasizes particularly on finding the relation of differ items in the database.

Transaction database DB, I={i1, i2, i3…in} is a set of items with n different itemsets in DB, each transaction T in DB is a set of item (i.e. itemsets), so T⊆ I[5].

Definition 1: Let I={i1, i2, … , in} be a set of items. D is a transactional database. Where (k=1,2,3…n) is an item. Tid is the exclusive identifier of transaction T in transactional database.

Definition 2: The implication of the form X=> Y is called an association rules. Where X⊆ I, Y⊆ I, and X∩Y= Ø.

Definition 3: Let D be a transactional database. If the percentage of transactions in *D* that contains X U Y is s%, the rule X =>Y holds in *D* with Support ***s*** **.** If the percentage of transactions in D containing X that also contain Y is c %, the rule X =>Y has Confidence c

Rules that satisfy both minimum support threshold (min-sup) and minimum confidence threshold (min-conf) are called strong rule .

Definition 4: If the support of item-sets *X* is greater than or equal to minimum support threshold, *X* is called frequent item-sets. If the support of item-sets *X* is smaller than the minimum support threshold, then *X* is called infrequent item-sets [6].

The design of association rule mining algorithm can be decomposed into two-step process [7]:

Step1: Find all frequent itemsets: By definition, each of these itemsets will occur at least as frequently as a pre-determined minimum support count.

Step2: Generate strong association rules from the frequent itemsets: By definition, these rules must satisfy minimum support and minimum confidence.

4. CLASSICAL APRIORI ALGORITHM SUMMARY

Apriori algorithm is one of the most influential algorithm to mine the frequent item sets of Boolean association rules.

Apriori employs an iterative approach known as a level-wise search, where k-itemsets are used to explore (k+1)-itemsets. First, the set of frequent 1-itemsets is found. This set is denoted L_1.L_1 is used to find L_2, the set of frequent 2-itemsets, which is used to find L_3, and so on, until no more frequent k-itemsets can be found. The finding of each L_k requires one full scan of the database. In order to find all the frequent itemsets, the algorithm adopted the recursive method. The main idea is as follows:

L_1 = {large 1-itemsets};
for (k=2; $L_{k-1}\neq\Phi$; k++) do
{
C_k=Apriori-gen (L_{k-1}); // the new candidates
for each transactions t∈ D do//scan D for counts
{
C_t=subset (C_k, t);
// get the subsets of t that are candidates
for each candidates c∈ C_t do
c.count++;
}
L_k={c∈ C_k |c.count≥minsup}
}
Return=$\cup_k L_k$;

all nonempty subsets of a frequent itemsets must also be frequent. To reduce the size of C_k, pruning is used as follows. If any (k-1)-subset of a candidate k-itemsets is not in L_{k-1}, then the candidate cannot be frequent either and so can be removed from C_k. The prune step reduces the cost of calculating all the support of candidate sets by reducing the size of candidate sets, which significantly improves the performance of finding frequent itemsets [9].

Apriori algorithm work in two step:

1. **Find all frequent itemsets:**

- Get frequent items:
  Items whose occurrence in database is greater than or equal to the min.support threshold.
- Get frequent itemsets:
  Generate candidates from frequent items.

Prune the results to find the frequent itemsets.

2. **Generate strong association rules from frequent itemsets**

Rules which satisfy the min.support and min.confidence threshold.

5. PROBLEM IDENTIFICATION AND PROPOSED METHOD

The classical Apriori algorithm is widely used for association mining technique by using bottom up approach. The classical Apriori algorithm performs well only when the frequent item sets are shorts. This algorithm is not useful for large amount of item set. Therefore, it is necessary to research on Apriori algorithm utilizing top down approach. The proposed algorithm uses top down approach, where in the rules are generated by avoiding generation of un-necessary patterns. The major advantage of this approach is that, the number of database scans is greatly reduced, since the number of combinations made is far less compared to the original Apriori algorithm and thus results in reduction of time and space.

Pseudo code:

Input: database (D), minimum support (min_sup).
Output: frequent item sets in D.
L1= frequent item set (D)
j=k; /* k is the maximum number of elements in a transaction from the database*/
for k= maxlength to 1
{
for i=k to 2
{
for each transaction Ti of order i
{
if (Ti has repeated)
{ Ti.count++; }
m=0;
while (i<j-m)
{**if** (Ti is a subset of each transaction Tj-m of order j-m)
{ Ti.count++; m++; }
}
If (Ti.count >=min_sup)
{ Rule Ti generated
/*store the transaction in Rule Table*/}
}
}
}

6. CONCLUSION

In this paper, the improved algorithm of Apriori algorithm is proposed to overcome the deficiency of the classical Apriori algorithm. The classical Apriori algorithm use the bottom up approach .This algorithm performs well only when the frequent item sets are shorts and it suffers from increased number of data base scan. The new proposed method use the top down approach which reduces the number of data base scans and it is useful for large amount of data base scan.

7. REFERENCES

[1] Langfang Lou, Qingxian Pan, Xiuqin Qiu , New Application of Association Rules in Teaching Evaluation System, International Conference on Computer and Information Application,2010, pp 13-16.

[2] Luo Fang, Qiu Qizhi ,The Study on the Application of Data Mining Based on Association Rules, International Conference on Communication Systems and Network Technologies 2012 pp.477-480.

[3] Karthiya Banu.R,Dr.Ravanan.R,Gopal.J ,Analysis and implementation of association rule mining 978-1-4244-8594-9/10,IEEE 2010 pp. 475-478.

[4] Agrawal, R., Imielinski, T., and Swami, A. N., Mining Association Rules Between Sets of Items in Large Databases. Proceedings of the ACM SIGMOD, International Conference on Management of Data,1993 pp.207-216.

[5]YanfeiZhou,Wanggen Wan*, Junwei Liu, Long Cai, Mining Association Rules Based on an Improved Apriori Algorithm, IEEE, 2010 pp.41s4-418.

[6] Agrawal.R and Srikant.R. , Fast algorithms for mining association rules. In Proc Int˝l Conf. Very Large Data Bases (VLDB), Sept. 1994, pages 487–499.

[7] Huiying Wang ,Xiangwei Liu,The research of improved association rules mining Apriori algorithm ,international conference on Fuzzy System and Knowledge Discovery IEEE, 2011 pp.961-964.

[8] Dr. Manish Shrivastava, Mr. Kapil Sharma, MR. Angad Singh Web Log Mining using Improved Version of Proposed Algorithm International Journal of Advanced Computer Research (ISSN (print): 2249-7277 ISSN (online): 2277-7970) Volume 1 Number 2 December 2011

[9] Guofeng Wang, Xiu Yu, Dongbiao Peng, Yinhu Cui, Qiming Li, Research of Data Mining Based on Apriori algorithm in Cutting Database, IEEE, 2010 pp.

[10] Ou Ping, Gao Y ongping, A New Improvement of Apriori Algorithm for Mining Association Rules, International Conference on Computer Application and System Modeling, 2010, pp.529-532.