A Survey On Improved Apriori Algorithm

Ekta Garg*, Meenakshi Bansal**

*Student of M.Tech
Department of Computer Engineering,
Yadavindra College of Engineering
Punjabi University, Talwandi Sabo, Punjab, India

**Assistant Professor
Department of Computer Engineering
Yadavindra College of Engineering
Punjabi Univerity, Talwandi Sabo, Punjab, India

Abstract— **As the advancement of information technology increases, the amount of data also increases. So, to handle the huge amount of data, Data mining plays an important role. Data mining is a process of extraction of valuable and unknown information from the large databases. There are various techniques and tasks but we will discuss about association rule mining and apriori algorithm. Association rule mining is a descriptive technique that is used to find out the interesting patterns among the data items stored in the database. Apriori algorithm mines the frequent itemsets and association rule learning over the transactional databases. It is used for pruning the transactions and optimizing the database. In this paper, a comprehensive overview of the previously existing work on apriori algorithm is presented and then a comparison is shown. Finally, some suggestions are given for future research.**

Keywords— **Data Mining, Association Rules, Apriori Algorithm, Frequent Itemsets**

I. INTRODUCTION

Data mining is to mine out the useful and hidden information from the large databases. It might help the user to take valuable decisions. As there is a rapid growth of technology, the problem of data explosion, rich data and poor information is also increasing day by day. To solve this, data mining is an effective method. Data mining techniques have emerged as a reflection on the association rule mining [1]. It is an important branch of data mining which is widely used in many applications like market basket analysis, medical transactions, telephonic transactions, e-banking and many more. Apriori algorithm is employed to find the association rules.

II. OVERVIEW OF ASSOCIATION RULES

Association rules were presented by R.Agarwal and others in 1993. Its main purpose is to find the association relationship among the large number of database items. Association rule mining is mainly used for market basket analysis [1]. Its main aim is to find out the interesting patterns among multiple domains based on a given degree of support and confidence.

If the support of itemsets is greater than or equal to minimum support threshold, X is frequent itemsets.

Definition: An association rule is a logical implication form of:

$$X \rightarrow Y$$

Where X and Y are predicates or set of items.
The support is the joint probability to find X and Y in the same group; the confidence is the conditional probability to find in a group Y having found X.

Confidence: Probability of set Y appear only if X appear
Support: Probability of set X and Y appear simultaneity

Support (X➔Y) = Support (XUY) = P (XY)
Confidence (X➔Y) = Support (XUY) / Support (X) = P (Y/X)

Association Rule is the rule, whose support degree and confidence degree respectively meet the threshold values given by users. Association rule mining process has two steps:

1) The first step is to find out the frequent itemsets where support degree is same or larger than the predefined minimum support degree.
2) Rules that satisfy both minimum support threshold (minsup) and minimum confidence threshold (minconf) are the strong association rules.

III. APRIORI ALGORITHM

Apriori Algorithm is the most popular & classical algorithm proposed by R. Agarwal in 1994 for mining frequent itemsets. Apriori algorithm uses an iterative method to search the rules layer by layer [1]. This proceeds by identifying the frequently occurring single item in the database and extending them by larger and larger itemsets appear sufficiently often in the database. This uses breadth first search approach to count candidate itemsets efficiently and prunes many set of items which are unlikely to be frequent [2].

A. *Classical Apriori Algorithm*

Apriori algorithm is proposed by [1]

(1) C1 = {candidate 1-itemsets};
(2) L1 = { c ϵ C1 |c.count≥minsup}};
(3) for (k=2; Lk−1≠Ø; k++) do begin
(4) Ck=apriori-gen (Lk-1);
(5) for all transactions t ϵ *D* do begin
(6) Ct = subset (Ck,t);
(7) for all candidates c ϵ Ct do
(8) c.count++;
(9) end
(10) Lk = {c ϵ Ck |c.count ≥ minsup}
(12) end
(13) Answer = ULk;

B. *Advantages and Disadvantages*

Advantages

1) Uses large itemset property.
2) Easy to implement.
3) Easily parallelized.

Disadvantages

1) Large number of candidate sets are generated.
2) Multiple database scans are required.
3) The system I/O cost increases due to multiple scanning of transactional database.

IV. LITERATURE SURVEY

A. *An Improved Aproiri Algorithm based on Pruning Optimization and Transaction Reduction*

Chen et. al introduced pruned optimization strategy. With this strategy, the generation of frequent itemsets is reduced and to compress the transaction of database, transaction reduction is used [4]. The value of the support selected is increasing which results in reducing number of frequent itemsets generated. The comparison of proposed BE-Apriori algorithm has higher efficiency than the pure apriori algorithm.

Advantages

- Decreases the system overhead.
- Reduces the running time.

Limitations

- Multiple database scans required.

B. *An Improved Apriori Algorithm*

Chang et.al proposed APRIORI-IMPROVE algorithm in which level L2 is directly generated from one scan over the database without generating candidate sets C1, L1 and C2. APRIORI-IMPROVE uses hash table and efficient horizontal data representation [6]. APRIORI-IMPROVE also optimized strategy of storage to save time & space. The performance of APRIORI-IMPROVE is higher as compared to apriori and fp-growth.

Advantages

- Requires only one scan.
- It saves the system time.

Limitations

- Large number of candidate sets are generated.

C. *The Research of Improved Association Rules Mining Apriori Algorithm*

Wang et. al proposed a new algorithm based on reducing the times of scanning candidate sets. It also uses hash structures for storage of candidate sets. In proposed algorithm [7], while generating set of candidate itemsets Ck, the Lk-1 itemset is scanned only once. It then checks whether any Lk-1 item X is subset of item Y in candidate set Ck or not. If yes, then count the number of Y itself. If the subset number of Y in Lk-1 is less than k, Y can be deleted from Ck. With this proposed algorithm, it is observed that run time is reduced gradually when the support increases. According to the results, the efficiency of the improved algorithm is better as compared to the classical algorithm.

Advantages

- Reduces the time of scanning candidate sets.
- Hash structure is used for storage of candidate sets.

Limitations

- Large number of candidate sets are generated.
- As the minconfidence value is increased, the time is also increased.

D. *An Improved Apriori Algorithm Based on Association Analysis*

Jia et.al proposed an improved algorithm based on a combination of data division and dynamic itemsets counting. The proposed algorithm [9] has improved the two main problems which are faced by classical apriori algorithm. First is the repeatedly scanning of transactional database and second is the generation of large number of candidate sets. To improve these problems, data division and dynamic itemsets counting is proposed. In data division, the transactional database is divided into n parts that don't intersect each other. In first scan, all the frequent sets of each division is mined which is called local frequent sets. In second scan, the whole database is scanned again, getting support degree of all candidate itemsets and then deciding the global frequent itemsets. After data division, dynamic itemsets counting is used to decide candidate itemsets before scanning database every time. So, the whole process needs only twice the entire database scan.

Advantages

- I/O load is reduced.
- Saves storage space.

Limitations
- Reduces the quality of candidate sets.

E. The Analysis on Apriori Algorithm Based on Interest Measure

Jingyao Hu proposed an improved apriori algorithm based on interest measure [10]. The proposed algorithm improves the readability of strong association rules by adding third threshold that is interest measure. The interest measure is based on both support and confidence of association rules. If interest measure is greater than minimum threshold of interest measure, then the rule is an interesting association rule, otherwise, the rule is of no value. The proposed algorithm reduces the set of candidates generated and also improves the speed.

Advantages
- Reduces the set of candidates generated.
- Strong association rules are produced.

Limitations
- Adding and calculating third threshold may increase the overhead.

TABLE OF COMPARISION I

Varaiations	Methodology	Input parameters	Problems to overcome	Results
A[4]	Pruning optimization	Same	Running time	Reduced running time but multiple scans still required.
B[6]	Hashing	Same	Scanning and generation of candidate sets	Requires only one scan but still large number of candidate sets generated.
C[7]	Hashing	Same	Time for scanning	Reduces the time for scanning but it depends on input parameter
D[9]	Combination of Data Division and Dynamic Itemsets Counting	Same	Scanning of database	The whole database needs twice to be scanned but it reduces the quality of candidate sets generated.
E[10]	Interest measure	Added interest measure threshold	Large number of candidate sets generated	Reduces the number of candidate sets generated but the added parameter may increase overhead.

V. CONCLUSIONS

In this paper, we analyzed and studied various existing improved apriori algorithm to mine frequent itemsets. Mainly common drawbacks are found in various existing apriori algorithm which is improved by using different approaches. It can be applied to many different applications like market basket analysis, telecommunication, network analysis, banking services and many others. In the future work, the problem of large number of candidate sets generated can still be improved and constraints can be applied.

REFERENCES

[1] Rakesh Agrawal and Ramakrishnan Srikant Fast algorithms for mining association rules in large databases. Proceedings of the 20th International Conference on Very Large Data Bases, VLDB, pages 487-499, Santiago, Chile, September 1994.

[2] J Han, "Data Mining Concepts and Techniques"Second Edition.Morgan Kaufmann Publisher, 2006,pp.123-134.

[3] Hu Ji-ming and Xian Xue-feng. The Research and Improvement of Apriori for association rules mining [J]. Computer Technology and Development 2006 16(4) 99~104.

[4] Zhuang Chen, Shibang Cai, Qiulin Song and Chonglai Zhu, "An Improved Aproiri Algorithm based on Pruning Optimization and Transaction Reduction," AMISEC 2011, 2nd IEEE International Conference, pp1908-1911.

[5] Agrawal.R, Srikant R. Fast Algorithms for Mining Association Rules in Large Databases Clin Proceedings of the 20th International Conference on Very Large Databases, I 994:487-499

[6] Rui Chang and Zhiyi Liu , " An Improved Apriori Algorithm," ICEOE 2011, IEEE International Conference, vol. 1, pp v1- 476 - v1- 478.

[7] Huiying Wang and Xiangwei Liu, "The Research of Improved Association Rules Mining Apriori Algorithm,"FSKD 2011, 8th IEEE International Conference, vol 2, pp961-964.

[8] Qiang Ma. Improved Algorithm based on Apriori Algorithm[J]. Development and application of computer,2010,23(2):6-7.

[9] Yubo Jia, Guanghu Xia, Hongdan Fan, Qian Zhang and Xu Li, "An Improved Apriori Algorithm Based on Association Analysis," ICNDC 2012, 3rd IEEE International Conference, pp208-211.

[10] Jingyao Hu, "The Analysis on Apriori Algorithm Based on Interest Measure," ICCECT 2012, IEEE International Conference, pp1010-1012.