Volume 5, Issue 6, June 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Enhanced Apriori Algorithm Using Top down Approach

[1]S. Neelima*, [2]N. Satyanarayana, [3]P. Krishna Murthy
[1]Department of CSE, Research Scholar, JNTUH, Hyderabad, India
[2]Department of CSE, Nagole Institute of Technology and Science, Hyderabad, India
[3]Principal, Swarna Bharathi Institute of Science and Technology, Khammam, AP, India

Abstract— Data mining is the process for generating frequent itemsets that satisfy minimum support. Mining frequent item set is very fundamental part of association rule mining. Number of algorithms are used for generating frequent itemsets. The Apriori algorithm generates the frequent item sets, there have been several methods proposed to improve its performance. Apriori algorithm consumes more time for scanning the database repeatedly. Enhanced apriori algorithm improves efficiency, which reduces a lot of time of scanning database and shortens the computation time of the algorithm.

Keywords— Apriori algorithm, Data Mining, Frequent itemset, Minimum support.

I. 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][7]. Data mining uses a variety of data analysis tools to discover patterns and relationships in data which is used to make predictions. Various data mining tasks are classified as: exploratory data analysis, descriptive modeling and predictive modeling, discovering patterns and rules and retrieval by content. It is currently used in a wide range of applications like healthcare, market basket analysis, education system, manufacturing engineering, scientific discovery, e-commerce and decision making [12]. Association mining is one of the most popular ways of data mining uses association rules that are an important class of methods of finding regularities/patterns in data. It is perhaps the most important model invented and extensively studied by databases and data mining community. Apriori utilizes a complete bottom up search with a horizontal layout and enumerate all frequent item sets [1]. 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.

II. 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.[7][8] for market basket analysis. Because of its important applicability, many revised algorithms have been introduced since then, and Association rule mining is still a widely researched area [1].

Agrawal et al. presented an AIS algorithm in [7] 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 [1].

Agrawal et. al.[8] 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 [5]. A further scalability study of data mining was reported by introducing a light-weight data structure called Segment Support Map (SSM) with the purpose of reduces the number of candidate item sets required for counting. Han et. al.,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.

III. ASSOCIATION RULE MINING

An association rule has the form R : X -> Y , where X and Y are two non-empty and non-intersecting itemsets. The support for rule R is defined as support(X U Y). A confidence factor (represented by percentage), defined as support(X U Y)=support(X) (assume support(X) > 0), is used to evaluate the strength of such association rules. The semantics of the confidence of a rule indicates how often it can be expected to apply, while its support indicates how trustworthy this rule is.

For example, if the minimum confidence is set to 100%, then the association rule {1,4}->{2,3} holds. But the {1,2}->{3,4} does not hold because its confidence is 67%. The goal of association rule mining is to discover all rules that have support and confidence greater than some user-defined minimum support and minimum confidence thresholds, respectively. The normally followed scheme for mining association rules consists of two stages :

1. The discovery of frequent itemsets, followed by
2. The generation of association rules [9].

IV. APRIORI ALGORITHM

Apriori is a classic algorithm for learning association rules in data mining. Apriori is an influential algorithm for mining frequent itemsets for Boolean association rules [5]. The Apriori Algorithm is a well-known approach which is proposed by Agrawal and Srikant (1994) [10]. It is an iterative approach and there are two steps in each iteration. The first step generates a set of candidate item sets. Then, in the second step we count the occurrence of each candidate set in database and prune all disqualified candidates (i.e. all infrequent item sets). Apriori uses two pruning technique, first on the bases of support count (should be greater than user specified support threshold) and second for an item set to be frequent , all its subset should be in last frequent item set. The iterations begin with size 2 item sets and the size is incremented after each iteration. The algorithm is based on the closure property of frequent item sets: if a set of items is frequent, then all its proper subsets are also frequent [3].The main idea is as follows

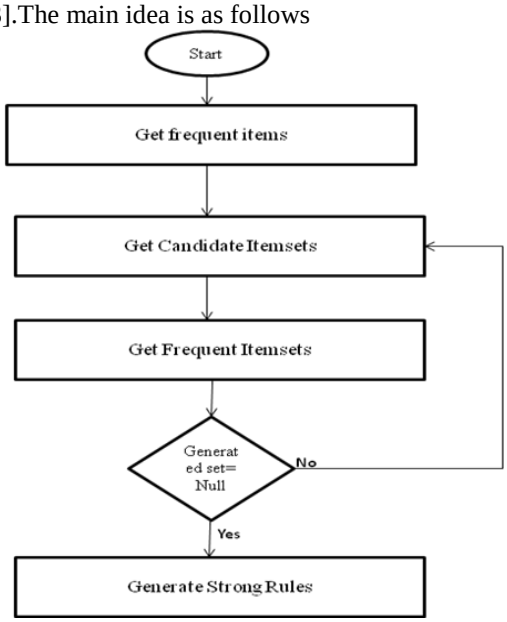

Fig:1 Flow chart for Apriori Algorithm

A. Advantages of Apriori:

1. Easy to implement
2. Uses large itemset property.
3. Easily parallelized.
4. Initial Information- transaction database D and user-defined minimum support threshold.
5. Algorithm uses information from previous steps to produce the frequent itemsets [2]

B. Disadvantages of Apriori:

1. Assumes transaction database is memory resident.
2. This algorithm is not efficient in large database.
3. This algorithm requires large number of dataset scans.
4. It only explains the presence and absence of an item in transactional databases.
5. In case of large dataset, Apriori algorithm produce large number of candidate itemsets. Algorithm scan database repeatedly for searching frequent itemsets, so more time and resources are required in large number of scans so it is inefficient in large data set [2]

V. ENHANCED APRIORI ALGORITHM

This algorithm reduces the number of scans while generating frequent itemsets. The main objective of this new approach is to build up new idea for generating frequent itemsets in transaction dataset. Top down approach is used for mining association rule. The top down Apriori algorithms uses large frequent item sets and generates frequent candidate item sets. The enhanced Apriori algorithm reduces unnecessary data base scans. This algorithm is useful for large amount of item set. Enhanced apriori algorithm uses less space and less number of iterations.

The main idea is as follows

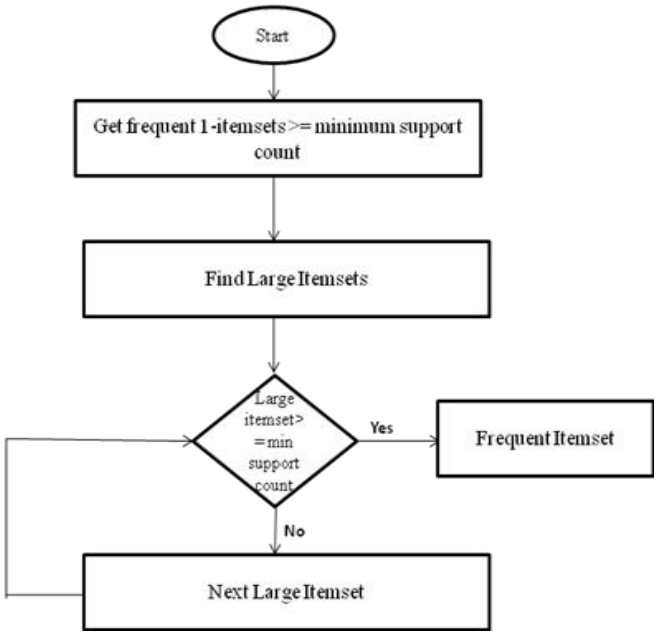

Fig:2 Flow chart for Enhanced Apriori Algorithm

A. Illustration To Enhanced Apriori Algorithm

Generate frequent item sets for the following transaction database in which there are nine sets, that is, |D|=9.
Minimum Support Threshold = $\frac{2}{9}$=22%. Minimum Support Count = $\frac{9*22}{100}$=2.

1. From the database select the large itemset.
2. Compare the large itemset >=min_support count.
3. Since {A,B,C.E} count is less than min_support count, it is not frequent itemset.

T1= {A,B,E}
T2={A,B}
T3={A,C}
T4={B,E}
T5={A,B,C}
T6={A,B,C,E}
T7={A,B,D}
T8={B,D}
T9={B,A}

} T6={A,B,C,E}=1>=2

4. Now consider next large itemset from the database.
5. Now next large itemset is large itemset.
6. Compare large itemset >=min_support count.

T1={A,B,E}=2 >=2
T5={A,B,C}=2>=2
T7={A,B,D}=1>=2

7. Since the count of {A,B,E} and {A,B,D} is greater than or equal to min_support count,{A,B,E}{A,B,D} are frequent itemset.

B. Advantages of Enhanced Apriori Algorithm:

1. Reduces unnecessary data base scans
2. Useful for large amount of item set.
3. Uses less space.
4. less number of iteration.

VI. CONCLUSION

In this paper, the enhanced Apriori algorithm is proposed to overcome the deficiency of the classical Apriori algorithm. The classical Apriori algorithm use the bottom up approach . The new proposed method use the top down approach which reduces the number of database scans and it is useful for large amount of database scan. Enhanced apriori algorithm is efficient than classical apriori algorithm and reduces the time.

REFERENCES

[1] Shikha Maheshwari Pooja Jain , Novel Method of Apriori Algorithm using Top Down Approach, International Journal of Computer Applications (0975 – 8887) Volume 77– No.10.

[2] Mr.kailash Patidar, Mr.Gajendra singh, Jatin Khalse, Improved Version of Apriori Algorithm Using Top Down Approach, IJSRD - International Journal for Scientific Research & Development| Vol. 2, Issue 10, 2014 | ISSN (online): 2321-0613.

[3] Geetha,Sk. Mohiddin , An Efficient Data Mining Technique for Generating Frequent Item sets, International Journal of Advanced Research in Computer Science and Software Engineering Volume 3, Issue 4, April 2013 ISSN: 2277 128X.

[4] Dao-I Lin, Fast Algorithms for Discovering the Maximum Frequent Set, New York University September 1998.

[5] Shikha Maheshwari Pooja Jain, The Research on Top Down Apriori Algorithm using Association Rule, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 4, April 2014 ISSN: 2277 128X.

[6] Ms Shweta, Dr. Kanwal Garg, Mining Efficient Association Rules Through Apriori Algorithm Using Attributes and Comparative Analysis of Various Association Rule Algorithms, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 6, June 2013 ISSN: 2277 128X..

[7] 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.

[8] 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.

[9] Dao-I Lin , Fast Algorithms for Discovering the Maximum Frequent Set.

[10] 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.

[11] Agrawal, R., Imielinski, T., and Swami, A. N. 1993. Mining association rules between sets of items in large databases. In Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, 207-216.

[12] Harkamal Kaur ,Mr. Abhishek Tyagi Enhancement in Apriori Algorithm using Transpose Technique to Improve Performance

