ISSN No. 0976-5697

Volume 3, No. 4, July- August 2012

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Matrix based Efficient Apriori Algorithm

[*1]Miss. Nutan Dhange, [2] Prof. Sheetal Dhande
[*1]Dept of ComputerEngineering, [2]Dept of Computer Sience and Engineering
Sipna's College of Engineering &Technology,
SGB Amravati University, Amravti,
Maharashtra, India
[*1]nutan10_dhange@yahoo.com, [2]dhande_123@rediffmail.com

Abstract: Apriori algorithm is a classical algorithm of association rule mining and widely used for mining association rule which uses frequent item. This algorithm produces overfull candidates of frequent itemsets, so the algorithm needs scan database frequently when finding frequent itemsets, so it must be inefficient. To address this problem the algorithm has improved based on the matrix. The matrix effectively indicate the affairs in the database and deal with the matrix to produce the largest frequent itemsets and others. It needn't scan the database time, again to lookup the affairs, also greatly reduce the number of candidates of frequent itemsets and improves the efficiency of computing.

I. INTRODUCTION

Association rule mining, one of the most important and well researched techniques of data mining. The associations between data are complicated and most of them are hidden. Association rule mining is the mostusually method in Association Knowledge Discovery which aim is to find out the hided information. The most famous is the Apriori algorithm.as But it has two deadly bottlenecks [2]:

a. (1) It needs great I/O load when frequently scans database. To each k circle, each element in the candidates of frequent itemsets Ck needs scan database one time to decide whether it can join the Lk. It needs scan database ten times if the frequent itemsets has ten elements.
b. (2) It may produce overfull candidates of frequent itemsets.

The number of the candidate of frequent itemsets Ck which were produced by frequent itemsets Lk - 1 increases in the speed of exponential.

The algorithm based on matrix, which takes full advantage of matrix. Matrix Apriori utilizes simple structures. It illustrates the apriori algorithm disadvantages and utilization of attributes which can improve the efficiency of apriori algorithm**.** It also minimizes the number of candidate sets, and achieving a more efficient computation than Apriori algorithm.

II. ASSOCIATION RULES

Let I=I1, I2, … , Im be a set of m distinct attributes, T be transaction that contains a set of items such that $T \subseteq I$, D be a database with different transaction records Ts. An association rule is an implication in the form of $X \Rightarrow Y$, where X, $Y \subset I$ are sets of items called itemsets, and $X \cap Y = \varnothing$. X is called antecedent while Y is called consequent, the rule means X implies Y. There are two important basic measures for association rules, support(s) and confidence(c). Since the database is large and users concern about only those frequently purchased items, usually thresholds of support and confidence are predefined by users to drop those rules that are not so interesting or useful. The two thresholds are called minimal support and minimal confidence respectively. Support(s) of an association rule is defined as the percentage/fraction of records that contain $X \cup Y$ to the total number of records in the database. Suppose the support of an item is 0.1%, it means only 0.1 percent of the transaction contain purchasing of this item. Confidence of an association rule is defined as the percentage/fraction of the number of transactions that contain $X \cup Y$ to the total number of records that contain X.Confidence is a measure of strength of the association rules, suppose the confidence of the association rule $X \Rightarrow Y$ is 80%, it means that 80% of the transactions that contain X also contain Y together.

In general, a set of items (such as the antecedent or the consequent of a rule) is called an itemset. The number of items in an itemset is called the length of an itemset. Itemsets of some length k are referred to as k-itemsets.

Generally, an association rules mining algorithm contains the following steps:

a. The set of candidate k-itemsets is generated by 1-extensions of the large (k -1)- itemsets generated in the previous iteration.
b. Supports for the candidate k-itemsets are generated by a pass over the database.
c. Itemsets that do not have the minimum support are discarded and the remaining itemsets are called large k-itemsets

III. ASSOCIATION ALGORITHMS

Association Rule provides a kind of simple form for the rule mode in data mining. Association Rule is one of the most popular methods of data mining. It simply presents the rate of some particular affairs in database taking place together, particularly is applicable to a sparse data sets.

In 1994 Agrawal etc. put forward famous Apriori algorithm[1] according to the property of association rule: the sub sets of the frequent item set is also frequent item set, the supersets of non-frequent item set is also non-frequent item set. The algorithm each time makes use of k-frequent item set carrying on conjunction to get k+1 candidate item set. Then get k+1 frequent item set through cutting. So keep on, until there is not frequent item set.

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 isused 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.

Column-wise apriori algorithm, column-wise approach to data access is often more efficient then the slandered row-wise approach. We also provide the result of empirical simulations to valid our analysis .The key idea in our approach is counting in the aproori algorithm with data access in column-wise manner,significantaly reduce the number of disk access required to identify itemset with a minimum support in the database-primarily by reducing the degree to which data and counter need to be repeatedly brought into memory[3]

Another way to improve Apriori is to use most suitable datastructure such as frequent pattern tree. Han et. al., in [4] introduced an algorithm known as FP-Tree algorithm forfrequent 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 egard 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 FPTree.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 thesechanges lead to a re-construction of the FP-Tree [5].

Dynamic itemset count (DIC) algorithm [6] for finding large itemset which uses fewer passes over the data then classic algorithm,uses fewer candidate itemsets than methods based on sampling. Dynamic Itemset Counting (DIC) [7] is an algorithm which reduces the number of passes made over the data while keeping the number of itemsets which are counted in any pass relatively low.

Enhanced version of Apriori algorithm is presented in [8] where, the efficiency is improved by scanning the database in forward and backward directions. Xiang-wei Liu et.al [9] presented an improved association rule mining algorithm that reduces scanning time of candidate sets using hash tree.

Another version of Apriori is reported in [10] as an algorithm called IApriori algorithm, which optimizes the join procedure of frequent item sets generated to reduce the size of the candidate item sets.

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 [11]. 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.

Association rules mining is one of the most well studied data mining tasks. It discovers relationships among attributes in different types of databases, producing if-then statements concerning attribute-values [2]. It was firstly introduced in [1] to discover association rules between items over basket data, an association rule describes the associations among items in which when some items are purchased in a transaction, the others are purchased, too. In order to find association rules, we need to discover all large or frequent itemsets from a large database of customer transactions.

A large itemset is a set of items which appear often enough within the same transactions.

IV. ANALYSIS OF PROBLEM

To an affair database, association rule mining is a process of finding out the strong associate rule base on the minsupport and minconfidence which were appoint by user, and it can divide into two small questions:

a. Find out the frequent itemsets: Use the minsupport to find all the frequent itemsets which not less than the minsupport. In fact, they will contain each other.
b. Produce the association rules: Find out the association rules which confidence not less than the minconfidence is in the largest frequent itemsets. So, how to find out the frequent itemsets as soon as quickly, is the focus of association rule mining.

There are two important natures in the association rule mining [12][13]:

a) 1: If itemsets X is a frequent itemsets, then all of its not-empty subsets
b) 2: If itemsets X is not a frequent itemsets, then all of its supersets are not frequent itemsets.

V. APPLICATION

a. The proposed method has the advantage of scalability to very large domains (up to a million of items are being considered), and reduces, the complexity of the original problem to the manageable sub-problem of mining association rules in much smaller sub-domains.
b. The scalability with respect to the number of records is also maintained, since no assumption is made for storing the data set in main memory and, moreover, the size of the corresponding data structures is not excessive (as in existing approaches). The problem of thrashing due to memory shortage is avoided.

c. A large domain can be replaced by a much smaller number of categories. Nevertheless, this approach requires the existence of a predetermined hierarchy and the knowledge of the items that belong in each category. In contrast, we focus on a kind of grouping that is determined by the in-between correlations of items, which is not predefined and not hierarchical

d. Complex objects such as transaction sequence, event logs, proteins and images are widely used in many fields. Efficient search of these objects becomes a critical problem for many applications. Due to the large volume of data, it is inefficient to perform a sequential scan on the whole database and examine objects one by one. High performance indexing mechanisms thus are in heavy demand in filtering objects that obviously violate the query requirement.

VI. FUTURE WORK

a. Apriori algorithm uses cut-technology when it generates item sets of candidates, it has to scan the entire database while scanning the transaction database each time. The scanning speed is very slow for its large amount of data. The improved Apriori algorithm based on matrix is improved from the Apriori algorithm and the matrix algorithm. Its basic idea is transforming the event database into matrix database so as to get the matrix item set of maximum item set. When finding the frequent k-item set from the frequent k-item set, only its matrix set is found. So only the corresponding data are calculated to get frequent k item set. Therefore the improved Apriori algorithm's computing time is very fast.

VII. REFERENCES

[1]. Agrawal R, Imielinski T, Swami A. Mining association rules between sets of items in large databases[C]. In: Proc. of the l993ACM on Management of Data, Washington, D.C, May 1993. 207-216.

[2]. Agrawal. R, and Srikant. R., Fast Algorithms for Mining Association Rules, Proceedings of 20th International Conference of Very Large Data Bases. Pp.487-499, 1994.

[3]. Data Organization and Access for Efficient Data Mining, Brian Dunkel, Nandit Soparkar, Data Engineering 1999, 15th international conference, Page(s): 522 - 529

[4]. Pattern without Candidate Generation: AFrequent-Pattern Tree Approach. Journal of Data Miningand Knowledge Discovery, 8, pp.53-87, 2004.

[5]. Han, J., Jian, Pei, and Yiwen, Yin. Mining Frequent Patterns without Candidate Generation. Proceedings of ACM International conference on Management of Data, 29(2), pp.1-12, 2000

[6]. Show-Jane Yen and Arbee L.P. Chen, A Graph-Based Approach for Discovering Various Types of Association Rules, IEEE TRANSACTIONS ONKNOWLEDGE AND DATA ENGINEERING, VOL. 13, NO. 5, SEPTEMBER/OCTOBER 2001

[7]. S.Brin and R.Motwani and J.Ullman and Shalom Tsur, Dynamic Itemset Counting and Implication Rules for Market Basket Data,SIGMOD Record, volume 6,number 2,pages 255-264,June 1997.

[8]. Wei Zhang, Zhang Wei, Dongme Sun Shaohua Teng and Haibin Zhu. An Algorithm to Improve Effectiveness of Apriori. Proceedings of 6th IEEE International Conference on Cognitive Informatics, pp.385-390, 2007.

[9]. Xiang-Wei Liu, and Pi-Lian He. The Research of mproved Association Rules Mining Apriori Algorithm.Proceedings of 3rd International Conference on Machine Learning and Cybernetics, pp.1577-1579, 2004.

[10]. Yiwu Xie, Yutong Li, Chunli Wang, and Mingyu Lu. The Optimization and Improvement of the Apriori Algorithm. Proceedings of IEEE International Symposium on Intelligent Information Technology Application Workshops, pp. 1101-1103, 2008.

[11]. Lakshmanan, V., S., Carson Kai-Sang, L., and T. Raymond. The Segment Support Map: Scalable Mining of Frequent Itemsets. Journal of ACM SIGKDD Explorations Newsletter, 2(2), pp.21-27, 2000.

[12]. Chen Wenwei. Data warehouse and data mining tutorial [M]. Beijing: Tsinghua University Press. 2006

[13]. Zhu Yixia, Yao Liwen, Huang Shuiyuan, Huang Longjun. A association rules mining algorithm based on matrix and trees[J]. Computer science. 2006, 33(7):196-198