An Association Ruled based Method for Mining Useful Information from Web Log

Vishal Gupta[1] Mahesh Malviya[2]
[1] P. G. Student [2]Assistant Professor
[1, 2]Department of Computer Science & Engineering
[1,2]JIT, Borawan(M.P.),India

Abstract: In data mining association rule is one of the most important and popular researched method for discovering interest between various large databases. The extract interest is correlation, frequent patterns and association rule among the different set of items in the transition database. In previous algorithm requires large amount of disk space for placing input/output subsystem. In order to reduce the multiple past over the disk space a new method is proposed in this paper which is top down approach instead of bottom down approach .The improved version of this algorithm will reduce the data base scan and avoid the generation of wanted patterns which will definitely reduce the data base scan, time and its space consumption.

Keywords:Data Mining, Apriori, Frequent Itemset Mining, Web Log Mining, Association rules mining.

I. INTRODUCTION

Today an association rule is the heart of data mining. It detects hidden linkages of unrelated data. These linkages are rules. That exceed a certain threshold are deemed to be interesting. The interesting rules allow actions to be taken based upon data pattern. These rules can also help making and justifying decisions.

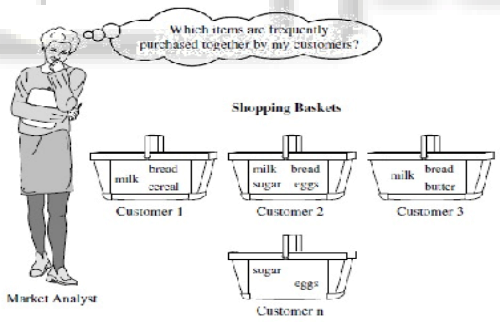

Figure Market basket analysis.

Association rules can be defined as suppose the statements is in the form $\{X1, X2, \ldots, Xn\} -> Y$ which means that Y may present in the transaction if all $X1, X2, \ldots, Xn$ are in the transaction. Note that there can be not just the single item but also it can be set of items. The probability of finding Y in a transaction with all $X1, X2, \ldots, Xn$ is called confidence. The percentage (threshold) that will hold in all transactions is called support. The level of confidence that a rule must exceed is called interestingness. There are different types of association rules. The simplest form is of association is the form that shows valid or invalid association. Market Basket Analysis is the example of the simple association rule (Boolean association rule). In this process analyzes customer buying habits by finding associations between the different items that customers place in their "shopping baskets" as shown in the figure

Rules that are generated from mining data at multiple levels of abstraction are called multiple- level or multilevel association rules. The multilevel association rules can be mined efficiently using concept hierarchies under a support- confidence framework. Another type of rule is Quantitative association rule which is a complicated type of rule. This type of rules deals with quantitive.

The Apriori Algorithm [2] is an influential algorithm for mining frequent itemsets for Boolean association rules. Frequent item sets [6] are the item sets which has minimum support . According to apriori property, any subset of frequent Itemset must be frequent. The join operation is to find Lk, which is a set of candidate k-itemsets is generated by joining Lk-1with itself. The objective is to find the frequent itemsets: the sets of items that have minimum support. A subset of a frequent Itemset must also be a frequent Itemset and iteratively find frequent itemsets with cardinality from 1 to k (k-Itemset) and use the frequent itemsets to generate association rules.

Web mining [10] is a term of applying data mining techniques for automatic discovery and extract useful information from the World Wide Web documents and its services. In brief, we can define Web mining is a technique for discovering and analyzing the useful information from the Web data. The web involves three types of data, data on the Web (web content), Web log data (usage) and Web structure data. The structure of the paper is as follows: Section 2 covers relative work. Section 3 existing and proposed Work. Conclusion is in Section 4.

II. RELATED WORK

Apriori algorithm for frequent pattern mining was proposed by Aggarwal *et al.* [7]. Though many variations of this algorithm exist till date, but Apriori is still an area where futher research is required. Many variations which are already done on the Apriori algorithm are presented in this section: One of the most important and well known popular data mining techniques is the Association rules or frequent item sets mining algorithm. This algorithm was originally proposed by Agrawal et al. [1] [2] for market basket analysis. Association rule mining is very useful thats why 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 [9]. In the context of web usage mining, the 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. The larger 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. [3] developed various versions of Apriori algorithm such as Apriori, and AprioriHybrid. Apriori 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. Support calculation in subsequent passes is one 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 [8] that reduces the size of candidate set by filtering any k-item set having support less then the minimum threshold. 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 a scalable Intelligent Data Distribution (IDD) algorithm and Hybrid Distribution (HD) algorithm respectively. The IDD algorithm 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.

III. Existing and Proposed Work

In the classical Apriori algorithm will follows bottom up approach? But in the proposed algorithm we will uses top down approach, where in the rules are generated by avoiding generation of unwanted patterns. The major advantage of this approach is that, this algorithm will reduce the data base scan and avoid the generation of wanted patterns which will definitely reduce the data base scan, time and its space.

A. Stepwise Descriptions of Pseudo code

1) Firstly scan the complete database to find the count of occurrences of each item. In the first iteration of the algorithm each item is a member of the set of candidate-1 item set C1. The compare candidate support count with minimum support count.In this way, the set of frequent-1 item set L1 is determined. It forms DB1.
2) Scan the Database DB1 for maximum element (k) transactions. The resulting database is termed DBn. The probability of forming the rule with maximum element transaction is very less because it is unique.
3) Initialize i=j=k. Now scan the database to find out i element transaction say Ti. If transaction has repeated then increment the respective counter by 1.
4) Check whether the Transaction Ti is a subset of each Transaction of bigger element Transaction set Ti+1 (Not applicable for k element transaction because it itself is a bigger set). If transaction Ti is a subset then increment the respective counter by 1. If the counts of the respective transaction are greater than the Minimum Support Count then Rule Ti is generated.
5) Repeat step 3 for all transactions of the order i (i element transactions).Decrement k by 1.Repeat step 2 until the value of k reduces to 1.

IV. Conclusions

In this paper, the improved version of Apriori algorithm is proposed. Which will overcome the deficiency of the basic Apriori algorithm? Since basic Apriori algorithm follows bottom up approach which increased number of data base scan. The newly proposed algorithm follows top down approach which reduces the number of database scans. It will take less time, less memory.

V. References

[1] 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, pp.207216, 1993.
[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] Agrawal. R., and Srikant. R. Mining Sequential Patterns. Proceedings of 11th International Conference on Data Engineering, IEEE Computer Society Press, pp.3-14, 1995.
[4] Eui-Hong Han, George Karypis, and Kumar, V. Scalable Parallel Data Mining for Association Rules. IEEE Transaction on Knowledge and Data Engineering, 12(3), pp.728-737, 2000.
[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] Han, J., Jian, Pei., Yiwen, Yin, and Runying, Mao. Mining Frequent Pattern without Candidate Generation: A Frequent-Pattern Tree Approach. Journal of Data Mining and Knowledge Discovery, 8, pp.53-87, 2004.
[7] Jong Park, S., Ming-Syan, Chen, and Yu, P. S. Using a Hash-Based Method with transaction Trimming for Mining Association Rules. IEEE Transactions on Knowledge and Data Engineering, 9(5), pp.813-825,1997.
[8] Srivastava, J., Cooley, R., Deshpande, M., and Tan, P. N. Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data. Journal of ACM Special Interest Group on Knowledge Discovery and Data Mining Explorations, 1(2), pp.12-23, 2000.

[9] Wang Tong, and He Pi-Lian. Web Log Mining by Improved Apriori All Algorithm. Transaction on Engineering Computing and Technology, 4, pp.97-100, 2005.

