Suhani Nagpal / International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 www.ijera.com
Vol. 2, Issue 3, May-Jun 2012, pp.2569-2572

Improved Apriori Algorithm using logarithmic decoding and pruning

Suhani Nagpal
Department of Computer Science and Information Technology,
Lovely Professional University, Punjab (INDIA)

ABSTRACT

In computer science and data mining, Apriori is a classic algorithm for learning association rules. Apriori is designed to operate on databases containing transactions. The purpose of the Apriori Algorithm is to find associations between different sets of data. It is sometimes referred to as "Market Basket Analysis". The aim of this research is to improve the performance of the conventional Apriori algorithm that mines the association rules. The approach is to attain the desired improvement is to create a more efficient new algorithm out of the conventional one by adding the encoding and decoding mechanisms to the latter in order to demonstrate the importance of the efficient decoding to high data mining performance and from various experiments it is proved that the logarithmic decoding method is the most efficient among the all methods it can speed up all the required processes.

Keywords – **Apriori algorithm, Data warehouse, Market-Basket Analysis**

I. INTRODUCTION

Apriori is a classic algorithm for learning association rules. Apriori is designed to operate on databases containing transactions. Other algorithms are designed for finding association rules in data having no transactions, or having no timestamps. Other algorithms are designed for finding association rules in data having no transactions, or having no timestamps. The purpose of the Apriori Algorithm is to find associations between different sets of data. It is sometimes referred to as "Market Basket Analysis". Each set of data has a number of items and is called a transaction. The output of Apriori is sets of rules that tell us how often items are contained in sets of data.

Frequent item set mining and association rule induction are powerful methods for so-called market basket analysis, which aims at finding regularities in the shopping behavior of customers of supermarkets, mail-order companies, online shops etc. With the induction of frequent item sets and association rules one tries to find sets of products that are frequently bought together, so that from the presence of certain products in a shopping cart one can infer (with a high Probability) that certain other products are present. Such information, especially if expressed in the form of rules, can often be used to increase the number of items sold, for instance, by appropriately arranging the products on the shelves of a supermarket or on the pages of a mail-order catalog (they may, for example, be placed adjacent to each other in order to invite even more customers to buy them together) or by directly suggesting items to a customer, which may be of interest for him/her.

II. RELATED WORK

In year 2011, Rui Chang, Zhiyi Liu have proposed the APRIORI-IMPROVE algorithm which reads transaction database by scanning only one time and does not generate candidate sets which reduces the task by reducing the response time as in this we need not generate the C2 which is known as candidate 2-itemset. It uses hash structure to generate L2 and uses an efficient horizontal data representation and optimized strategy of storage to save time and space. [6] In year 2010, Guoling Liu, Runian Geng discussed the positive association rules mining over multi-databases for this they have purposed an algorithm which avoids the I/O congestion and focuses on the number of item set as well as support degree, thus greatly heightening the accurate of the global association mining and it scans the whole database only once and avoids re-scanning the old database to acquire new knowledge from incremental addition to dataset.[1]

In year 2010, Yongge Shi, Yiqun Zhou have purposed some association rules to increase the efficiency of the Apriori algorithm which can improve the speed of data mining effectively, enhance the ability of ADSL line quality's analysis and solving. This improvement Apriori algorithm can be used for the telecom operators in broadband line and in substandard line's analysis. [2]

In year 2010, Xue Xing, Yao Chen, Yan-en Wang

focuses on the association rules which applied in data mining that aims to analyze large data and reveal knowledge hidden in the database. It applies on the association rule mining to the software of the examination paper evaluation system. [3]

In year 2010, Libing Wu, KuiGong, Fuliang Guo, XiaohuaGe have given a c# code which achieves the improved algorithm confirmed by many experiments, this algorithm is better than traditional algorithms in time consuming and reduce the frequency we scan the database and reduce the unnecessary duplication of effort. [4] In year 2009, Liu Jing, Qiu Chu, Lu Yongquan, Ji Haipeng, Wang Jintao, Li Nan, Gao Pengdong, Yu Wenhua have worked on the Apriorialgorithm for improving its efficiency because with large database efficiency is the most important and the promising factor they have done by reduced the number of scanning data base, reduced the number of candidate item-set which might become frequent item. [5]

III. PROPOSED METHOD

It is hypothesize that the existing techniques have either only improved the efficiency but do not give any effect to the change in the complexity collectively. So I have decided to propose a technique by which we can scan the database only once and will work on the tables which are created by the user only at the time of analysis so this will reduce the execution speed because by this we need not scan the whole database again and again as we can work on the subsets made i.e. by making the temporary table in between such that to do the work in half way around such as to reduce the execution time of the required algorithm. Secondly, by using the logarithmic technique of decoding the algorithm such that to reduce the complexity of the algorithm and by truncating the value at the integer part only and leaving the float part of the algorithm.

The main objective focuses on designing a single algorithm in which we can have both efficiency and can have less storage space. So that we can work in an efficient environment which is suitable to each person for working and it may be the most effective way to find the frequent itemsets. This scheme gives the good response and performance means it takes less time to execute the instruction and also increases throughput by finding the desired frequent item set. As we know when the execution time is reduced the cost is also less.

Algorithm design for the proposed Apriori algorithm

1. Generate the candidate itemsets in C_k from the frequent itemsets in L_{k-1}.
    1. Join L_{k-1} p with L_{k-1}q, as follows:
    insert into C_k
    select $p.\text{item}_1, p.\text{item}_2, \ldots, p.\text{item}_{k-1}, q.\text{item}_{k-1}$
    from L_{k-1} p, L_{k-1}q
    where $p.\text{item}_1 = q.\text{item}_1, \ldots p.\text{item}_{k-2} = q.\text{item}_{k-2}, p.\text{item}_{k-1} < q.\text{item}_{k-1}$
    2. Generate all (k-1)-subsets from the candidate itemsets in C_k
    3. store the values in the log form such that the result is in float value
    4. Prune all candidate itemsets from C_k where some (k-1)-subset of the candidate item set is not in the frequent item set L_{k-1}
2. Take the values from the temporary table and move to the new iteration
3. Scan the transaction database to determine the support for each candidate item set in C_k
4. Save the frequent itemsets in L_k.

The first pass of the algorithm simply uses the function getL1 to count item occurrences to determine the frequent 1- itemsets. A subsequent pass, for example pass k, consists of two phases:

a) First, the frequent itemsets Lk-1 found in the (k-1)th pass are used to generate the candidate itemsets Ck using the function getCk described below.
b) Next, the data matrix is scanned and the support of candidates in Ck is counted. For fast counting, we use the function be_subset to efficiently determine whether the candidates in Ck are contained in a given transaction or not.

There are two steps in the function getCk.

a) First, in the join step, we join Lk-1 and Lk-1 to generate potential candidates.
b) Next, in the prune step, we use the function infqn_subset to remove all candidates that have a subset that is not frequent. The pruning is based on the Aprioriproperty that "all non-empty subsets of a frequent item set must be frequent as well".

The function getCk returns a superset of the set of all frequent k-itemsets. We also use the function display_itemsets to save all the frequent itemsets.

IV. RESULTS AND DISCUSSIONS

After doing all the research work it is found that the Apriori algorithm is used in the market basket analysis which will help us to find the minimal subsets of all the items i.e. to find the most frequent subset of all the

items of the set of transactions as in the conventional Apriori algorithm is stated as Apriori is an influential algorithm to find frequent itemsets. This is the main interface of the Apriorialgorithm designed in c# which will help us to find the minimal subsets of the transactions given so for that we need to explain the working of each and everything used in this interface. We have the concept of the support and confidence which we need to discuss in this as:-

$$Rule: X \Rightarrow Y \quad \begin{cases} Support = \dfrac{frq(X,Y)}{N} \\ Confidence = \dfrac{frq(X,Y)}{frq(X)} \end{cases}$$

From this we can see that support is something which gives the surety of if X occurs then Y will also occur and in confidence we can say that if X occurs then there are chances for Y to occur.

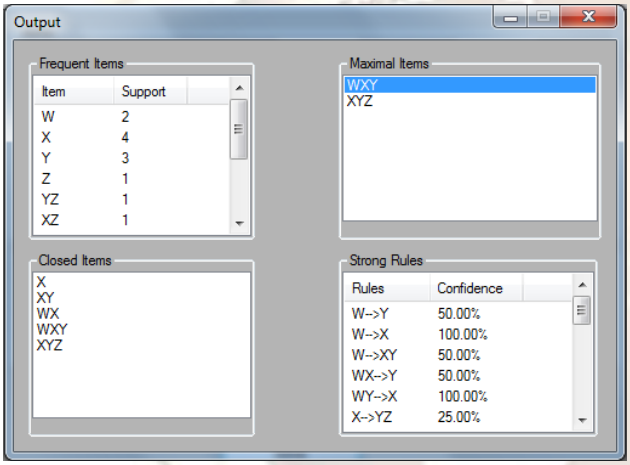

Fig. 1. Output after execution

Fig.1. shows the output window and will solve for the goal of the algorithm i.e. to find the minimal subset of all the itemsets i.e. the minimal subsets are those from which we have the maximum support at least crossing the minimum support like the most selling items of the grocery store are this. The most selling items are finding by this method and the shopkeepers have an idea of whatever the need to buy and in what quantity. So from here we can see that all the frequent items are listed in this using the concept of C1 and C2 and hence we have closed items in which we have all the items which satisfy the minimum support defined i.e. the items under the L1 and L2 table.

Then on the right hand side we have the maximal items which is the result basically which will tell us what is the result at the end i.e. the items like the {W, X, Y} and the item set {X, Y, Z} are included under this as these are the only itemsets which are able to satisfy the minimum support and able to define the max subset.

V. CONCLUSION AND FUTURE SCOPE

From the above all discussions we are able to say that the aim of this project is to improve the performance of the conventional Apriori algorithm that mines the association rules. The approach is to attain the desired improvement is to create a more efficient new algorithm out of the conventional one by adding the encoding and decoding mechanisms to the latter in order to demonstrate the importance of the efficient decoding to high data mining performance and from various experiments it is proved that the logarithmic decoding method is the most efficient among the all methods it can speed up all the required processes. Several different styles of the major improvement strategies. Then studies the pruning optimization and transaction reduction strategy, finally put forward the improved Apriori algorithm based on pruning optimization and transaction reduction. The experiments about data set retail show that ,compared with the pure Apriori algorithm, the improved Apriori algorithm has decreased the number of frequent itemsets generated, reduced the running time obviously. The improved algorithm has a good advantage of low system overhead and good operating performance, its efficiency is significantly higher than the Apriori algorithm. With the expansion of the scale of data, this advantage will become obvious increasingly.

It is concluded that we have found a method which helps in reducing the complexity of the Apriori algorithm by using the logarithmic decoding technique by which we can decode the occurrence of each and every item from a normal integer form to a specific logarithmic form. As we know that each and everything has its required advantages and disadvantages
So the Apriori algorithms also has its advantages and its various uses are given by

- Initial information: transactional database D and user-defined numeric minimum support threshold min_sup
- Algorithm uses knowledge from previous iteration phase to produce frequent itemsets
- This is reflected in the Latin origin of the name that means ”from what comes before”

The various limitations of the Apriori algorithm are given by

- Needs several iterations of the data
- Uses a uniform minimum support threshold
- Difficulties to find rarely occurring events

- Alternative methods (other than Apriori) can address this by using a non-uniform minimum support threshold
- Some competing alternative approaches focus on partition and sampling

So to overcome these limitations we can work on any of the particular limitation so as to give the future scope for this work because as we can use some technique to use the variable minimum support threshold in this particular algorithm such that we change the priority for each iteration as we were doing some for calculating the C1, L1 and hence calculating the C2, L2 and so till the minimal subsets of the required itemsets are found.

As we have seen in the particular example explained earlier which gives a list of items with the support and everything explained under this. In that we are using a minimum support = 2 which is uniform till the end and the particular set of items which are used to find the minimal subsets of all the itemsets in that particular transaction set. The itemsets in that particular transaction are calculated so we need to do the change as we need to make the modifications in the conventional Apriori algorithm we need to use variable support so as to do the market basket analysis of a particular place. The market basket analysis is a method by which we can analyze what is being sold in a particular store which either a grocery store a milk bar or apparels store. The analysis is done for each and everything as we take an example of grocery store Milk, eggs, biscuits, bread butter, chocolates, jam and many more items which are sold on the daily basis to various customers as these are the daily used items at home.

REFERENCES

1. " Guoling Liu, Runian Geng" 2010, *2nd international conference on computer engineering and technology*.
2. "Yongge Shi, Yiqun Zhou" 2010,*IEEE International Conference on Granular Computing.*
3. "Xue Xing, Yao Chen, Yan-en Wang" 2010, *International Conference on Innovative Computing and Communication and 2010 Asia-Pacific Conference on Information Technology and Ocean Engineering*
4. " Libing Wu, KuiGong, Fuliang Guo, XiaohuaGe" 2010 *computer science and information technology ICCSIT*
5. "Liu Jing, Qiu Chu, Lu Yongquan, Ji Haipeng, Wang Jintao, Li Nan, Gao Pengdong, Yu Wenhua" 2009 *international conference on computer engineering and technology*
6. "Rui Chang, Zhiyi Liu" 2011 *International Conference on Electronics and Optoelectronics (ICEOE 2011)*
7. " Dexin Zhou, Jiacang Kang, Zhicheng Fan, Wenlin Zhang" 2011 *International Conference on Electronics and communication*
8. "Zhuang Chen, Shibang Cai, Qiulin Song and Chonglai Zhu", 2011 *2nd International Conference on Artificial Intelligence, Management Science and Electronic Commerce (AIMSEC)*
9. "Allan K.Y. Wong, S.L. Wu and L. Feng", 2010 *IEEE International Conference on Systems, Man, and Cybernetics, 2010. IEEE SMC '10 Conference Proceedings*
10. "Ehsan Saboori, Shafigh Parsazad, Yasaman Sanatkhani" 2010 *3rd International Conference on Advanced Computer Theory and Engineering (ICACTE).*
11. "Yanguang Shen, Jie Liu, Jing Shen", 2010 *International Conference on Intelligent Computation Technology and Automation (ICICTA).*
12. "Introduction to Data Mining with Case Studies", G.K. Gupta.
13. "Data mining: concepts and techniques", Jiawei Han, Micheline Kamber.
14. "Introduction to data mining and its applications" S. Sumathi, S. N. Sivanandam.
15.