International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 5, Issue 2, No 02, 2017 ISSN: 2348-6600
Reference ID: IJCS-309 PAGE NO: 2065-2071

An Innovative Technique for Enhancing the Performance of Association Rules Mining for Apriori Algorithm

S.BASKARAN[#1], S.ANBUSELVAN[*2]

[#]**Head**
Department of Computer Science
Tamil university
Thanjavur 613 010
sbaskarantj@gmail.com

[*]**M.Phil Scholar**
Department of Computer Science
Tamil university
Thanjavur 613 010
psanbu81@gmail.com

Abstract - The art of data mining has been constantly evolving. There are a number of innovative and intuitive techniques that have emerged that fine-tune data mining concepts in a bid to give companies more comprehensive insight into their own data with useful future trends. Businesses can use data mining for knowledge discovery and exploration of available data. This can help them predict future trends, understand customer's preferences and purchase habits, and conduct a constructive market analysis. They can then build models based on historical data patterns and garner more from targeted market campaigns as well as strategize more profitable selling approaches. Data mining helps enterprises to make informed business decisions, enhances business intelligence, thereby improving the company's revenue and reducing cost overheads. We consider the problem of discovering association rules between items in a large database of sales transactions. We proposed a new algorithm for solving this problem that is fundamentally based on and improved upon Apriori Algorithm and its other updated versions. The proposed method is to find out all the frequent Item sets without scanning DB so many times. Count the support of the frequent itemsets by scanning the DB another time; Output the association rules from the frequent itemsets. The proposed algorithm in this paper reduces the times of scanning DB.

Index Terms—Association Rules, Apriori Algorithm, Association Analysis, Frequent Pattern.

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 5, Issue 2, No 02, 2017 ISSN: 2348-6600
Reference ID: IJCS-309 PAGE NO: 2065-2071

I. INTRODUCTION

Data mining automates the detection of relevant patterns in a database, using defined approaches and algorithms to look into current and historical data that can then be analyzed to predict future trends. Because data mining tools predict future trends and behaviors by reading through databases for hidden patterns, they allow organizations to make proactive, knowledge-driven decisions and answer questions that were previously too time-consuming to resolve.

Many techniques have been used for data mining. Association Rule Mining [ARM] is an important technique for discovering interesting associations among the entities in a dataset. ARM can be defined as the relation and dependency between the itemsets by given support and confidence in database. The major challenges associated with ARM, especially for parallel and distributed data mining, are minimizing I/O, increasing processing speed and reducing communication cost. These are the reasons for the need of improvement of ARM. In the algorithms of the association rules mining, Apriori is the ancestor which offered by Agrawal R in 1993. According to theory, the main idea of this algorithm is that the subset of a frequent itemset is a frequent itemset and the superset of an infrequent itemset is an infrequent itemset. This algorithms scan the database repeatedly to mining the association rules. Finding all such rules is valuable for cross-marketing and attached mailing applications. Other applications include catalog design, add-on sales, store layout, and customer segmentation based on buying patterns. The databases involved in these applications are very large. It is imperative, therefore, to have fast algorithms for this task.

II. BRIEF INFORMATION ON EXISTING SYSTEM

From the above discussion, it can be understood that the key problem of the Apriori is that it takes too much time for mining the frequent itemsets. This is due to costs in two areas. One is the time for scanning the large database repeatedly and the other is for generating frequent itemsets with JOIN. There are a lot of improved algorithms for Apriori such as AprioriTID, Apriori Hybri, Multiple joins, Reorder and Direct etc.

The feature of AprioriTID is that the support of the candidate frequent itemsets is calculated only at

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 5, Issue 2, No 02, 2017 ISSN: 2348-6600
Reference ID: IJCS-309 PAGE NO: 2065-2071

the first time it scanned the database D and in addition it generated candidate transaction database D' which only includes the candidate frequent itemsets. Then the latter mining is based on the database D'. Since D' is smaller than D, this technique reduces the time of I/O operation and consequently it enhances the efficiency of the algorithm.

Algorithm Apriori Hybri is the combination of algorithm Apriori and AprioriTID. In this case, when the candidate transaction database D' couldn't be contained in the RAM, it mines with algorithm Apriori otherwise with algorithm AprioriTID.

III. PROPOSED METHOD AND TECHNIQUE

The most important step in mining association is generation frequent itemsets. In algorithm apriori the most time is consumed by scanning the database repeatedly. It would reduce the running time of the algorithm by reducing the times it scans the database far and away. In this paper we proposed a technique which mine frequent itemsets by evaluating their probability of supports based on association analyzing. First, it gained the probability of every 1-itemset by scanning the database, the 1-itemset with the larger support than the probability the user sets would be frequent 1-itemsets. Second, it evaluates the probability of every 2-itemset, every 3-itemset, and every k-itemset from the frequent 1-itemsets. Third, it gains all the candidate frequent itemsets. Fourth, it scans the database for verifying the support of the candidate frequent itemsets, last the frequent itemsets are mined and association rules also do. In the method it reduces a lot of times of scanning database and shortened the calculate time of the algorithm

IV. PROBLEM DECOMPOSITION

The problem can be decomposed into the following sub problems:

1. Find all sets of items that have transaction support above minimum support. The support for an itemset is the number of transactions that contain the itemset. Itemsets with minimum support are called large itemsets, and all others are called small itemsets.

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 5, Issue 2, No 02, 2017 ISSN: 2348-6600
Reference ID: IJCS-309 PAGE NO: 2065-2071

2. This is a straightforward algorithm for generating the desired rules by using the large itemsets. For every large itemset l, find all non-empty subsets of l. For every such subset a, output a rule of the form a =) (l ─ a) if the ratio of support (l) to support (a) is at least minconf. It is needed to be considered that all subsets of l to generate rules with multiple consequents.

V. DISCOVERING LARGE ITEMSETS

Algorithms for discovering large itemsets make multiple passes over the data. In the first pass, the support of individual items is counted and then determined which of them are large, i.e. have minimum support. In each subsequent pass, it is to be started with a seed set of itemsets found to be large in the previous pass. This seed is set for generating new potentially large itemsets, called candidate itemsets, and count the actual support for these candidate itemsets during the pass over the data. At the end of the pass, it will be determined which of the candidate itemsets are actually large, and they become the seed for the next pass. This process continues until no new large itemsets are found.

The algorithms we propose differ fundamentally from the AIS and SETM algorithms in terms of which candidate itemsets are counted in a pass and in the way that those candidates are generated. In both the AIS and SETM algorithms, candidate itemsets are generated on-they during the pass as data is being read. Specifically, after reading a transaction, it is determined which of the itemsets found large in the previous pass are present in the transaction. New candidate itemsets are generated by extending these large itemsets with other items in the transaction. However, as we will see, the disadvantage

VI. CORE CONCEPTS OF PROPOSED ALGORITHM

Let P_1, $P_2 \ldots P_n$ are the independent probability of every item A_1, $A_2 \ldots A_n$, the probability for any two item A_k, A_m ($P_k < P_m$) both appeared in one transaction is $P_{km.}$

If A_k and A_m are total non-correlation, from definition, it can be concluded that $P_{km} = P_k * P_m$, if

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 5, Issue 2, No 02, 2017 ISSN: 2348-6600
Reference ID: IJCS-309 PAGE NO: 2065-2071

A_k and A_m are total correlation, then P_{km} is the minimum of the P_k and P_m that is P_k, so, $P_k * P_m \leq P_{km} \leq P_k$.

Now the problem is that: Given P_k and P_m, and $P_k * P_m \leq P_{km} \leq P_k$, then evaluate P_{km}. In this paper we discussed a method by association analysis to confirm the formula.

Let parameter a be the probability which A_k and A_m are total correlation, and parameter b for total non-correlation. $a + b = 1$, $0 < a, b < 1$, then P_{km} can be defined as formula below:

$$P_{km} = a * P_k + b * P_k * P_m$$

There are a lot methods to confirm the value of parameter a and b, in this paper we provide a way to define "a" and "b" by association analysis.

A) Confirm the Parameter "a", "b" by association analysis

There are a series of criterion about environment. The most ingredients of the pollution can be confirmed based on the source. So we can consider the criterion as a referenced list and the list needed to find the correlation as a comparison list. Then we get the correlation coefficient which is the parameter "a" in our formula, and b=1-a.

Let $S=\{S_1, S_2, \ldots S_m\}$ be the value list of item A_m, $S_1, S_2, \ldots S_m$ are sample extracted from the DB and $X=\{X_1, X_2, \ldots X_m\}$ be the value list for item A_k, $X_1, X_2, \ldots X_m$ are sample extracted from the DB.

VII. DESCRIPTION OF PROPOSED ALGORITHM

After generating the 1-dimensional frequent itemsets, every round of the original Apriori algorithm can be separated into, pruning candidates and counting candidate itemsets occurrence. Our algorithm improves prune operation by using a count-based method; the count occurrence operation is improved by decreasing the mount of scan data using generation record. Details of IAA algorithm are given below.

SD: support difference (User input),
LS: least support among all items
Set Max_value=0

a. for i = 1; i ≤ | C1 |; ++i do
    if max_value > Support(i)

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 5, Issue 2, No 02, 2017 ISSN: 2348-6600
Reference ID: IJCS-309 PAGE NO: 2065-2071

max_value=Support(i);
M(i) = Support(i) − SD(n);
if M(i) < LS then MIS(i) = LS;
else MIS(i) = M(i);
end if
end for

b. return MIS;

VIII. CONCLUSION

Data mining in association rule is an important research topic and Apriori is the core algorithm in mining association rule. We propose a method that enhances the efficiency of the existing algorithm by evaluating the probability of candidate frequent itemsets. It reduces the runtime of algorithm by reducing the times of scanning database. In this paper, we presented an efficient method for confirming the parameter “a” and b” in formula by association analysis. If the parameter “a” and “b”are unseemliness, it omits some association rules. The solution for this problem is multi enactment for parameter “a” and “b”, and then chooses the best. The method of grey relational analysis is widely used in the association analysis of environment databases. There are a lot of other methods to confirm the parameter “a” and ”b”. In this paper we implemented the algorithm and the performed experiments show significant benefits for different number of the itemsets and tuples in the database. The result shows that probability apriori with multiple minimum supports algorithm improves the performance over the existing approaches.

REFERENCES

1) R. Agrawal, T. Imielinski, and A. Swami, “Mining Association Rules Between Sets of Items in Large Databases”, Proceedings of the ACM SIGMOD Conference on Management of data, pp.207-216, May 1993.
2) R. Agrawal, and R. Srikant, "Fast Algorithms for Mining Association Rules", In Proc. VLDB 1994, pp.487-499.
3) Sujni Paul, and V. Saravanan, “Hash Partitioned Apriori in Parallel and Distributed Data Mining Environment with Dynamic Data Allocation Approach”, Computer Science and Information Technology, 2008. ICCSIT'08.

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 5, Issue 2, No 02, 2017 ISSN: 2348-6600
Reference ID: IJCS-309 PAGE NO: 2065-2071

International Conference on Aug. 29 2008-Sept. 2 2008, pp.481-485

4) Lei Ji, Baowen Zhang, and Jianhua Li, "A New Improvement on Apriori Algorithm", Computational Intelligence and Security, 2006 International Conference on Volume 1, Nov. 2006, pp.840-844
5) Kun-Ming Yu, Jia-Ling Zhou, "A Weighted Load-Balancing Parallel Apriori Algorithm for Association Rule Mining", Granular Computing, 2008. GrC 2008. IEEE International Conference on 26- 28 Aug. 2008, pp.756-761
6) R. Agrawal, T. Imielinski, and A. Swami, "Mining association rules between sets of items in large databases", SIGMOD, pg 207 - 216, 1999.
7) Bing Liu, Wynne Hsu and Yiming Ma, "Mining association rules with multiple minimum supports. In ACM Special Interest Group on Knowledge Discovery and Data Mining Explorations, pp.337–341 (1999).
8) Michael Hashler, "A Model-Based frequency constraint for mining associations from transaction data" at Springer Science & Business Media, LLC. Manufactured in the United States, Data Mining and Knowledge Discovery, 13, 137–166, 2006.
9) Ramya, V & Baskaran S. (2917), Design and Implementation of an Online Social Network Application with Privacy Violation Detector Based on text Mining Technique, International Journal of Computer Science, Vol 5, Issue 2 , pp. 2020-2026.