An Efficient Compressed Data Structure Based Method for Frequent Item Set Mining

Devendra Verma[1] **Mr. Gajendra Singh**[2]
[1,2]SSSSIT, Sehore

Abstract--Frequent pattern mining is very important for business organizations. The major applications of frequent pattern mining include disease prediction and analysis, rain forecasting, profit maximization, etc. In this paper, we are presenting a new method for mining frequent patterns. Our method is based on a new compact data structure. This data structure will help in reducing the execution time.

I. INTRODUCTION

In recent years, the size of database used for storing valuable information in an organization has increased exponentially. This has led to a increasing researchers interest in the development of tools capable in the automatic extraction of knowledge from data. Data mining or knowledge discovery in database is a field of research dealing with the automatic discovery of implicit information or knowledge within the databases. The implicit information within databases, mainly the interesting association relationships among sets of objects that lead to association rules may disclose useful patterns for rain forecast, decision support, , disease prediction, financial forecast, , attribute prediction, marketing policies, even medical diagnosis and many other applications.

With the increase in Information & Communication Technology, the size of the databases created by the organizations for information storage is also increasing. Some of such organizations include retail, telecommunications, petroleum, utilities, manufacturing, credit cards, transportation, insurance, banking etc. There are many more such organizations, involved in extracting the valuable data, it necessary to explore the databases completely and efficiently. Knowledge discovery in databases (KDD) helps to identifying precious information in such huge databases. This valuable information can help the decision maker to make accurate future decisions. KDD applications deliver measurable benefits, including profit maximization, increase in sales and reduced cost of doing business, enhanced profitability, and improved quality of service. So it is clear that the Knowledge Discovery in Databases has become one of the most active and exciting research areas in the database community.

II. RELATED WORK

The first and foremost algorithm for mining all frequent itemsets and strong association rules was the AIS algorithm by [3]. After that, the algorithm was improved and renamed Apriori. Apriori algorithm is, the most popular, classical and important algorithm for mining frequent itemsets.

It is observed that reducing the candidate items from the database is one of the important task for increasing the efficiency. If we can reduce the number of candidate sets then it will result in lower time and space complexity. To do the same a DHP technique was proposed [5] to reduce the number of candidates in the early passes. In this method, the support of an item is counted by mapping the items from the candidate list into the buckets which is divided according to support known as Hash table structure. When a new itemset is encountered if item exist earlier then increase the bucket count else insert into new bucket. At the end the bucket whose support count is less the minimum support is removed from the candidate set.

Partitioning algorithm [1] is based on the concept partitioning. To find the frequent elements on the basis partitioning of database in n parts, it uses memory efficiently. It overcomes the memory problem for large database which do not fit into main memory because small parts of database easily fit into main memory. This algorithm is divided into two main passes,

The sampling algorithm [6] is used to overcome the limitation of I/O overhead by not considering the whole database for checking the frequency. This algorithm is based in the idea to pick a random sample of itemset R from the database instead of whole database D. The small sample is picked in such a way that whole sample is accommodated in the main memory. In this algorithm we try to find the frequent elements for the sample only and there is chance to miss the global frequent elements in that sample therefore lower threshold support is used instead of actual minimum support to find the frequent elements local to sample. In the best case , the sampling algorithm requires only one pass to find all frequent elements if all the elements included in sample and if elements missed in sample then second pass are needed to find the itemsets missed in first pass or in sample [7].

The DIC algorithm [4] was also used to reduce the number of database scan. It is also based upon the downward disclosure property in which adds the candidate itemsets at different point of time during the scan.

It was observed in [8] [7] [9] [10] that the improved Apriori algorithm is based on the combination of forward scan and reverse scan of a given database. If certain conditions are satisfied then the improved algorithm can greatly reduce the iteration, scanning times required for the discovery of candidate itemsets.

III. PROBLEM DEFINITION

Let I = {I1, I2, In} be a set of all items. Then a k-item set α, which consists of k items from I, is frequent if α occurs in a transaction database D no lower than θ |D| times, where θ is a user-specified minimum support threshold (called min_sup), and |D| is the total number of transactions in D.

IV. PROPOSED WORK

Scan the transaction database to find the frequency of all size - 1itemsets. . In this step, we count each item's support by using compressed data structure, i.e. head and body of the database. Here body of the database contain itemset with their support and arranges in the lexicographic order, i.e. sorted order. Then erase all those size-1 itemsets of step 1 whose support is less than the MST. Then eliminate the infrequent item from each transaction. We will get a modified data base. Repeat the same procedure again and again, until there are elements to mine.

V. CONCLUSION

In this paper, we presented a novel compressed data structure based algorithm for mining frequent patterns. This algorithm efficiently mines all the possible frequent item sets from a transaction data base. The time required in the overall frequent item set mining is less in comparison to the existing algorithm.

REFERENCES

[1] E. Omiecinski, and S. Navathe. "An efficient algorithm for mining association rules in large databases". In Proc. Int'l Conf. Very Large Data Bases (VLDB), Sept. 1995, pages 432–443.

[2] Imielinski.t, Swami.A. "Mining Association Rules between Sets of Items in Large Databases". In Proc. Int'l Conf. of the 1993 ACM SIGMOD Conference Washington DC, USA.

[3] Agrawal.R "Fast algorithms for mining association rules". In Proc. Int'l Conf. Very Large Data Bases (VLDB), Sept. 1994, pages 487–499.

[4] Brin.S, Motwani. J.D, and S. Tsur. "Dynamic itemset counting and implication rules for market basket analysis". In Proc. ACM-SIGMOD Int'l Conf. Management of Data (SIGMOD), May 1997, pages 255–264.

[5] Park. J. S, M.S. "An effective hash-based algorithm for mining association rules". In Proc. ACM-SIGMOD Int'l Conf. Management of Data (SIGMOD), San Jose, CA, May 1995, pages 175–186.

[6] Toivonen "Sampling large databases for association rules". In Proc. Int'l Conf. Very Large Data Bases (VLDB), Sept. 1996, Bombay, India, pages 134–145.

[7] By Jiawei Han, Micheline Kamber, "Data mining Concepts and Techniques" by Morgan Kaufmann Publishers, 2006.

[8] Shaohua Teng, Wei Zhang, Haibin Zhu. "An Algorithm to Improve the Effectiveness of Apriori". In Proc. Int'l Conf. on 6th IEEE Int. Conf. on Cognitive Informatics (ICCI'07), 2007.

[9] Gu, C.-K., Dong, X.-L. "Efficient mining of local frequent periodic patterns in time series database", International Conference on Machine Learning and Cybernetic, pp. 183–186, 2009.

[10] Chowdhury Farhan Ahmed, Syed Khairuzzaman Tanbeer, ByeongSoo Jeong, Young-Koo Lee a,Ho-Jin Choi(2012) "Single-pass incremental and interactive mining for weighted frequent patterns", Expert Systems with Applications 39 pp.7976–7994, ELSEVIER 2012.