INTEGRATING COMPRESSION TECHNIQUE FOR ASSOCIATION RULE MINING

Shubham Singh[1], Sumitra Menaria[2]

[1] *Student,Computer Science & Engg., PIET Vadodara,Gujarat,India*
[2] *Asst. Prof., Computer Science & Engg.,PIET Vadodara,Gujarat,India*

ABSTRACT

Data Reduction in the field of process becomes a challenging issue in the data mining. Everyone wants that there data must be processed quickly in order to get the result. There are various methods that can allow the large data to be processed quickly or in lesser amount of time. In order to do so, one of the method which is widely used in most of the fields is, "Data Compression" which can easily made availability of the required space. Data mining has many useful applications in recent years because it can help users to discover the interesting knowledge in large databases. If the Data is already pre-processed then data compression over such type of data can easily be done in quick succession of time, Here we analyze the methods simple Apriori, and the Apriori over compressed database using the compression technique, existing compression algorithms are not appropriate for data mining due to lack of consistency of maintain the compressed database. Two different approaches were proposed in which, first we need to compress databases and then perform the data mining process that can promises to increased up the processing time in large databases by using the compression techniques. In this research an approach called Mining Merged Transactions with the Quantification Table is used by using the sorting techniques to put the same data types into one to solve these problems. The Apriori is used to find the frequent item sets in the database then the compressed database is maintained in the future as this table relationship of transactions is used to merge related transactions and builds a quantification table to prune the candidate item sets..

Keyword : - *Association Rule Mining, Apriori, Compression, quantification table, sorting*

1. INTRODUCTION

Now a day's uses of large data increases everywhere day by day. Data compression is one of good solutions to reduce the size of the data that can save the time of discovering useful knowledge by using appropriate methods, for example, data mining. Data mining is used to help users to discover interesting and useful knowledge from large database more easily.[1][13] There are many mining techniques available to discover the interesting and useful knowledge in which association rule mining is one of the ancient techniques which compromises of various different algorithms for finding the relation between two or more data sets present in a database, Apriori was the first algorithms that is useful for finding the frequent item sets in a database. It is more and more popular to apply the association rule mining in recent years because of its wide applications in many fields such as stock analysis, web log mining, medical diagnosis, customer market analysis, and bioinformatics. In this research, the main focus is on association rule mining and data pre-process with data compression.[11] Data pre-process transforms the original database into a new data representation. Eventually, it generates a new transaction database at the end of the data pre-process step. It uses an Apriori like algorithm of association rule mining to find frequent item sets. There are some problems in this approach. First, the database that get compressed is not reversible, after the original database is transformed by the data pre-process step by applying data compression. It is very difficult to maintain this database in the future because the availability of the frequent items in the database make it harder.. Second, although some rules can be mined from the new transactions, it still needs to scan the database again to verify the result.[7] This is because the data mining step produces potentially ambiguous results. It is a serious problem to scan the database multiple times because of the high cost of re-checking the frequent item sets[12].

2. LITERATURE SURVEY

Discussing brief about the existing systems with their benefit and limitation. The market-basket problem [1] assumes we have some large number of items, e.g., bread," "milk." Customers their market baskets with some subset of the items, and we get to know what items people buy together, even if we don't know who they are. Marketers use this information to position items, and control the way a typical customer traverses the store. Words appearing frequently together in documents may represent phrases or linked concepts. Can be used for intelligence gathering. Data Mining is the discovery of hidden information found in databases and can be viewed as a step in the knowledge discovery process [10]. Data mining functions include clustering, classification, prediction and link analysis (associations). One of the most important data mining applications is that of mining association rules. Association rules [7], Association rule mining [7] finds interesting association or correlation relationship among a large set of data items Association rule are used to predict the associability of two or more things together on the basis of analysis of available facts and figures. Most algorithms used to identify large itemsets can be classified as either sequential or parallel in most cases, it is assumed that the itemsets are identified and stored in lexicographic order On the other hand, parallel algorithms focus on how to parallelize the task of finding large itemsets. In the following subsections we describe important features of previously proposed algorithms. These different algorithms are used in association rule mining for finding the frequent items, data set partitioning which are

1. Apriori Algorithm
2. Partitioning
3. Dynamic Item set Counting
4. Parallel and Distributed Algorithms
5. Hybrid Distribution.

Mingjun et al. presents a novel algorithm for mining complete frequent item sets, and also introduce transaction mapping algorithms means each itemsets is mapped and compressed to a continuous transactions intervals in a different space and evaluated against FP Growth and dEclat.[2]. Marghny et al. presents technique for mining frequent itemsets, and a table techniques are also introduce here CountTableFI and BinaryCountTableF. All the transaction were represented in binary and decimal number in this new table technique. Hence, it is simple and fast to use subset and identical set properties. [4]. Mahmoud et al. suggest a new algorithm to compress transactions from uncertain database based on modified version of M2TQT (Mining Merged Transactions with the Quantification Table) approach and fuzzy logic concept. The algorithm bands the uncertain data to set of clusters using K-Mean algorithm and exploits fuzzy membership function to classify the transaction items as one of those clusters. uncertain data is probabilistic in nature and frequent item set is counted as expected values so, compressed transactions will give us approximate values for the item set's support. This technique focuses on compressing related uncertain transactions by collecting the uncertain data into set of clusters, [7]. Tekin Bicer et al. introduced the methodology to improve the performance of large scale data analytics applications. This methodology performs the data processing at middleware, which exploits the similarities between spatial and temporal neighbors in a popular climate simulation dataset and enables high compression ratios and low decompression costs. The compression methodology that is introduced and the framework that is required had been applied to three applications over two datasets including the Global Cloud-Resolving Model (GCRM) climate dataset [9]. Daniel J. et al. represents that how the use of column-oriented database system architecture invites us to re-evaluate the process of how and when data in databases is compressed. C-Store includes column-oriented versions of most of the familiar relational operators. [8]. From the studies ensuring that Apriori-like algorithms generate a lot of candidate item sets and need to check the candidate item sets by scanning the database. It is very time-consuming. And local transaction variation was not supported in the existing systems that were used in different methods.

3. A FRAMEWORK OF INTEGRATING TECHNIQUES TO PRUNE REDUNDANCY

The goal of the proposed algorithm is to take the advantages over Apriori like algorithm without suffering from the problem of checking candidate item sets again and again that can be done by applying sorting techniques before compression

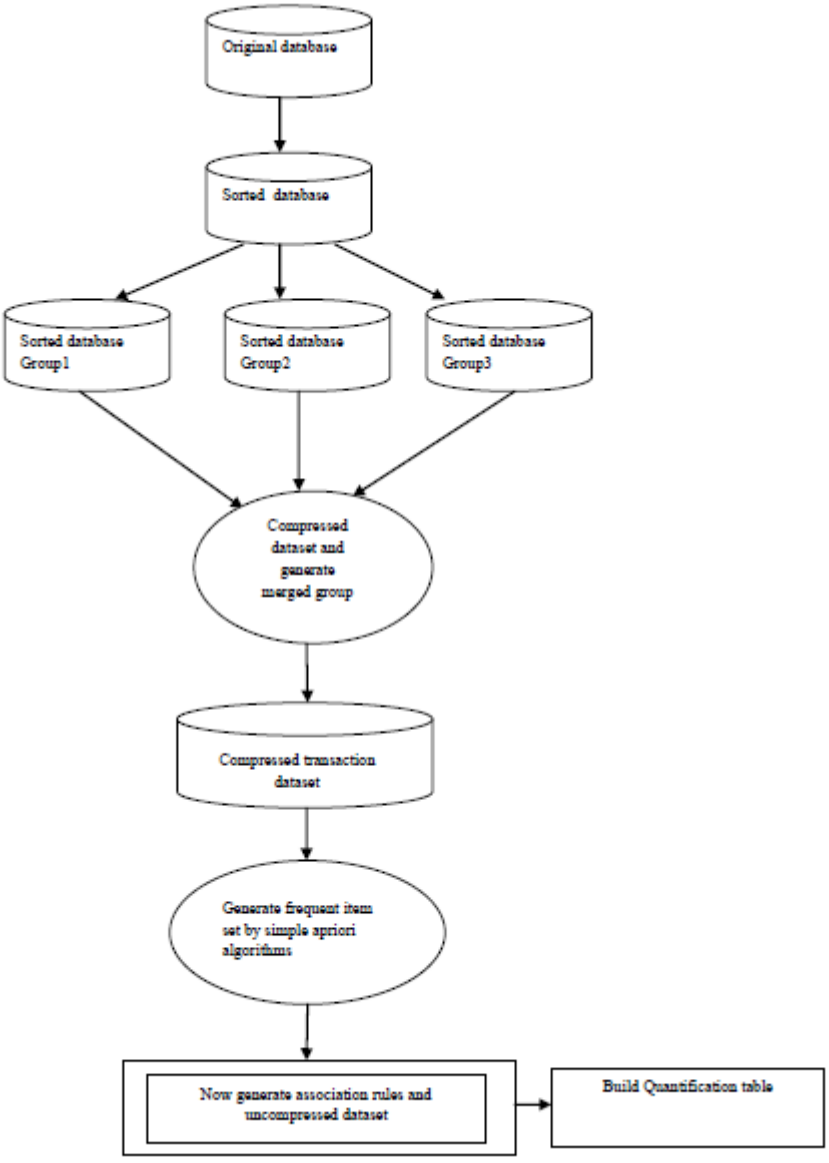

Fig. 1. Architectural work flow of proposed approach

The original transaction is taken as it is for performing the data preprocess i.e, on the original database we perform the data transformation (Data Pre-Process) and merge transaction. Then we will perform sorting by the use of sorting technique on the original database by performing data preprocess consisting of two sub processes. One sub-process transforms the original database into a new data representation. Another sub-process is sorting all the transactions to various groups of transactions and then merges each group into a new transaction. Then the sorted databases are compressed into one compressed dataset on which we need to generate the merged group that we will done in two phases: firstly, it is use to find the frequent itemsets. And in the second phase we need to prune redundancy. After then we will get a compressed transaction dataset over which we will discover frequent itemsets

using simple apriori algorithm. Then we will generate the association rule mining over the simple apriori algorithm and gets decompressed dataset to Support local transaction variation. Once the frequent itemsets gets generated then a quantification table needs to be built so that the number of items that are frequent in the dataset, can be pruned so that it is impossible to become frequent After performing the proposed scenario we can reduce the process time of association rule mining by using a quantification table. Reduce I/O time by using only the compressed database to do data mining

4. CONCLUSIONS

As the rapid increase of large data in all fields when come under process then the run time of entire process becomes a major part of concern while dealing with such type of large data sets. Regularly use of compression comes under play such type of role i.e, by compression we can reduce the size of the same amount of large data that can be processed easily. Hence, the proposed algorithm can enhancing compression technique while dealing with such large data sets as it can require such a effective sorting technique that can be useful to keep the same data types into one memory space in the data base by neglecting decompressed technique to reverse into original form. It can also reduce I/O time by using only the compressed database to do data mining, which can provide a better runtime of entire process by increasing the efficiency

.

5. REFERENCES

[1] Chanchal Yadav, Shuliang Wang, Manoj Kumar:" An Approach to Improve Apriori Algorithm Based On Association rule Mining". IEEE – 31661, 4th ICCCNT - 2013 July 4-6, 2013.

[2] Mingjun Song, and Sanguthevar Rajasekaran: "A Transaction Mapping Algorithm for Frequent Itemsets Mining" IEEE transaction on knowledge and data engineering revised october5,2005.

[3] Zalak V Vyas, Amit P. Ganatra2, Dr. Y.P.Kosta, C. K. Bhesadadia:"Modified RAAT (Reduced AprioriAlgorithm using Tag) for Efficiency Improvement with EP(Emerging Patterns) and JEP(Jumping EP)". 2010 International Conference on Advances in Computer Engineering

[4] Marghny H. Mohamed • Mohammed M. Darwieesh:" Efficient mining frequent itemsets algorithms". Received: 7 March 2012 / Accepted: 29 April 2013 Springer-Verlag Berlin Heidelberg 2013

[5] Wael A. AlZoubi, Azuraliza Abu Bakar, Khairuddin Omar:" Scalable and Efficient Method for Mining Association Rules". IEEE 2009 International Conference on Electrical Engineering and Informatics 5-7 August 2009, Selangor, Malaysia.

[6] Michail Vlachos ,Nikolaos M. Freris, Anastasios Kyrillidis:" Compressive mining: fast and optimal data mining in the compressed domain". Received: 18 December 2013 / Revised: 8 April 2014 / Accepted: 15 May 2014 © Springer-Verlag Berlin Heidelberg 2014

[7] Mahmoud M. Gabr, Saad M. Darwish, and Sayed A. Mohsin:" An efficient compression algorithm for uncertain databases aimed at mining problem". Lecture notes on software engineering, Vol. 3, No.2, May 2015, DOI:10.7633/LNSE.2015.V3.182

[8] Daniel J. Abadi, Samuel R. Madden, Miguel C. Ferreira:" Integrating Compression and Execution in Column Oriented Database Systems". SIGMOD 2006, June 27–29, 2006, Chicago, Illinois, USA.Copyright 2006 ACM 1595932569/06/0006.

[9] Tekin Bicer, Jian Yin, David Chiu, Gagan Agrawal, and Karen Schuchard:" Integrating Online Compression to Accelerate Large-Scale Data Analytics Applications". 2013 IEEE 27th International Symposium on Parallel & Distributed Processing.

[10] Fan Zhang, Yan Zhang, Jason Bakos:" GPApriori: GPU-Accelerated Frequent Itemset Mining". 2011 IEEE International Conference on Cluster Computing.

[11] Qiankun Zhao, Sourav S. Bhowmick:" Association Rule Mining: A Survey". Technical Report, CAIS, Nanyang Technological University, Singapore, No. 2003116 , 2003.

[12] Qian Wan and Aijun An:" Compact Transaction Database for Efficient Frequent Pattern Mining".

[13] Jiawei Han and Micheline Kamber: "Data Mining; Concepts and Techniques". Second Edition. 2006