IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 16, Issue 4, Ver. III (Jul – Aug. 2014), PP 112-115
www.iosrjournals.org

A Fast & Memory Efficient Technique for Mining Frequent Item Sets from a Data Set

Richa Mathur[1], Virendra Kumar[2]
[1]*(Computer Science, Suresh Gyan Vihar University, India)*
[2]*(Computer Science, Suresh Gyan Vihar University, India)*

Abstract: *Frequent/Periodic item set mining is a extensively used data mining method for market based analysis,privacy preserving and it is also a heart favourite theme for the resarchers. A substantial work has been devoted to this research and tremendous progression made in this field so far. Frequent/Periodic itemset mining is used for search and to find back the relationship in a given data set. This paper introduces a new way which is more efficient in time and space frequent itemset mining. Our method scans the database only one time whereas the previous algorithms scans the database many times which utilizes more time and memory related to new one. In this way,the new algorithm will reduced the complexity (time & memory) of frequent pattern mining. We present efficient techniques to implement the new approach.*

Keywords: *Incremental Association Rule Mining, Minimum Support Threshold(MST),Transactional Data set.*

I. Introduction

Data mining is the process of discovering and analyzing useful data from a large data set. The goal of the data mining process is to extract the useful information from a data set and transform it into an understandable structure for further use. It allows the user to analyze the data from various dimensions, categorize it and summarize the relationships identify. Data mining has emerged in various areas such as Customer relationship management (identify those who are likely to leave for a competitor), Banking (loan/credit card approval predict good customers based on old customers), Targeted marketing (identify likely responders to promotions), Fraud detection (telecommunications, financial transactions) etc. Data mining is the key part of Knowledge Discovery in Database (KDD)[1] [4] process. Data selection, data cleaning, data transformation, Data mining, finding presentation, finding interpretation, and finding evaluation are the steps involve in KDD process.

There are different kinds of method and techniques for data mining. Tasks in data mining can be classified as Summarization (relevant data is summarized and abstracted, resulting a smaller set which gives a overview of a data and usually with complete information) , Classification (it determines the class of an object based on its attributes), Clustering (identification of classes), Trend analysis, Regression and Deviation (Predictive mining), Association Rule Discovery[1] [2], Sequential Pattern Discovery (Descriptive mining). Data mining adopted its techniques from various research areas, including Statistical approach (Bayesian network), Machine learning, database systems, neural networks, rough sets, and visualizations. Predictive mining is the technique which is used to predict the unknown variables or future values of other variable and Descriptive mining is technique which is used to find the human-interpretable patterns that describes the data.

One of the major technique in data mining are Association rules. The most important task in association rule mining is to find the frequent/periodic patterns, associations, correlations, or casual structures among sets of items or objects in transaction or relational databases, and other information repositories [13]. In a given set of transactions, where transaction consists of items such as P and R then association rules are denoted as P=>R and intersection between them is null. The association rule can be useful for commodity management, marketing, etc. The support of this rule is defined by percentage of transaction that contains set P. And the Confidence of this rule is defined as percentage of these P transactions that also contain R. In Association rule mining, Frequent item set is an item set whose support is greater than the Minimum Support Threshold (MST). Minimum support threshold is a user defined support which is used to generate frequent items. Previously algorithms which are used to discover frequent patterns are static in nature. These algorithms are not able to work efficiently whenever any change happens to original database as in real world data is growing continuously. One solution of this algorithm is to reapply the algorithm on new database but in this case CPU utilization and time is very high and this approach is costly whenever small amount of data is inserted. Efficiency of these algorithms is based on number of passes as well as scans required for processing. A new algorithm was introduced to discover frequent items whenever new data is added dynamically to the original database. This algorithm was based on Generate and Test Method. In this method all possible candidates are generated and then tested for minimum support threshold (MST).

This paper presents a new incremental algorithm which is incremental in nature, Pattern- Growth approach is used. In Pattern-Growth approach, a frequent pattern of minimum size(1) is generated and then those patterns are used for finding frequent itemset. This may reduce the number of scans to the original database and the execution time is faster than the previous method.

II. Related Work

Apriori algorithm [3] and FP-tree [5] algorithm are the simplest approaches that were proposed to generate or discover the frequent item/pattern. Apriori algorithm is repetitive in nature, it is candidate-generation and test approach. FP-tree approach requires only two passes of processing whereas apriori algorithm required multiple passes for processing, so FP-tree amends the apriori disadvantage. It is based on FP-growth [5] algorithm which is based on divide conquer approach to gestate frequent itemsets.

Tree-based incremental Association rule mining algorithm (TIARM) was proposed by G.Pradeepini and S.Jyothi. Two different process has been used by this algorithm. One is to generate INC-tree to make tree more close-packed in nature and Second, frequent patterns of various sizes are recognized by applying TIARM on INC-tree. In this frequent pattern are generated without the use of candidate items. Liu Jian-ping et al [7] introduced an algorithm based on Fast Updated Frequent Pattern [6] concept. The main idea of FUFP is reuse of earlier mined frequent items to update with incremental database. In FUFP, all the links are bidirectional that is, it is to remove/add child node without much reconstruction whereas in FP-tree all links are in single direction. This algorithm was known as FUFP-tree based incremental association rule mining algorithm (Pre-FP). Chowdhury Farhan Ahmed et al. [9] have given two frequent itemset mining algorithm with single database scan. First one is weight, in which weight of different items are organized in upward order (i.e. IWFPwa). Another algorithm is dependent frequency which saves more memory space related to the earlier method as number of nodes are minimum and in this method frequency is arranged in decreasing order (IWFPfd).

Depend on Constant Incremental Updating Technique Siqing Shan et al. introduced Incremental Association Rules Mining Method. T-tree algorithm is applied on transaction database which works as FP-tree. Finally T-tree is inserted into the FP-growth algorithm to find out frequent pattern. D. Kerana Hanirex and Dr. M. A. Dorai Rangaswamy [10] have proposed clustering based incremental algorithm to discover frequent patterns. This algorithm has better efficiency than previous Apriori algorithm by reducing number of passes and memory space. Liu Han-bing, Zhang Ya-juan, Zheng Quan-lu and Ye Mao-gong [11] has proposed Incremental Frequent pattern mining algorithm based on AprioriTidlist Algorithm. When new data is added it discover frequent pattern using old frequent pattern. Shih-Sheng Chen et al. [8] have proposed a method for discovery of frequent periodic pattern using multiple minimum supports. All the items are arranges according to their MIS in transactions, it uses sorted closer property. Then it uses PFP (Periodic Frequent Pattern). This algorithm is more efficient but it does not support attribute uncertain data.

Jyoti Jadhav, Lata Ragha and Vijay Katkar [2] have recently proposed a Incremental Frequent Pattern Mining. This method presents a parallel frequent pattern generation mechanism using MMDBMS. It is divided into two types of modules: Central Co-ordinator module which divides the available dataset into number of available processors, and Parallel Processor module which generate frequent patterns and return the result to Central co-ordinator. In this Counting Based approach is used. It is based on Generate and Test method. In this all the possible candidates are generated first and then tested for MST (Minimum Support Threshold).

III. Proposed Method

In this paper, Pattern-Growth approach is used. In which a frequent pattern of minimum size(1) is generated and then these patterns are used for finding Frequent itemset.

Problem Statement:

Let I = {I1, I2, In} be a set of all items. A size k item set α, which consists of k items from I, is said to be 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 MST), and the |D| is the total number of transactions in D.

Proposed Algorithm:

Input:

- A Transaction Database D
- MST – Minimum support Threshold

Step1: First of all, scan the transaction database D to find out the number of occurrences of all size 1itemsets. Then the support of each item is computed & stored in a data structure. This data structure has two parts: the head part & the body part. The body of this data structure contains all single items with their support. 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.

Step 2: In this step all the size 1 items are arranged in the decreasing order of there support count. This is the candidate set of all size 1 items.
Step 3: In this step, the support of each size 1 item is compared with the minimum threshold known as the MST (Minimum Support Threshold). Eliminate all those size-1 itemsets of step 1 whose support is less than the MST. It will result in a compressed table, which consists of all the frequent items of size1. It is known as compressed data base (CDB)
Step 4: Now sort all the itemsets of last step in descending order of their item count (frequency)
Step 5: Create a 2 dimensional data structure (Table) and store the transaction and the correspondent frequency in that table.
Step 6: Then scan the data structure created in step 5 to locate all the K size itemsets. Select only those item sets whose support is greater than the minimum threshold (MST).
Step 7: If the support count of the k size item sets is less than the MST then take K size itemsets and k – 1 size itemsets together to generate a k - 1 size item set. Continue this step until no item set having support greater than the MST is found.
Step 8: All the largest possible size item sets are found in step 7 then by applying the downward closure property all the subsets are also frequent.
Step 9: Now scan the compressed table of step 5. It may contain some frequent item sets of smaller size which have yet not been included in the list of frequent item sets. Now reduce the data base of step 5 by considering only those transactions which contain frequent 1-itemset element but not contain the maximal frequent transaction.
Step 10: If no such transaction exists in table of step 5 then go to step 11 otherwise repeat step 5 to 10.
Step 11: exit.

Output: All the frequent item sets

IV. Experimental Result

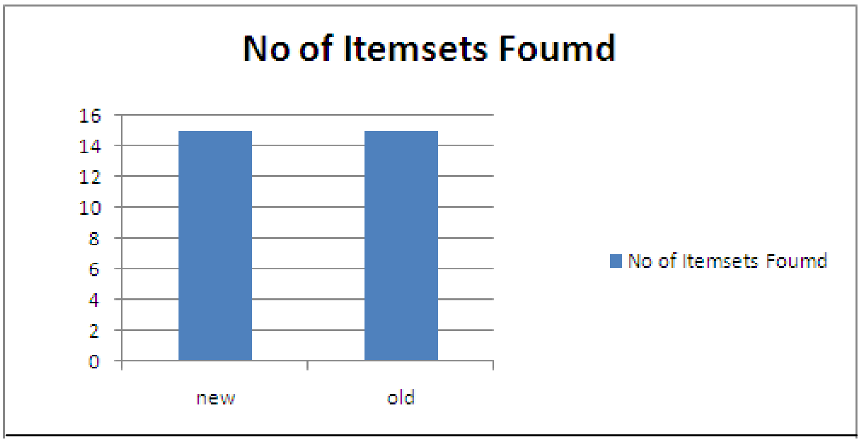

Figure1: Result Comparison

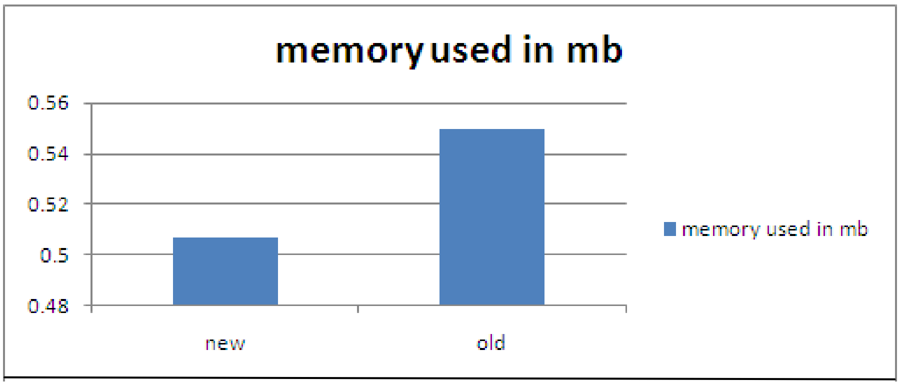

Figure2: Memory Comparison

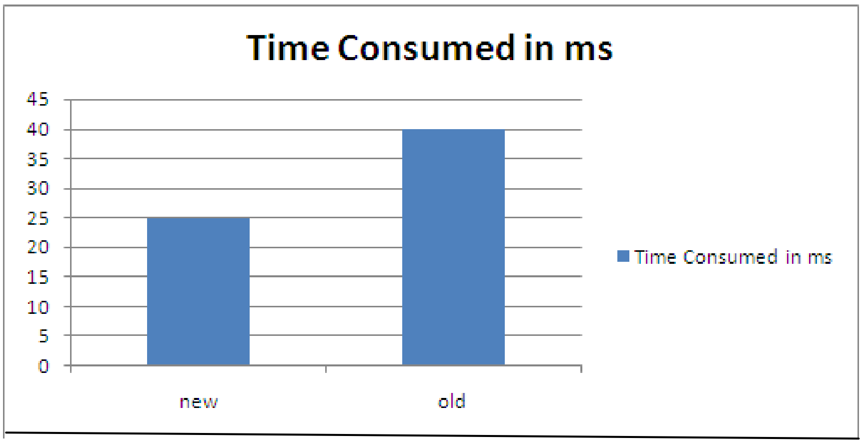

Figure3: Time Consumption Comparison

From the above comparison analysis it could be seen that the number of itemsets found in both the proposed/new and old method are same but the memory and time consumption is different. It could be seen clearly from the above graph that the time consumption in new algorithm/method is less as compared to the old method which is based on parallel processing and the memory is also consumed less by the new technique than the old method. So from all these it is clear that the proposed method is more effective than the old (Frequent Pattern Mining) method.

V. Conclusion

We review the list of existing frequent/periodic item set mining techniques. We limited ourselves to the classic frequent/periodic item set data mining problem. It is the generation of all frequent/periodic item sets that is available in market basket like data on the subject of minimal thresholds for support & confidence.Frequent itemset mining is crucial for association rule mining. We have assessed the execution of our proposed algorithm. It is fast. Also it is taking less main memory for computation in comparison to previous algorithm.

Refrences

[1]. Dr. Yongjian Fu," Data Mining: Applications,Tasks and Techniques ".

[2]. Jyoti Jadhav, Dr.Lata Ragha and Mr. Vijay Katkar, "Incremental Frequent Pattern Mining", IJEAT-2012

[3]. R. Agrawal, and R. Srikant, "Fast algorithms for mining association rules," Derive of 20th International Conference on Very Large Data Bases, Morgan Kaufmann, 1994.

[4]. Agrawal, R. and Psaila, G. Active Data Mining,Proceeding of 1st international Conference knowledge discovery and database, Montreal, 1995.

[5]. "Mining frequent patterns without candidate generation," J. Pei, Y. Yin and J.Han, The ACM SIGMOD International Conference on Management of Data, 2000

[6]. "The Pre-FUFP algorithm for incremental mining" Lin, C.-W., Hong, T. –P., & Lu, W. –H. (2009).Expert Systems with Applications.

[7]. Wang Ying, Yang Fan-ding and Liu Jian-ping and Wang Yang, "Incremental-Mining algorithm Pre-FP in association rules based on FP-tree", ICNDC, First international Conference, IEEE 2010.

[8]. "New and efficient knowledge discovery of partial periodic patterns with multiple minimum supports", Tony Cheng-Kui Huang, Zhe-Min Lin and Shih-Sheng Chen, The Journal of Systems and Software, 2011.

[9]. "Single-pass incremental and interactive mining for weighted frequent patterns", Syed Khairuzzaman Tanbeer, Byeong-Soo Jeong, Young-Koo Lee a,Ho-Jin Choi and Chowdhary Farhan Ahmed, Expert Systems with Applications 39 pp. ELSEVIER 2012.

[10]. D.Kerana Hanirex, Dr.M.A.Dorai Rangaswamy "Efficient Algorithm For Mining Frequent Itemsets Using Clustering Techniques", IJCSE, Vol. 3 No. 3 March-2011.

[11]. Zhang Ya-juan, Zheng Quan-lu, Ye Mao-gong and Liu Han-bing "New Incremental Updating Algorithm for Mining Association Rules Based on Apriori Transactional idList Algorithm", Cross Strait Quad-Regional Radio Science and Wireless Technology Conference (CSQRWC), Vol. 2, pp 1611 – 1614, IEEE 2011.