Mining of Frequent Closed Itemsets: A Review

Darshan Modi[1], Namrata Shroff[2]

[1] *Darshan Modi, M.E. Student, Computer Engineering, Government Engineering College Gandhinagar, Gujarat, India*
[2] *Namrata Shroff, Asst. Professor, Computer Engineering, Government Engineering College Gandhinagar, Gujarat, India*

ABSTRACT

Our capabilities of both generating and collecting data have been increasing rapidly. Mining frequent patterns is one of the most important concepts of data mining. Several algorithms have been developed for finding frequent item sets from the databases. The efficiency of these algorithms is a major issue since a long time. In data mining, association rule mining is one of the important techniques for discovering meaningful patterns from large collection of data. Discovering frequent item sets play an important role in mining association rules, sequence rules, web log mining and many other interesting patterns among complex data. The problem of mining association rules has attracted lots of attention in the research community. The most time consuming operation in discovering association rule, is the computation of the frequency of the occurrences of interesting subset of items (called candidates) in the database of transactions. Existing Apriori, FP Growth and Closet+ used with aim of improving the performances of high utility item sets. Main aim of this review is developing an efficient algorithm for finding frequent closed patterns. In this review paper we review some existing algorithms for frequent closed item set mining. The present paper provides an overview of various research papers on frequent closed itemset mining algorithms.

Keyword : - *Data mining, frequent closed itemset mining, frequent closed itemsets, Association rules*

1. Introduction

Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems. Data mining tools can answer business questions that traditionally were too time consuming to resolve. They scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations.

Association rules are one of the major techniques of data mining. Association rule mining finding frequent patterns, associations, correlations, or causal structures among sets of items or objects in transaction databases, relational databases, and other information repositories [13]. The volume of data is increasing dramatically as the data generated by day-to-day activities. Therefore, mining association rules from massive amount of data in the database is interested for many industries which can help in many business decision making processes, such as cross-marketing, Basket data analysis, and promotion assortment. The techniques for discovering association rules from the data have traditionally focused on identifying relationships between items telling some aspect of human behavior, usually buying behavior for determining items that customers buy together. All rules of this type describe a particular local pattern. The group of association rules can be easily interpreted and communicated.

Frequent patterns, such as frequent itemsets, substructures, sequences term-sets, phrasesets, and sub graphs, generally exist in real-world databases. Identifying frequent itemsets is one of the most important issues faced by the knowledge discovery and data mining community. Frequent itemset mining plays an important role in several data mining fields as association rules, warehousing, correlations, clustering of high-dimensional biological data and classification [13]. Given a data set d that contains k items, the number of itemsets that could be generated is 2k - 1,

excluding the empty set [11]. In order to searching the frequent itemsets, the support of each itemset must be computed by scanning each transaction in the dataset. A brute force approach for doing this will be computationally expensive due to the exponential number of itemsets whose support counts must be determined. There have been a lot of excellent algorithms developed for extracting frequent itemsets in very large databases. The efficiency of algorithm is linked to the size of the database which is amenable to be treated. There are two typical strategies adopted by these algorithms: the first is an effective pruning strategy to reduce the combinatorial search space of candidate itemsets (Apriori techniques). The second strategy is to use a compressed data representation to facilitate in-core processing of the itemsets (FP-tree techniques).

The problem of mining frequent itemsets are essentially, to discover all rules, from the given transactional database D that have support greater than or equal to the user specified minimum support.

A *closed itemset* refers to an itemset with support that does not equal that of any of its proper supersets. A *frequent closed itemset* [12] refers to a closed itemset with support that satisfies the minimum support. For any two closed itemsets, for example, X and Y and $X \subset Y$, if a closed itemset X' does not exist for $X \subset X' \subset Y$, then X is an *immediate closed subset* (IC subset) of Y, and Y is an *immediate closed superset* (IC superset) of X. For example, supposing that {ABCD}, {ABC}, {AB}, and {D} are closed itemsets, then {ABC} and {D} are IC subsets of {ABCD}, and {AB} is an IC subset of {ABC}. In other words, {ABC} is an IC superset of {AB}, and {ABCD} is an IC superset of {ABC} and {D}.

The *closure of an itemset* X is denoted as a closed itemset and is a superset of X with support equaling that of X [6]. Therefore, the support for the closure of Itemset X exceeds the support for all supersets of X. All frequent itemsets and their supports can be obtained from frequent closed itemsets. Therefore, by mining frequent closed itemsets, the mining time, storage space, and redundant information can be reduced.

2. Literature Survey

Marghny H. Mohamed et al in "Efficient mining frequent itemsets algorithms" [1] proposed countTable method. The countTable is used to employ subsets property for compressing the transaction database to new lower representation of occurrences items. In this paper, authors have developed this method to avoid the costly candidate generation-and-test processing completely. Moreover, the proposed methods also compress crucial information about all itemsets, maximal length frequent itemsets, minimal length frequent itemsets, avoid expensive, and repeated database scans. The proposed named CountTableFI and BinaryCountTableFI are presented, the algorithm has significant difference from the Apriori and all other algorithms extended from Apriori. The idea behind this algorithm is in the representation of the transactions, where, authors represent all transactions in binary number and decimal number, so it is simple and fast to use subset and identical set properties. It construct a highly compact count table, which is usually substantially smaller than the original database and discover frequent itemsets with Intersection And operation is faster than the traditional item comparing method used in many Apriori-like algorithm. These techniques are efficient and scalable. It generate super set maximum length frequent itemsets, works best in merge transaction and compact representation of count table of all occurrences each item table. It performs closely artificial dense datasets.

In **"Graph Based New Approach For Frequent Pattern Mining," Anurag Choubey [2]** used graph based approach. Graph is an efficient way to represent and understand the complex data, which has the potential to depict the values in a precise diagrammatic manner for dense and sparse databases. This research intends to propose a graph structure that captures only those itemsets that needs to define a sufficiently immense dataset into a submatrix representing important weights and does not give any chance to outliers. Authors have devised a strategy that covers significant facts of data by drilling down the large data into a succinct form of an Adjacency Matrix at different stages of mining process. The graph structure is so designed that it can be easily maintained and the trade off in compressing the large data values is reduced. In this paper, Authors used a graph-based approach so as to find the frequent pattern that repeatedly appears across various transactions. This Algorithm runs faster than the existing algorithms Apriori and FP growth because, it required only one database scan at each level. In this algorithm minimum support decreases and running time increases. Traditional algorithms are followed, less efficient in IO overheads. We take the base of FP-Growth for comprising with proposed algorithm.

"A Fast & Memory Efficient Technique for Mining Frequent Item Sets from a Data Set" by Richa Mathur et al in [3] introduces a new way which is more efficient in time and space frequent itemset mining. Method used in

this paper scans the database only one time whereas the previous algorithms scan 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. This paper presents efficient techniques to implement the approach. In this paper Minimum Support Threshold (MST) is used to generate frequent items. This 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. They support parallel processing and less memory consumption and more effective than old frequent mining pattern. We will follow the same for frequent itemset generation but using association rule mining classification with large database support.

Bay Vo et al in "A new method for mining Frequent Weighted Itemsets based on WIT-trees", [4] proposes algorithms for the fast mining of Frequent Weighted Itemsets (FWI) from weighted item transaction databases. Firstly, an algorithm for directly mining FWI using WIT-trees (Weighted Itemset-Tidset trees) is presented. After that, some theorems are developed concerning the fast mining of FWI. Based on these theorems, an advanced algorithm for mining FWI is proposed. Finally, a Diffset strategy for the efficient computation of the weighted support for itemsets is described, and an algorithm for mining FWI using Diffsets presented. In this paper WIT-FWI, WIT-FWI-MODIFY and WIT-FWI-DIFF algorithms are used & compared with Apriori. This algorithm is best suited for real time datasets but based on weighted item transaction datasets. This algorithm works only for weighted transaction item datasets and works only for high utility item sets.

Show-Jane Yen, "An efficient algorithm for incrementally mining frequent closed itemsets" [5] proposes an efficient algorithm for incrementally mining frequent closed itemsets without scanning the original database. The proposed algorithm updates closed itemsets by performing several operations on the previously closed itemsets and added/deleted transactions without searching the previously closed itemsets. The experimental results show that the proposed algorithm significantly outperforms previous methods, which require a substantial length of time to search previously closed itemsets. In this paper, Authors propose an algorithm named Maintenance of Frequent Closed Itemsets (MFCI), which includes two functions, MFCI-add and MFCI-del, for maintaining closed itemsets when transactions are added or deleted. When a transaction is added or deleted from a transaction database, the MFCI algorithm performs a few operations on the added/deleted transactions. Authors also proposed a closed itemset generator that generates closed itemsets that must be updated when a transaction is added. These algorithms works on generated closed itemsets and memory usage not clearly investigated. It uses more memory to the as compared to other algorithms so we will use some other reduced memory algorithms.

Jianyong Wang et al in "CLOSET+: Searching for the Best Strategies for Mining Frequent Closed Itemsets", [6] have mining frequent closed itemsets provides complete and non-redundant results for frequent pattern analysis. This paper also have proposed various strategies for efficient frequent closed itemset mining, such as depth-first search vs. breadth-first search, vertical formats vs. horizontal formats, tree-structure vs. other data structures, top-down vs. bottom-up traversal, pseudo projection vs. physical projection of conditional database, etc. This paper gives systematic study of the search strategies and develops a winning algorithm CLOSET+. A thorough performance study on synthetic and real data sets has shown the advantages of the strategies and the improvement of CLOSET+ over existing mining algorithms, including CLOSET, CHARM and OP, in terms of runtime, memory usage and scalability. CLOSET+ follows the popular divide-and-conquer paradigm and the depth-first search strategy. It uses FP-tree as the compression technique. In Closet+, a hybrid tree-projection method is introduced to improve the space efficiency. In CLOSET+, a hybrid tree-projection method is developed, which builds conditional projected databases in two ways: bottom-up physical tree-projection for dense datasets and top-down pseudo tree-projection for sparse datasets. CLOSET+ is a highly scalable and both runtime and space efficient algorithm for dense and sparse datasets, on different data distributions and support thresholds. This algorithm is suggested as generalized formation of mining frequent closed itemsets where tree building is purposed by FP tree and provides better scalability in terms of distinct items. This algorithm is employed to mining non redundant association rules which is yielding some complex results. We will try to modify it in terms of large number of itemsets including linkage analysis in large databases without yielding complex results.

Sujatha Dandu et al in [7] proposed modified APFT algorithm in "Improved Algorithm for Frequent Item sets Mining Based on Apriori and FP-Tree", which combines the Apriori algorithm and FP-tree structure of FP-growth algorithm. Authors have proposed to go one step further & modify the APFT to include correlated items & trim the non correlated itemsets. This additional feature optimizes the FPtree & removes loosely associated items from the frequent itemsets. Authors call this method as APFTC method which is APFT with correlation. APFTC includes the concept of correlation to filter (reduce) the association rules that not only satisfy the minimum support but also have liner relationships among them. The computational results verify the good performance of APFTC algorithm. The algorithm APFTC is working efficiently and in many cases, it's much faster than FP-Growth. In this paper, Complexity is managed in better way & works faster and set up with correlation concepts.

Hannu Toivonen et al in "Data Mining Applied to Linkage Disequilibrium Mapping", [8] introduce a new method for linkage disequilibrium mapping: haplotype pattern mining (HPM). The method, inspired by data mining methods, is based on discovery of recurrent patterns. The haplotypes are ordered by their strength of association with the phenotype, and all haplotypes exceeding a given threshold level are used for prediction of disease susceptibility–gene location. The method is model-free, in the sense that it does not require (and is unable to utilize) any assumptions about the inheritance model of the disease. Linkage disequilibrium mapping can be applied to data mining for simulated data sets, an implementation of the algorithm, and more detailed results can still is a topic to search.

3. CONCLUSIONS

Mining complete set of itemsets often suffers from generating a very large number of itemsets and association rules. Mining frequent closed itemsets provides an interesting alternative since it inherits the same analytical power as mining the whole set of frequent itemsets but generates a much smaller set of frequent itemsets and leads to less and more interesting association rules than the former.

The aim of this review paper is study how efficiently generating the frequently closed dataset, where we save time and memory both. In this paper, we study the principles for mining closed frequent sequential patterns and review some existing algorithms for frequent closed item set mining.

4. REFERENCES

[1] Marghny H. Mohamed, Mohammed M. Darwieesh, "Efficient mining frequent itemsets algorithms" Springer-Verlag Berlin Heidelberg 2013

[2] Anurag Choubey, Dr. Ravindra Patel, Dr. J.L. Rana, "Graph Based New Approach For Frequent Pattern Mining" International Journal of Computer Science & Information Technology (IJCSIT) Vol 4, No 1, Feb 2012

[3] Richa Mathur, Virendra Kumar, "A Fast & Memory Efficient Technique for Mining Frequent Item Sets from a Data Set" IOSR Journal of Computer Engineering (IOSR-JCE) , Volume 16, Issue 4, Ver. III (Jul – Aug. 2014), PP 112-115

[4] Bay Vo, Frans Coenen, Bac Le, "A new method for mining Frequent Weighted Itemsets based on WIT-trees" Elsevier, Expert Systems with Applications 40 (2013) 1256–1264

[5] Show-Jane Yen, Yue-Shi Lee, Chiu-Kuang Wang, "An efficient algorithm for incrementally mining frequent closed itemsets" Springer Science+Business Media New York 2013

[6] Wang J, Han J, Pei J (2003), "CLOSET+: searching for the best strategies for mining frequent closed itemsets." In: Proc of 9th ACM SIGKDD international conference on knowledge discovery and data mining, pp 236–245

[7] Sujatha Dandu, B.L.Deekshatulu, Priti Chandra, "Improved Algorithm for Frequent Item sets Mining Based on Apriori and FP-Tree" Global Journal of Computer Science and Technology Software & Data Engineering, Volume 13 Issue 2 Version 1.0 Year 2013

[8] Hannu Toivonen, Paivi Onkamo, Kari Vasko, Vesa Ollikainen, Petteri Sevon, Heikki Mannila, Mathias Herr and Juha Kere, "Data Mining Applied to Linkage Disequilibrium Mapping", The American Journal of Human Genetics 67.1 (2000): 133-145.

[9] Aggrawal.R, 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.

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

[11] A. Savasere, 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.

[12] Pasquier N, Bastide Y, Taouil R, Lakhal L, (1999) "Discovering frequent closed itemsets for association rules." In: Proc of 7th international conference on database theory, pp 398–416

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

[14] J. Pei, J. Han, and R. Mao., "CLOSET: An efficient algorithm for mining frequent closed itemsets." In DMKD'00, May 2000.