ISSN: 2321-7782 (Online)
Volume 2, Issue 1, January 2014
International Journal of Advance Research in Computer Science and Management Studies
Research Paper
Available online at: www.ijarcsms.com

A Review paper for mining Frequent Closed Itemsets

Dungarwal Jayesh M[1]
S.V.C.S.E.
Alwar
Rajasthan - India

Neeru Yadav[2]
Prof.
S.V.C.S.E.
Alwar
Rajasthan - India

Abstract: ***A problem faced at the time of finding a frequent closed item sets is duplication occurred. In this paper we are finding such algorithm which is finding and removing the duplicate items. Which is implemented this technique by using depth-first closed itemset mining algorithm. There is no need to stored whole itemset in the memory. We used vertical representation of data technique. This algorithm performs better than other algorithm. Before this algorithm CLOSET, CLOSET+, FPCLOSED algorithm is used.***

Keywords: Support, Confidence, Lift, Frequent closed itemset, vertical dataset, dense or sparse datasets.

I. INTRODUCTION

The frequent closed itemsets is the important part of data mining.so some concepts are required to find out frequent closed itemset.The first number is called the **support** for the rule. The support is simply the number of transactions that include all items in the antecedent and consequent parts of the rule. (The support is sometimes expressed as a percentage of the total number of records in the database.) The other number is known as the confidence of the rule. **Confidence** is the ratio of the number of transactions that include all items in the consequent as well as the antecedent (namely, the support) to the number of transactions that include all items in the antecedent. **Lift** is one more parameter of interest in the association analysis. Lift is nothing but the ratio of Confidence to Expected Confidence. It is a frequent itemset that is both closed and its support is greater than or equal to minsup. An itemset is closed in a data set if there exists no superset that has the same support count as this original itemset. First identify all frequent itemsets. Then from this group find those that are closed by checking to see if there exists a superset that has the same support as the frequent itemset, if there is, the itemset is disqualified, but if none can be found, the itemset is closed.An alternative method is to first identify the closed itemsets and then use the minsup to determine which ones are frequent.

The Algorithm CHARM is presented by Mohammed j. Zaki and Ching-jui Hsiao for mining all frequent closed itemset. It also uses a very important concept called diffsets to reduce the memory of intermediate computations it means that it does not required large memory to storing results of calculation required for this algorithm. It also uses a fast hash-based technique to remove any "nonclosed" item sets found during computation. In CHARM algorithm simultaneously studied about the itemset space and transaction space, rather than only the itemset search space. It also uses a highly efficient hybrid search method that skips many levels of the IT-tree to identify the frequent closed itemsets within a very less time, instead of having to enumerate many possible subsets. An extensive set of experiments confirms that CHARM provides orders of magnitude improvement over existing methods for mining closed itemsets. It makes a very less database scans than the longest closed frequent set found, and it scales linearly in the number of transactions and also is also linear in the number of closed itemsets found. But the problem of duplicate items is not solved and this will make this algorithm is not that much effective.

The Close algorithm is presented by Nicolas Pasquier, Yves Bastide, Rafik Taouiland Lotfi Lakhal for an efficient mining of association rules using closed itemsets lattices. Close is also another Apriori-like algorithm which is directly mines frequent

closed itemsets. At the start of this algorithm is to use bottom-up search technique to identify the smallest frequent itemset that determines a closed itemset. After finding the frequent k-itemsets, Close compares the support of each itemset set with its subsets at the previous level. When the support of an itemsets matches the support of any of its subsets, the itemsets is pruned. After that in Close algorithm is to calculate the closure of all the itemsets which is found at the very first step. The authors also created a variation of that algorithm, called A-CLOSE, which is also generates a reduced set of association rules without having to determine all frequent itemsets, thus automatically reduces the algorithm calculation costs. A-CLOSE calculates the closure of all the minimal generators previously found. Since a single equivalence class may have more than one minimal itemsets, duplicate closures may be calculated .because of that A-CLOSE performance suffers from the high cost of the off-line closure calculation and the large number of subset searches.

The CLOSET algorithm is a very efficient but highly complex technique. The CLOSET algorithm was designed to extract frequent closed itemsets from large databases. Presented by Jain Pei, Jiawei Han and Runying Mao, it offers a very efficient method for producing association rules. It reduces both the computational and cognitive cost in association rule analysis, by limiting the results to just frequent closed itemsets.

These algorithms start with scanning for frequent items. The algorithm that divides the frequent items by finding just the frequent closed itemsets. The CLOSET technique continues by recursively mining the subsets of the frequent item closed sets. The algorithm then effectively creates conditional databases of the frequent closed-items separately from the initial transactional database.

The actual mechanics of this process can become little complex. It begins by calculating the amount of support for items, and including any item above a minimum support level in a list, defined by the particular data being studied. The list of items meeting these criteria becomes the "f list" of frequently occurring items. The process of dividing the search space then takes each item and produces a new set for it, excluding each of the previous items. After that, the algorithm populates these sets by searching for items which fulfill the criteria for exclusion. This can be conceived of as creating a number of conditional databases of frequent closed itemsets.

The CLOSET approach is mainly identify for its efficiency, which is a result of total four optimization methods. The first of these methods is compressing both the original transactional database and the generated conditional databases into an FP-tree structure. FP-trees, also known as prefix trees, are constructed such that transactions with the same prefix share portions of the path down the tree. The details of this structure are complex; suffice to say that this effectively compresses the databases.

The next two optimizations extract items in different ways. The second extracts every item appearing in the conditional databases of the frequent item subsets. This helps reduce the size of the FP-tree and improves the overall speed of the recursive process by combining some items. The third exploits the natural structure of the FP-tree by directly extracting frequent closed itemsets. Since the items have been arranged by prefix, this allows for natural harvesting of the closed itemsets. The final method prunes out frequent items which have the same level of support and can be expressed as a subset of other itemsets.

The CLOSET algorithm is a very efficient but highly complex technique. It allows for reasonably fast mining of frequent itemsets from data with control over the number of rules or itemsets generated. For more details and a more rigorous description of the underlying mathematics, consult "CLOSET: An Efficient Algorithm for Mining Frequent Closed Itemsets," written by the originators of this technique.but the problem of duplication occurs at the time of finding frequent closed itemsets. And due this reason only CLOSET algorithm is not used that much in many application.

The CLOSET+ algorithm is presented by Jian Pei, Jiawei Han, and Jianyong Wang for Searching for the best strategies for mining frequent closed itemsets.CLOSET+ algorithm is a fast algorithm using the widely popular FP-tree (FrequentPattern tree) structure. It is composed of two phases. At the start of this algorithm, a compact representation of the database using FP-tree structure is built. After that second phase commences where the FP-tree is mined and frequent patterns are found. CLOSET+

makes use of several novel techniques for pruning the search space and thus increasing the mining speed.CLOSET+ performs quite well on the connect dataset with low supports, but in any other case it is about two orders of magnitude slower. CLOSET+ memory occupation grows exponentially because of the huge number of closed itemsets generated. Such technique is not used because of its inefficiency, and CLOSET+ steps back using the same strategy of CLOSET, i.e. storing every mined closed itemsets.

The DCI Closed: a Fast and Memory Efficient Algorithm to Mine Frequent Closed Itemsets is presented by Claudio Lucchese, Salvatore Orlando, Raffaele Perego. One of the main problem is occurs at the time of mining the frequent closed itemsets is the duplication. In this algorithm a general technique is used for find out and removes the duplicate closed itemsets, without the need of storing the whole closed itemsets in main memory. This is one of the very important features of this algorithm their approach can be exploited with substantial performance benefits by any algorithm that adopts a vertical representation of the dataset. This algorithm contains mainly three functions CLOSED SET, PRE SET, POST SET. From CLOSED SET new closed set, new generators and corresponding closed sets can be building. While the composition of POST SET guarantees that the various generators will be produced according to the lexicographic order. The composition of PRE SET guarantees that duplicate generators will be pruned by function is dup ().

II. CONCLUSION

DCI-Closed a Fast and Memory Efficient Algorithm to Mine Frequent Closed Itemsets algorithm gives best result because this technique is finding and removing duplicate itemsets without keeping to stored all itemsets in memory. Because this technique used vertical bitmap for represent the dataset. So this technique is very effective.

SR NO	Algorithms	Result
I	CHARM	It reduces the memory of intermediate computations
II	Close	It generates a reduced set of association rules without having to determine all frequent itemsets
III	CLOSET	Duplication occurs at the time of finding frequent closed Itemsets
IV	CLOSET+	CLOSET+ memory occupation grows exponentially because of the huge number of closed itemsets generated
V	DCI Closed	A general technique is used for find out and remove the duplicate closed itemsets

References

1. Mohammed J. Zaki and Ching-Jui Hsiao. Charm: An efficient algorithm for closed itemsetsmining. In 2nd SIAMInternational Conference on Data Mining, April 2002.
2. Jian Pei, Jiawei Han, and Runying Mao. Closet: An efficientalgorithm for mining frequent closed itemsets. In SIGMOD International Workshop on Data Mining and Knowledge Discovery, May 2000.
3. Jian Pei, Jiawei Han, and Jianyong Wang. Closet+: Searching for the best strategies for mining frequent closed itemsets. In SIGKDD '03, August 2003.
4. Claudio Lucchese, Salvatore Orlando, Raffaele Perego. DCI Closed: a Fast and Memory Efficient Algorithmto Mine Frequent Closed Itemsets.
5. Nicolas Pasquier, Yves Bastide, Rafik Taouil, Lotfi Lakhal Close algorithm for an efficient mining of association rules using closed itemset lattices.

AUTHOR(S) PROFILE

Dungarwal Jayesh M Pursuing M.tech (CSE) from S.V.C.S.E.alwar, Rajasthan and received the degree in Computer Engineering from Sinhgad Institute of Technology lonavala, Pune in 2011.