# A Compact Data Structure Based Technique for Mining Frequent Closed Item Sets

Kamlesh Ahuja, Durgesh Kumar Mishra(✉), and Sarika Jain

Sri Aurobindo Institute of Technology, Indore, India
Ahujakamlesh24@gmail.com, drdurgeshmishra@gmail.com,
sarika.jain@sait.ac.in

**Abstract.** Frequent pattern mining is top chart research field for young researchers. It has a huge array of real world applications. Although many algorithms, tools, techniques are available for performing the task of frequent pattern mining. Apriori and fp growth are very popular frequent pattern mining techniques. This paper presents an updated methodology for frequent closed item set mining. The proposed model is based on the concept of data reduction. Useless data is eliminated from the transaction data base. The experimental results have shown that the proposed updated method is outperforming the existing methods.

**Keywords:** Data mining · Frequent pattern mining · Frequent closed item sets · Data mart · Data warehouse

## 1 Introduction

In many cases it is useful to use support thresholds are minimum as possible. But, unfortunately, the number of extracted patterns grows exponentially as we decrease. It thus happens that the collection of discovered patterns is so large to require an additional mining process that should filter the really interesting patterns. Various data bases scattered around the world are integrated in to a data ware house. It is huge data repository this new database functions as a type of data mart (Fig. 1).

The same holds with dense datasets, such as census data. These contain strongly correlated items and long frequent patterns. In fact, such datasets are hard to mine even with high minimum support threshold. The Apriori property [2] does not remove the extent of candidates: every subset of a candidate is likely to be frequent. In conclusion, the complexity of the mining task becomes rapidly intractable by using conventional algorithms. Closed item sets are a solution to the problems described above. These are obtained by partitioning the lattice of frequent item sets into equivalence classes according to the following property: two distinct item sets belong the same class if and only if they occur in the same set of item sets. Closed item sets are the collection of maximal item sets of these equivalence classes. When a dataset is dense, the number of closed item sets extracted is order of magnitudes smaller than the number of frequent ones. This leverages the problem of the analyst of analyzing a large collection of patterns. Also, they reduce the complexity of the problem, since only a reduced search space has to be visited. For example, the pattern found within the sales knowledge of a food market


A. Unal et al. (Eds.): SmartCom 2016, CCIS 628, pp. 503–508, 2016.
DOI: 10.1007/978-981-10-3433-6_60

would indicate that if a client buys onions and potatoes along, he or she is probably going to additionally get hamburger meat. Such information are often used because the basis for decisions regarding marketing activities like, e.g. promotional evaluation or product placements. In addition to the above example from market basket analysis association rules are used these days in several application areas as well as web usage mining, bioinformatics and intrusion detection. As against sequence mining, association rule learning generally doesn't take into account the order of things either inside a transaction or across transactions (Fig. 2).

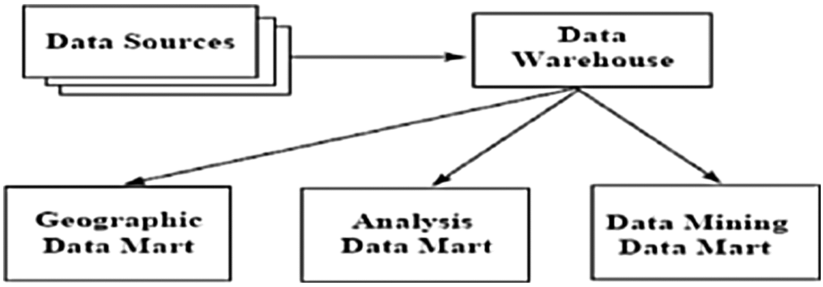


**Fig. 1.** Depicts that data warehouse and its relations with other streams

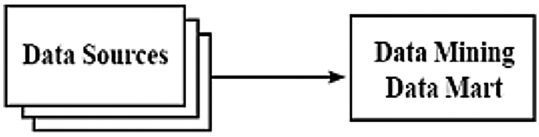


**Fig. 2.** Depicts that data warehouse and data mart

High performance data mining often tries to solve an expensive problem looking for an equivalent one that it is easier to solve. In fact, from closed item sets it is trivial to generate the whole gathering of frequent item sets along with their supports. In other words, frequent and closed frequent item sets are two different representations of the same knowledge. Moreover, recent FIM algorithms, use the concept of closed item sets to speed up their computation, and when possible they explicitly extract closed item sets and then generate frequent ones in a sort of post processing phase. The first of these kind of algorithms was Pascal [1, 7–9], and now any FIM algorithm uses a similar expedient. More importantly, association rules extracted from closed item sets have been proved to be more meaningful for analysts, because many duplicate items are discarded [2]. Suppose to have two frequent rules r1 : {diapers} -> {milk, beer} and r2 : {diapers} -> {milk} having the same support and confidence. In this case, the rules r1 is more informative since it includes r2: it tells something more about the implications of item diapers. Note that supp(diapers, milk) = supp(diapers, milk, beer), i.e. the two item sets occur in the same set of transactions and therefore they belong to the same equivalence class, but since r2 incudes r1 then {diapers, milk} is not closed. Thus, an algorithm based on closed item sets will not generate

the redundant rule r2. Something more about the implications of item diapers. Note that supp(diapers, milk) = supp(diapers, milk, beer), i.e. the two item sets occur in the same set of transactions and therefore they belong to the same equivalence class, but since r2 incudes r1 then {diapers, milk} is not closed. Thus, an algorithm based on closed item sets will not generate the redundant rule r2. This is why many algorithms for mining closed frequent item sets have been proposed, and why the idea of closed item sets has been borrowed by other frequent pattern mining tasks: there are algorithm for the extraction of closed sequences [6], closed trees [3], closed graphs [5], etc. The idea of closed item sets come from the application of formal concept analysis (Fig. 3).

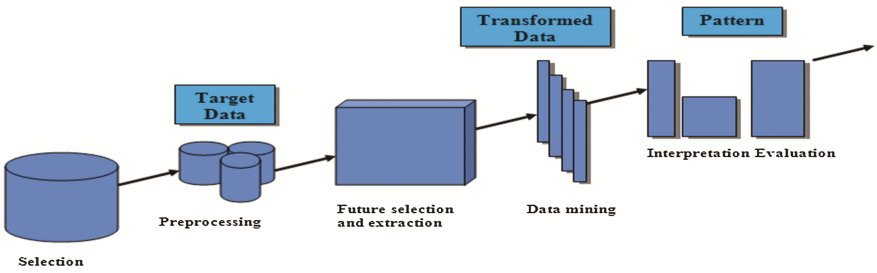


**Fig. 3.** Depicts that general concept of data mining

This was formalized in the early 80 s by Rudolf Wille [4] and years later it has found many application in data mining, information retrieval and artificial intelligence. Guo et al. [11] proposed a vertical variant of the a priori algorithm. In apriori, several scans of the data base are required. The author proposed a version of the improved a priori algorithm. In this version lesser scans of the data base are required.

## 2 Problem Definition

We are given a transaction data base D with user defined threshold. The problem is to find all the frequent closed patterns from d in such a way that they satisfies the minimum user defined threshold & also it uses less computational resources as compared to the existing technique.

## 3 Proposed Solution

Step 1: input:

1. a transaction database.
2. User defined Threshold.

Step 2: The transaction database scanned the whole database once and the count of each item is found.

Step 3: If count of any item of step 2 is less than user defined threshold then eliminate the infrequent item.

Step 4: Now arrange the frequent items found in step 3 in decreasing order of their count. It will be used in construction of the compact tree (CF-Tree).
Step 5: Construct CF tree by reading one transaction at a time.
Step 6: Extract a sub tree ending in an item (Suppose X).
Step 7:
- Check that the item of step 5 is frequent or not.
- If it is frequent then extract it as frequent item.
- New item X is frequent so now find the other frequent items ending with X.
- Continue this recursive procedure until no item found.

Step 8: Arrange the frequent item sets in the decreasing order of their size.
Step 9: For each frequent item sets having support more than the MST.

- Find all the super sets(S) of the frequent item stets.
- If any super set of the frequent item set is not having the same support as the frequent item set then add both in CFIS list.
- Otherwise add only superset in the CFIS list.

Step 10: Delete duplicate items from the CFIS list if any.
Step 11: Return CFIS.

## 4 Comparison Between Existing and Proposed Algorithm

The existing method work on the basis of generate and test method. It means that the algorithm first generates all the candidates of size 1 and then performs the pruning according to the MST. Then it generates all the candidates of size 2 and then perform the pruning according to the MST. The same process is repeated for the subsequent size elements.

The proposed method generates all the candidates of size 1 and then performs the pruning according to the MST. After that it eliminates all the infrequent items of size 1 from the data set to generate a new compact data set. Then this compact data structure is used to generate the subsequent size elements. So it will save time n space.

As shown in Figs. 4 and 5 Comparison based on the existing and proposed algorithm. This experiment use a Traffic Accidents Data Set. This data set of traffic accidents is obtained from the National Institute of Statistics (NIS) for the region of Flanders (Belgium) for the period 19912000. The traffic accident data contain a rich source of information on the different circumstances in which the accidents have occurred.

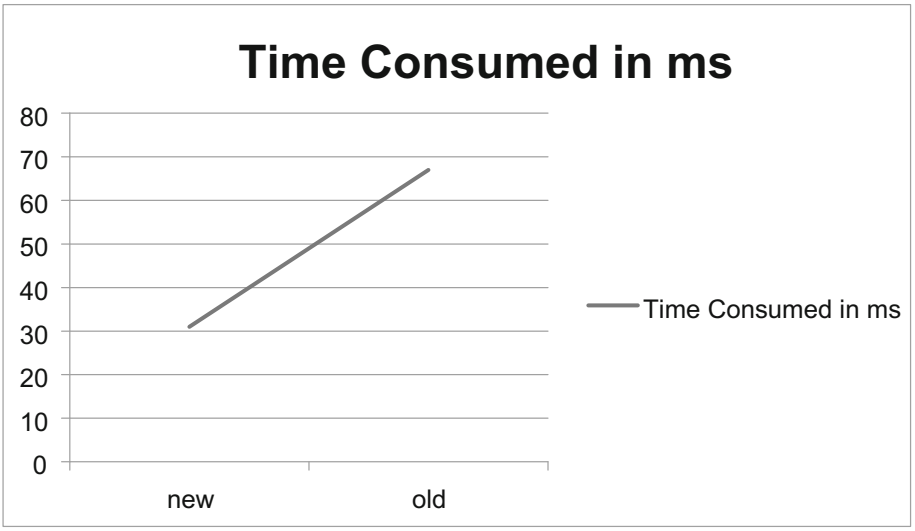


**Fig. 4.** Depicts the time consumption comparison

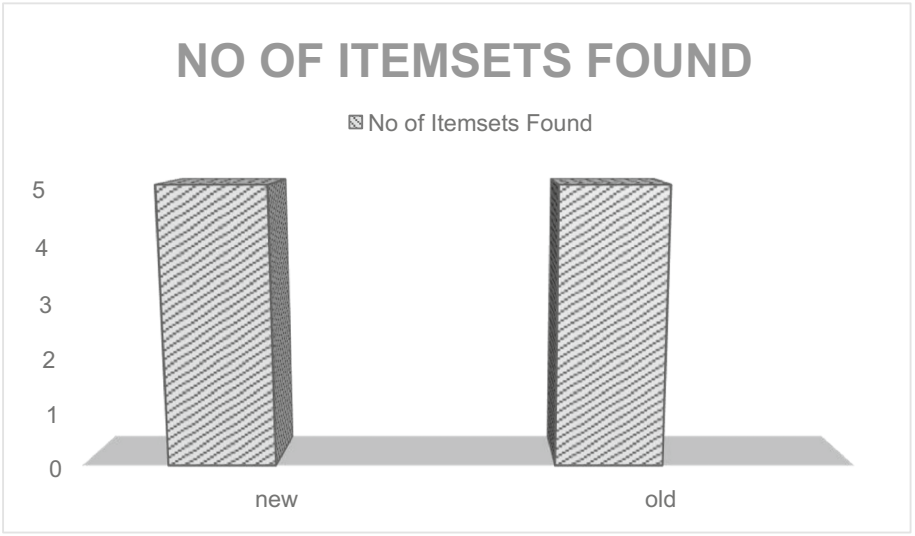


**Fig. 5.** Depicts the result comparison

## 5 Conclusion

The basic objective of frequent closed item set mining cum association rule mining is to find strong correlation among the items in the transaction data set. All the researchers are aware of the fact that they are required to deal with the voluminous data while performing mining on the data. So the goal is to device such algorithms which are time

and memory efficient. In this paper, we presented a novel algorithm for mining frequent closed item sets from a data sets. Frequent closed mining of data mining is used for that purpose. Frequent closed item set mining is crucial for association rule mining. We have evaluated the performance of our proposed algorithm. It is fast.

## References

1. Bastide, Y., Taouil, R., Pasquier, N., Stumme, G., Lakhal, L.: Mining frequent patterns with counting inference. SIGKDD Explor. Newsl. **2**(2), 66–75 (2000)
2. Chi, Y., Yang, Y., Xia, Y., Muntz, R.R.: CMTreeMiner: mining both closed and maximal frequent subtrees. In: Dai, H., Srikant, R., Zhang, C. (eds.) PAKDD 2004. LNCS (LNAI), vol. 3056, pp. 63–73. Springer, Heidelberg (2004). doi:10.1007/978-3-540-24775-3_9
3. Wille, R.: Restructuring lattice theory: an approach based on hierarchies of concepts. In: Rival, I. (ed.) Ordered Sets, pp. 445–470. Reidel., Dordrecht (1982)
4. Yan, X., Han, J.: Closegraph: mining closed frequent graph patterns. In: KDD 2003: Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 286–295, August 2003
5. Yan, X., Han, J., Afshar, R.: Clospan: mining closed sequential patterns in large datasets. In: SDM 2003: Proceedings of the Third SIAM International Conference on Data Mining, pp. 166–177, May 2003
6. Pasquier, N., Bastide, Y., Taouil, R., Lakhal, L.: Discovering frequent closed itemsets for association rules. In: ICDT 1999: Proceeding of the 7th International Conference on Database Theory, pp. 398–416, January 1999
7. Pei, J., Han, J., Mao, R.: Closet: an efficient algorithm for mining frequent closed itemsets. In: DMKD 2000: ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery, pp. 21–30, May 2000
8. Zaki, M.J., Hsiao, C.-J.: Charm: an efficient algorithm for closed itemset mining. In: SDM 2002: Proceedings of the Second SIAM International Conference on Data Mining, April 2002
9. Gouda, K., Zaki, M.J.: Genmax: an efficient algorithm for mining maximal frequent itemsets. Data Min. Knowl. Disc. **11**(3), 223–242 (2005)
10. Grahne, G., Zhu, J.: Efficiently using prefix-trees in mining frequent itemsets. In: FIMI 2003: Proceedings of the ICDM 2003 Workshop on Frequent Itemset Mining Implementations, November 2003
11. Guo, Y., Wang, Z.: A vertical format algorithm for mining frequent item sets. In: 2nd International Conference on Advanced Computer Control (ICACC), vol. 4, pp. 11–13 (2010)