A COMPACT DATA STRUCTURE BASED TECHNIQUE FOR MINING FREQUENT PATTERNS FROM WEB LOG DATA SET

Akanksha Gupta, Prof. Balwant Prajapat

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 web usage 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. . In this paper, we have developed a method to discover frequent web item sets from the web transaction database. The proposed method is fast in comparison to older algorithms. Also it takes les main memory space for computation purpose.**

Keywords**: data mining, frequent pattern mining, frequent closed item sets, data mart, data warehouse.**

1. Introduction

In many cases it is useful to use low minimum support thresholds. 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.

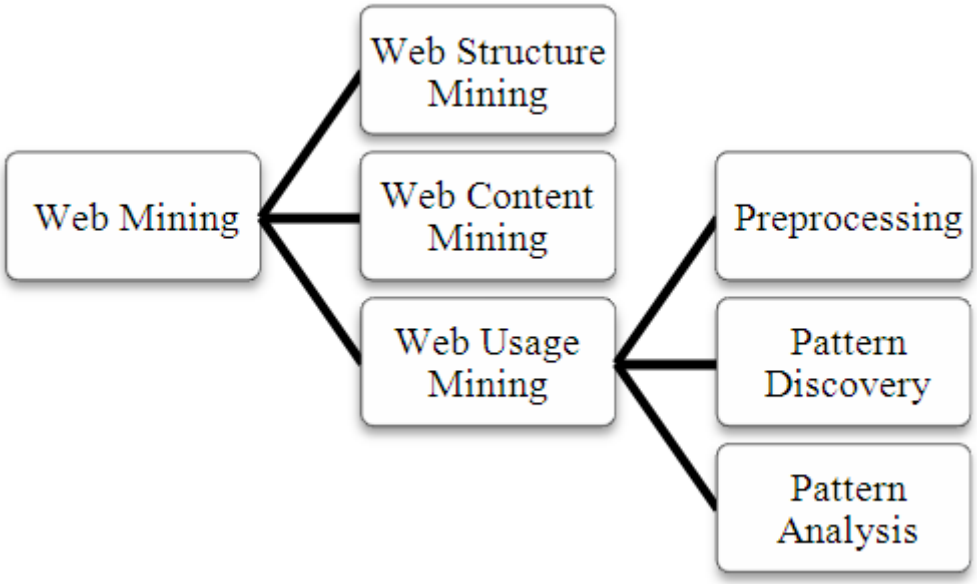

Figure 1: Web Mining Categorization

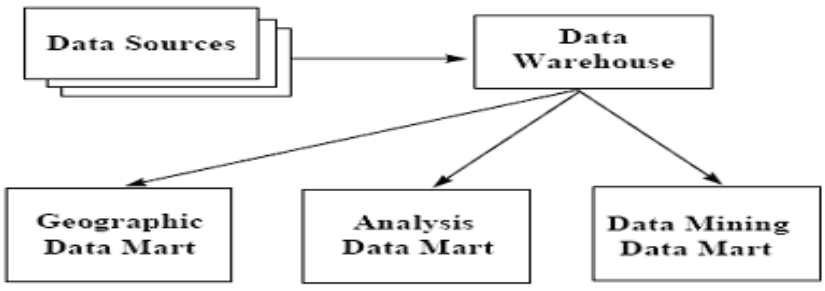

Figure. 2 Depicts that Data Warehouse and its Relations with Other Streams

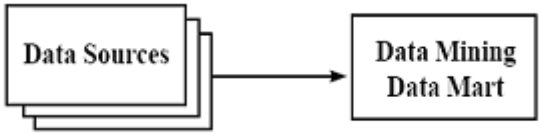

Figure. 3 Depicts that Data Warehouse and Data Mart

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 provide an effective pruning 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 transactions. 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 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.

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 collection 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 [1,7,8,9], 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 redundancies 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.
This was formalized in the early 80s 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: START
STEP 2: INPUT TRANSACTION DATA SET & MINIMUM SUPPORT THRESHOLD
STEP 3: FIRST THE PROPOSED ALGORITHM SCANS THE TRANSACTION DATA BASE AND CALULATES THE SUPPORT OF EACH SINGLE SIZE ITEM.
STEP 4: IN THIS STEP A LIST OF FREQUENT ITEM AND INFREQUENT ITEM IS PREPARED ON THE BASIS OF MINIMUM SUPPORT THRESHOLD.
IF AN ITEM IS HAVING SUPPORT GREATER THAN THE MINIMUM SUPPORT THRESHOLD THEN ITEM IS PLACED IN FREQUENT ITEM LIST AND ALSO IN EXPANSION LIST. OTHERWISE IT IS PLACED IN INFREQUENT ITEM LIST
STEP 5: REMOVE THE TRANSACTION WHICH DOES NOT CONTAIN ANY FREQUENT ITEM

STEP 6: IN THIS STEP, ALL THE MEMBERS OF THE INFREQUENT ITEM LIST ARE REMOVED FROM THE TRANSACTION DATA BASE BECAUSE THEY WILL NOT APPEAR IN ANY FREQUENT ITEM SET. IN THIS WAY, THE ORIGINAL TRANSACTION DATA BASE IS CONVERTED INTO REDUCED SIZE DATA BASE. NOW THIS REDUCED DATA BASE WILL BE USED IN THE CALCULATION OF LARGER SIZE FREQUENT ITEM SETS.
STEP 7: WHILE EXPANSION LIST IS NOT EMPTY
PERFORM LEFT EXPANSION OF SMALLER SIZE ITEMS TO GENERATE HIGHER SIZE ITEMS AND THEN REPEAT STEP 4 FOR THEM
OR
PERFORM RIGHT EXPANSION OF ELEMENTS AND THEN REPEAT STEP4 FORTHEM.
STEP 8: WRITE THE LIST OF FREQUENT ITEM SETS
STEP 9: STOP

4. Comparison between existing and proposed algorithm

The existing method is based on the concept 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.

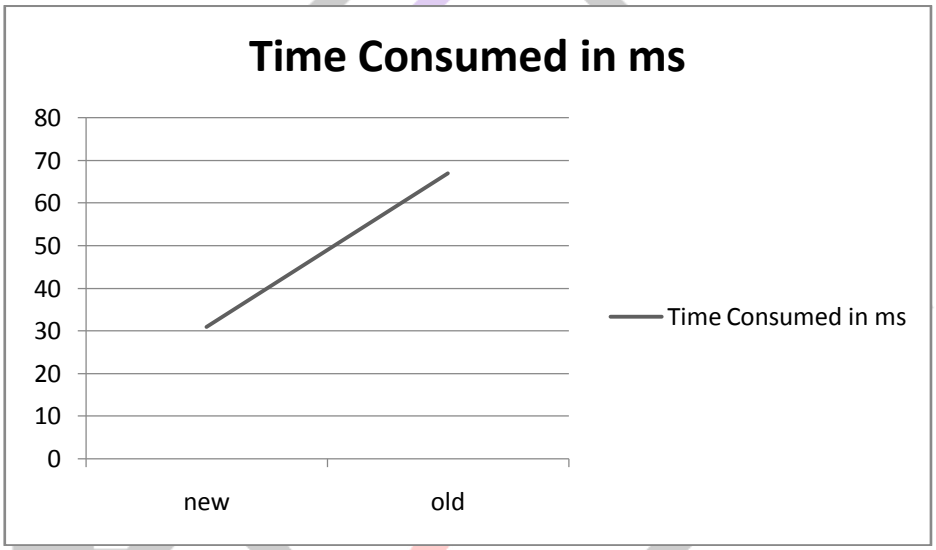

Figure. 4 Depicts the Time Consumption Comparison

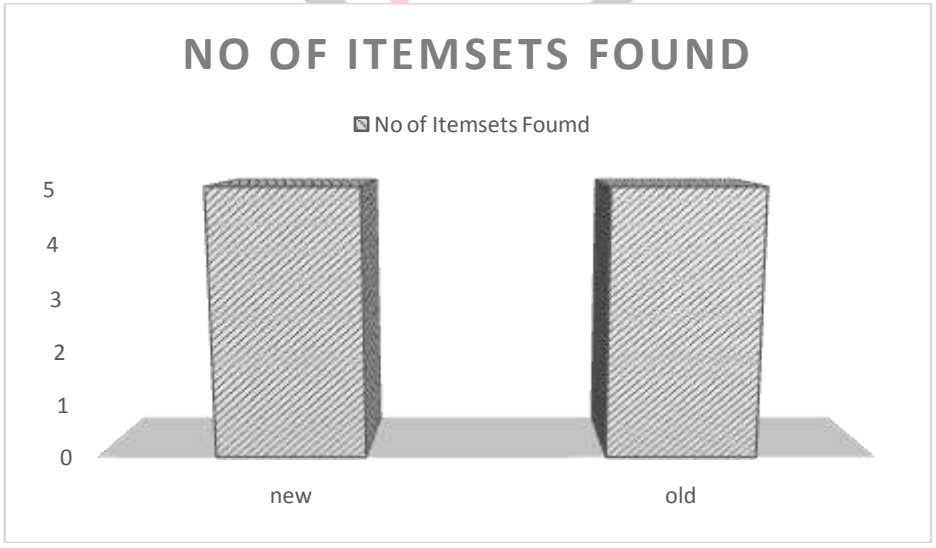

Figure. 5 Depicts the Result Comparison

As shown in fig.4 and fig.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 1991-2000. The traffic accident data contain a rich source of information on the different circumstances in which the accidents have occurred.

5. Conclusion

The basic objective of web usage 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. Y. Bastide, R. Taouil, N. Pasquier, G. Stumme, and L. Lakhal. Mining frequent patterns with counting inference. SIGKDD Explorations Newsletter, 2(2):66–75, December 2000.
2. Y. Chi, Y. Yang, Y. Xia, and R. R. Muntz. CMTreeMiner: Mining both closed and maximal frequent subtrees. In PAKDD ’04: Proceeding of the eighth Pacific Asia Conference on Knowledge Discovery and Data Mining, pages 63–73, May 2004.
3. R. Wille. Restructuring lattice theory: an approach based on hierarchies of concepts. In I. Rival, editor, Ordered sets, pages 445–470, Dordrecht–Boston, 1982. Reidel.
4. X. Yan and J. Han. Closegraph: mining closed frequent graph patterns. In KDD ’03: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, pages 286–295, August 2003.
5. X. Yan, J. Han, and R. Afshar. Clospan: Mining closed sequential patterns in large datasets. In SDM ’03: Proceedings of the third SIAM International Conference on Data Mining, pages 166–177, May 2003.
6. N. Pasquier, Y. Bastide, R. Taouil, and L. Lakhal. Discovering frequent closed itemsets for association rules. In ICDT ’99: Proceeding of the 7th International Conference on Database Theory, pages 398–416, January 1999.
7. J. Pei, J. Han, and R. Mao. Closet: An efficient algorithm for mining frequent closed itemsets. In DMKD ’00: ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery, pages 21–30, May 2000.
8. M. J. Zaki and C.-J. Hsiao. Charm: An efficient algorithm for closed itemset mining. In SDM ’02: Proceedings of the second SIAM International Conference on Data Mining, April 2002.
9. K. Gouda and M. J. Zaki. Genmax: An efficient algorithm for mining maximal frequent itemsets. Data Mining and Knowledge Discovery, 11(3):223–242, 2005.
10. G. Grahne and J. Zhu. Efficiently using prefix-trees in mining frequent itemsets. In FIMI ’03: Proceedings of the ICDM 2003 Workshop on Frequent Itemset Mining Implementations, November 2003.
11. Guo Yi-ming and Wang Zhi-jun, “A vertical format algorithm for mining frequent item sets,” 2nd International Conference on Advanced Computer Control (ICACC), Vol. 4, pp. 11 – 13, 2010.