Middle-East Journal of Scientific Research 23 (Sensing, Signal Processing and Security): 377-380, 2015
ISSN 1990-9233
© IDOSI Publications, 2015
DOI: 10.5829/idosi.mejsr.2015.23.ssps.207

A Novel Approach for Infrequent Weighted Itemset Mining Without Candidate Generation

T. Praveen and J. Wisely Joe

St. Peter's College of Engineering and Technology, Chennai, India

Abstract: Mining association rules is an important problem in the field of data mining. Though most of the previous work has been done on finding frequent itemsets for rule generation, infrequent itemsets mining has grown with its applications in knowledge discovery, business intelligence and other fields. Infrequent itemset mining is discovering itemsets whose frequency of occurrences in the given dataset is less than or equal to a maximum threshold instead of minimum threshold in frequent itemset mining. The number of occurrences of the item is expressed in terms of the support count and only that decides the rule is strong or weak. In this paper we also added a weight constraint for each item based on it's significance. The weight for each item is given with input dataset and plays a major role in identification of profitable itemsets. Infrequent weighted itemset mining has acquired significant usage in data analysis in recent years. Our proposed system concentrates on infrequent weighted itemset mining without candidate generation to improve the performance of existing algorithm.

Key words: Associations · Pattern Mining · Infrequent Itemset Mining

INTRODUCTION

Data mining is the process of extracting interesting and previously unknown information, useful patterns from a large set of databases. Data mining tasks are Classification-predicting the rules, Clustering-finding the clusters in data, Associations, Visualization, Summarization, Deviation Detection, Estimation and Link analysis. Itemset mining is an important data mining association technique widely used for discovering valuable relations among data. The first attempt to perform itemset mining was focused on discovering frequent itemsets, *i.e*, patterns whose support in given data is above a given threshold. Frequent itemsets are used in a number of real-life scenarios (e.g., market basket analysis, healthcare, DNA analysis, biological data analysis). However, many traditional approaches ignore the influence of each item in the set of transactions while analyzing data. A new measure weight is associated with each data item and characterizes its significance within transactions. Weight is the, most important component and without which it may lead to lose of useful information. In some papers weights are not considered during frequent itemset generation. It is introduced only when we generate rules and the algorithm produces high quality association rules.

The frequent itemset mining algorithms can be broadly classified into two categories: (i) candidate generation paradigm and (ii) pattern-growth paradigm. In previous studies, experimentally it has been shown that pattern-growth based algorithms are computationally faster on dense datasets as it uses a tree data structure to store data and has no candidate generation. However, significantly less attention has been paid to mining of infrequent itemsets, even though it has got important usage in mining of negative association rules from infrequent itemsets, statistical disclosure risk assessment where rare patterns in anonymous census data can lead to statistical disclosure [1], fraud detection where rare patterns in financial data may suggest unusual activity associated with fraudulent behavior [1] and bioinformatics where rare patterns in microarray data may suggest genetic disorders. In proposed method, we concentrate on pattern-growth paradigm to find infrequent weighted itemsets. We propose an optimization on the Apriori algorithm to mine minimally infrequent

Corresponding Author: T. Praveen, St. Peter's College of Engineering and Technology, Chennai, India.

weighted itemsets, *i.e*, the infrequent weighted itemsets, from weighted transactional data sets. To address this problem, the IWI-support measure instead of support measure is defined as a weighted frequency of occurrence of an itemset in the analyzed data. In paper [2], weight is associated with each data item and shows its local significance within each transaction. In our approach. Occurrence weights are derived from the user defined weights associated with each item in transaction dataset [3]. There are two different IWI-support measures are used:

- The IWI-support-min measure, which is based on a minimum cost function, *i.e*, the occurrence of an itemset in a given transaction is weighted by the weight of its least interesting item,
- The IWI-support-max measure, which is based on a maximum cost function, *i.e*, the occurrence of an itemset in a given transaction is weighted by the weight of the most interesting item.

Minimum and maximum are the main cost functions in optimization problems. So it is very much suitable for finding infrequent weighted data correlations. The following problems have been addressed and discussed:

- IWI and Minimal IWI mining driven by a maximum IWI-support-min threshold and
- IWI and Minimal IWI mining driven by a maximum IWI-support-max threshold.

To achieve above methods, we represent two algorithms which are, improved infrequent weighted itemset miner and improved minimal infrequent weighted itemset miner. The algorithm performs the IIWI and IMIWI mining driven by the IWI support thresholds. Works done on these, discussed in following sections.

Problem Description: Consider I = {x_1, x_2,.. . , x_n} to be a set of items. An itemset X ◎ I is a subset of items. If its length is k, it is referred to as a k-itemset. A transaction T is a tuple (tid, X) where tid is the transaction identifier and X is an itemset. X is said to contain an itemset Y if and only if Y ◎ X. A transaction database T D is simply a set of transactions.

Each itemset has an associated measure called support. For an itemset X, support (X, T D) = X.count where X.count is the number of transactions in T D with item X. For a user defined support threshold S, an itemset is frequent if and only if its support is greater than or equal to S. Otherwise it is infrequent.

In some T D, the number of infrequent itemsets may be large. It may be difficult to generate and report all of them. A key observation here is the fact that if an itemset is infrequent, so will be all its supersets. Thus, it is enough to generate only minimal infrequent itemsets, *i.e*, those which are infrequent but whose all subsets are frequent.

Definition 1: *(Minimally Infrequent Itemset)*- An itemset X is said to be minimally infrequent for an user defined support threshold S if it is infrequent and all its proper subsets are frequent itemsets, *i.e*, supp (X) < S and Y ◎ X, supp(Y) = S.

Given a particular support threshold, our goal is to effectively generate all the minimally infrequent itemsets (MIIs) using the pattern-growth algorithm.

Related Works: Frequent itemset mining is an important data mining task that has been introduced in [4]. This is very much used in knowing the purchasing behaviour of the customers in turn increasing the profit. In the traditional itemset mining problem items present in a transactional database are treated equally. To give importance or preference to the items based on their interest or profit within each transaction, in [2] the authors focus on finding more informative association rules, *i.e*, the weighted association rules (WAR), which include weights for every item.Weight shows the significance of item. Only after performing the traditional frequent itemset mining process, weights are introduced during the rule generation. The first attempt to give item weights into the itemset mining procedure has been done in [2]. It proposes to exploit the downward closure property of the proposed weighted support constraint to drive the Apriori-based itemset mining phase. However, in [2], [5] weights have to be previously known before doin analysis, while, in many real-life cases, this might not be the case. To solve this problem, in [6] the transactional data set is represented as a bipartite graph and weight is calculated by using well-known indexing strategy, *i.e*, HITS [6], in order to automate item weight assignment. Weighted item support and confidence quality indexes are defined accordingly and used for driving the itemset and rule mining phases. This paper differs from the above-mentioned approaches because it focuses mainly on mining infrequent itemsets from weighted data instead of frequent ones.. A new effort has been done to discover rare relations among data, *i.e*, the infrequent itemset mining problem [7, 3, 8 ,9].

In [7, 3] a recursive algorithm for finding minimal, distinct itemsets from well-structured data sets with exact support value 1, is proposed. in [6] the issue of discovering minimal infrequent itemsets was addressed, *i.e*, the itemsets that satisfy a maximum support threshold and do not contain any infrequent subset, from transactional data sets. Recently, in [9] an FP-Growth-like algorithm for mining minimal infrequent itemsets has also been proposed with a slightly modified tree structure. Unlike all the discussed approaches, we discuss the issue of treating items differently, based on their importance in every transaction or in entire T D, in the extraction of infrequent itemsets from weighted items.

Proposed Work

The Improved Infrequent Weighted Itemset Miner Algorithm: Transactional data set with weights for every item, a maximum IIWI-support (IIWI-support-min or IIWI-support-max) threshold are given. This algorithm extracts all IIWIs whose IIWI-support satisfies the threshold. We enforce either IIWI-support-min or IIWI-support-max thresholds, steps are same for the IIWI Miner mining algorithm. IIWI Miner is an algorithm for mining infrequent weighted itemsets like FP-Growth algorithm used for mining frequent patterns. The same tree structure is used like FP-Tree, which reduces the computational cost as it has only limited number of scans on the entire T D.To achieve IIWIM, some modifications are done in FP-Growth algorithm.The changes done in algorithm makes better pruning of itemsets and produce strong associations between items. To reduce the complexity of the pruning and mining process, IIWI Miner adopts an FP-tree pruning strategy to discard items that could never be a part of itemset satisfying the IWI-support threshold. If an item is pruned if it appears only in tree paths from leaf to the root node characterized by IWI-support value greater IWI-support threshold.

The Improved Minimal Infrequent Weighted Itemset Miner Algorithm: Transactional data set with weights for every item, a maximum IIWI-support (IIWI-support-min or IIWI-support-max) threshold are given.The Minimal Infrequent Weighted Itemset Miner algorithm extracts all the MIWIs that satisfy the threshold.This algorithm focuses on generating only minimal infrequent patterns.The recursive extraction in the MIWI Mining procedure is stopped when an infrequent itemset comes.For every infrequent itemset, all its extensions are not minimal.

CONCLUSION

This paper solves the problems in discovering infrequent itemsets by using weights. Weights are used for differentiating between related items and not within each transaction. Two FPGrowth- like algorithms that accomplish IIWI and IMIWI mining efficiently are also proposed. The usefulness of the pruned patterns has to be validated on data coming from a real-life scenario. As future work, we plan to add multiple minimum support thresholds for pruning good patterns that really focus on profit. Furthermore, the application of different aggregation functions besides minimum and maximum will be studied.

REFERENCES

1. Haglin, D.J. and A.M. Manning, 2007. “On Minimal Infrequent Itemset Mining,” Proc. Int’l Conf. Data Mining (DMIN ’07), pp: 141-147.
2. Tao, F., F. Murtagh and M. Farid, 2003. “Weighted Association Rule Mining Using Weighted Support and Significance Framework,” Proc. nineth ACM SIGKDD Int’l Conf. Knowledge Discovery and Data Mining (KDD ’03), pp: 661-666,.
3. Han, J., J. Pei and Y. Yin, 2000. “Mining Frequent Patterns without Candidate Generation,” Proc. ACM SIGMOD Int’l Conf. Management of Data, pp: 1-12.
4. Agrawal, R., T. Imielinski and Swami, 1993. “Mining Association Rules between Sets of Items in Large Databases,” Proc. ACM SIGMOD Int’l Conf. Management of Data (SIGMOD ’93), pp: 207-216.
5. Sun, K. and F. Bai, 2008. “Mining Weighted Association Rules Without Preassigned Weights,” IEEE Trans. Knowledge and Data Eng., vol. 20, 4: 489-495.
6. J.M., 1999. “Authoritative Sources in a Hyperlinked Environment,” J. ACM, vol. 46, 5: 604-632.
7. Chui, C.K., B. Kao and E. Hung, 2007. “Mining Frequent Itemsets from Uncertain Data,” Proc. 11th Pacific-Asia Conf. Advances in Knowledge Discovery and Data Mining (PAKDD ’07), pp: 47-58.

8. Manning, A. and D. Haglin, 2005. “A New Algorithm for Finding Minimal Sample Uniques for Use in Statistical Disclosure Assessment,” Proc. IEEE Fifth Int’l Conf. Data Mining (ICDM ’05), pp: 290-297, 2005. International Conference on, vol. 3. IEEE, pp. 486-491.

9. Gupta, A. Mittal and A. Bhattacharya, 2011. “Minimally Infrequent Itemset Mining Using Pattern-Growth Paradigm and Residual Trees,” Proc. Int’l Conf. Management of Data (COMAD), pp: 57-68.