ISSN: 2321-7782 (Online)

Volume 2, Issue 11, November 2014

International Journal of Advance Research in Computer Science and Management Studies

Research Article / Survey Paper / Case Study

Available online at: www.ijarcsms.com

A Review on Efficient Mining Approach of Infrequent Weighted Itemset

Sonia Jadhav[1]
P.G. Scholar Department of CSE
BSIOTR, Wagholi
Pune, India

G. M. Bhandari[2]
Professor, Department of CSE
BSIOTR, Wagholi
Pune, India

Abstract: Itemset mining is a data mining method extensively used for learning important correlations among data. The aim of Association Rule Mining is to find the correlation between data Items based on frequency of occurrence Infrequent Itemset mining is a variation of frequent itemset mining where it finds the uninteresting patterns i.e., it finds the data items which occurs very rarely. Considering weight for each distinct items in a transaction independent manner adds effectiveness for finding frequent itemset mining. This paper focus on review various Existing Algorithms related to frequent and infrequent itemset mining which creates a path for future researches in the field of Association Rule Mining.

Keywords: Clustering, association rule, weighted itemset, infrequent itemset mining, weight, Correlation.

I. INTRODUCTION

Data is a facts, contents, numbers or text that can be processed by a machine. The patterns, associations or the relationship among all this data can provide information. Information can be converted into knowledge about historical patterns and future trends. Data Mining is the process of finding correlation or patterns among dozens of fields in large relational databases. However in some context, it is to extract information or patterns from data in huge databases. Data mining is the procedure for discovering data from different viewpoints and summarizing it into valuable information. This information can be used to improve costs and profits of data information or both.

Association rule mining [1] is the one of most popularly used research in data mining and has much application. It illustrates the relationship among the various data attributes. The extraction of interesting correlations, frequent patterns, associations or casual structures among sets of items in the transaction databases or other data repositories is the main objective of Association rule mining. Association rule mining extracts interesting correlation and relation between large volumes of transactions. Itemset mining is an exploratory data mining technique widely used for discovering valuable correlations among data or information. Infrequent itemsets are produced from very big or huge data sets by applying some rules or association rule mining algorithms like Apriori technique, that take larger computing time to compute all the frequent itemsets. Extraction of frequent itemsets is a core step in many association analysis techniques. The frequent occurrence of item is expressed in terms of the support count.

However, significantly less attention has been paid to mining of infrequent itemsets, but it has acquired significant usage in mining of negative association rules from infrequent itemset, fraud detection, where rare patterns in financial or tax data may suggest unusual activity associated with fraudulent behavior, market basket analysis and in bioinformatics where rare patterns in microarray data may suggest genetic disorders. Several frequent items set mining including Apriori, FP-Growth algorithm, FP-GROWTH* algorithm, Enhanced FP-Growth algorithm, and Transaction mapping algorithm were proposed. And this paper discuss about literature review on various infrequent itemset mining algorithms.

II. Related Work

TECHNIQUES USED FOR INFREQUENT ITEM MINING

a) *Rare Association Rules generation*

In [4] Laszlo et.al presented generation of rare association rules for mining of infrequent itemsets. In this work presented a method to taking out rare association rules that stay hidden for traditional frequent itemset mining algorithms. When compared with other method the presented method finds strong but rare associations that are local regularities in the data are found. These rules are said to be “mRI rules” .Apriori computes the support of minimal rare itemsets (mRIs), i.e. rare itemsets such that all proper subsets are frequent. Instead of pruning the mRIs, they are retained. In addition, it is shown that the mRIs form a generator set of rare itemsets, i.e. all rare itemsets can be restored from the set of mRIs which have two merits. Firstly, they are highly informative in the case that they have an ancestor which is a producer itemset while adding up the resultant to give ways for a closed itemset. Secondly, the amount of these rules is minimal, that is the mRG rules comprise a dense illustration of all largely confident associations that can be taken from the least rare itemsets.

b) *Positive and Negative Association rule*

In [2] X. wu Efficient mining of both positive and negative association rules .They focused on identifying the associations among frequent itemsets. They designed a new method for efficiently mining both positive and negative association rules in databases. This approach is novel and different from existing research efforts on association analysis. Some infrequent itemsets are of very interest in this method but not in existing research efforts. They had also designed constraints for reducing the search space, and had used the increasing degree of the conditional probability relative to the prior probability to estimate the confidence of positive and negative association rules.

c) *Minimal infrequent itemset mining*

In [3] David et.al presented a new algorithm of MINIT, for finding minimal τ -infrequent or minimal τ -concurrent item sets. Firstly, a ranking of items is organized by estimating the need of each of the items and then generating a record of items in rising order of support. Minimal τ –infrequent itemsets are determined by using each item in rank order, iteratively calling MINIT on the maintained set of the dataset with regard to items using only those items with superior rank than current items , after that checking each candidate of minimal infrequent items (MII) against the original dataset is performed. A system that can be utilized to judge only superior-ranking items in the iteration is to preserve a “liveness” vector representing which items stay feasible at each level of the iteration.

d) *Pattern-Growth Paradigm and Residual Trees*

In[5] Ashish Gupta e.al presented pattern-growth paradigm to discover minimally infrequent itemsets. They recommend a new algorithm based on the pattern-growth paradigm to find minimally infrequent itemsets. It has no subset which is also infrequent. This work uses novel algorithm of IFP min for mining minimally infrequent itemsets. Then the residual tree concept has been incorporated by using a variant of the FP-Tree structure which is known as inverse FP-tree. In order to mine the minimally infrequent itemsets, optimization of Apriori algorithm is performed. Finally the presented tree are used for mining of frequent itemset as well.

Sr. No.	Author & Year	Techniques	Merits	De-Merits
1	David J. Haglin and Anna M. Manning-2007	Minimal infrequent itemset mining	Better Performance	Improved running time is not observed
2	Ashish Gupta, Akshay Mittal, Arnab Bhattacharya-2011	Pattern- Growth Paradigm and Residual Trees	Improved performance is obtained with less computational time	Better scalability is not achieved for mining maximally frequent itemsets

III. ALGORITHM

a) Apriori Algorithm

Apriori [6] means a "Prior Knowledge of data" was the first proposed algorithm in Association rule mining, to identify the frequent itemsets in the large transactional database. Apriori works in two phases. During the first phase it generates all possible Itemsets combinations. These combinations will act as possible candidates. The candidates will be used in subsequent phases. In Apriori algorithm, first the minimum support is applied to find all frequent itemsets in a database and second, these frequent itemsets and the minimum confidence constraint are used to form rules.

Apriori Algorithm:

```
procedure Apriori (T, min Support)
{
    //T is the database and min Support is the minimum support
    L1= {frequent items};
    for (k= 2; Lk-1 !=∅; k++)
    {
        Ck= candidates generated from Lk-1
        for each transaction t in database
        do
        {
            #increment the count of all candidates in Ck that are contained in it
            Lk = candidates in Ck with min Support
        }//end for each
    }//end for
    return ∪k  Lk ;
}
```

The main drawback of Apriori is the generation of large number of candidate sets. The efficiency of apriori can be improved by Monotonicity property, hash based technique, Partioning methods and so on.

b) FP-growth Algorithm

The drawback of Apriori can be improved by Frequent pattern Growth algorithm[3].This algorithm is implemented without generating the candidate sets. This algorithm proposes a tree structure called FP tree structure, going to collect information from the database and creates an optimized data structure as Conditional pattern. Initially it Scans the transaction database DB once and Collects the set of frequent items F and their supports and then Sort the frequent itemsets in descending order as L, based on the support count. This algorithm reduces the number of candidate set generation, number of transactions, number of comparisons.

Algorithm FP-growth:

Input:

-A transactional database *DB* and a minimum support

threshold ξ.

Output:

- frequent pattern tree, FP-tree

/*phase1: */

(1) Scan the transactional database.

(2) Collect the set of frequent items *F* and their supports.

Sort *F* in support descending order as *L*. /* the *list* of frequent items*/

(3) [8] Create the root of an FP-tree, *T*, and label it as “root” /* for each transaction do */

(4) Select and sort the frequent items in *Trans* according to the order of *L*.

(5) perform the insert-tree function /* call insert-tree function recursively */ /*phase2: */

Input:

An FP-tree constructed in the above algorithm,

D – transaction database;

s – minimum support threshold.

Output:

-The complete set of frequent patterns.

(1) call the FP-Growth function

(2) check if the tree has a single path,

(3) then for each combination (denoted as *B*) of the nodes

in the path *P* do

(4) generate pattern B ∪ A with support=minimum

support of nodes in B

(5) else

(6) construct B’s conditional pattern base and B’s

conditional FP-tree.

c) The Infrequent Weighted Itemset Miner Algorithm

IWI Miner is a FP-growth-like mining algorithm [7] that performs projection-based itemset mining. Hence, we performs the main FP-growth mining steps:

(a) FP-tree creation and

(b) recursive itemset mining from the FP tree index. Unlike FP-Growth, IWI Miner discovers infrequent weighted itemsets instead of frequent (unweighted) ones. To accomplish this task, the following main modifications with respect to FP-growth have been introduced:

1. A novel pruning strategy for pruning part of the search space early and
2. a slightly modified FP-tree structure, which allows storing the IWI-support value associated with each node.

Algorithm (IWI Miner(T, E))

Input:

-T, a weighted transactional dataset

Input:

-E, a maximum IWI-support threshold

Output:

-F, the set of IWI satisfying E

(1) F=0 /*Initialization*//*scan T and count the IWI support of each item */

(2) count the infrequent weighted item sets with the support value.

(3) create header table which is a data structure which holds information about total weight values.

(4) for each transaction, create equivalent transaction.

(5) create an FP-Tree, for each transaction.

(6) Iterate the process until all transactions are traced.

(7) create conditional pattern base

(8) calculate weight value.

(9) obtain the infrequent item sets.

To reduce the complexity of the mining process, IWI Miner adopts an FP-tree node pruning strategy to early discard items (nodes) that could never belong to any itemset satisfying the IWI-support threshold. Hence, an item(i.e., its associated nodes) is pruned if it appears only in tree paths from the root to a leaf node characterized by IWI-support value greater than E.

d) Minimal Infrequent Weighted ItemSet Miner

The MIWI Mining procedure is similar to IWI Mining[1][7]. However, since MIWI Miner focuses on generating only minimal infrequent patterns, the recursive extraction in the MIWI Mining procedure is stopped as soon as an infrequent itemset occurs. It finds both the infrequent itemsets and minimal infrequent itemset mining. The advantage of MIWI algorithm is reduction in generating the candidate sets, reducing the computational Time, improved the efficiency of algorithm performance compared to FP-Growth algorithm.

IV. Conclusion

In this paper the problem of generating the infrequent itemset by using the weights for distinguishing among applicable items and not within each transaction. We proposed the two FP growths like algorithms which achieve IWI and MIWI mining competently. The value of the exposed patterns has been authenticated on the data coming from a real life context with the help of domain experts.In future, this topic will incorporate the proposed system in an advanced decision making system which sustain domain expert targeted actions based on the characteristics of the discovered IWIs.

References

1. Agrawal R, Imielinski T, &Swami , A. “Mining association rules between sets of items in large databases”.In proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, pages 207-216, Washington, DC, 1993.
2. X. Wu, C. Zhang, and S. Zhang, “Efficient Mining of Both Positive and Negative Association Rules,” ACM Trans. Information Systems,vol. 22, no. 3, pp. 381-405, 2004.
3. D.J. Haglin and A.M. Manning, “On Minimal Infrequent Itemset Mining,” Proc. Int’l Conf. Data Mining (DMIN ’07), pp. 141-147,2007.
4. Laszlo Szathmary, PetkoValtchev, and Amedeo Napoli,” Finding Minimal Rare Itemsets and Rare Association Rules” Proceedings of the 4th International Conference on Knowledge Science, Engineering and Management (KSEM 2010).
5. A. Gupta, A. Mittal, and A. Bhattacharya, “Minimally Infrequent Itemset Mining Using Pattern-Growth Paradigm and Residual Trees,”Proc. Int’l Conf. Management of Data (COMAD), pp. 57-68,2011.
6. R. Agrawal and R. Srikant, “Fast Algorithms for Mining Association Rules,” Proc. 20th Int’l Conf. Very Large Data Bases (VLDB ’94), pp. 487-499, 1994.
7. Han, J. , Pei, J. , & Yin, Y. “Mining frequent patterns without candidate generation”. In Proc. ACM-SIGMOD Int. Conf. Management of Data (SIGMOD ’96), Page 205-216, 2000.
8. Luca Cagliero and Paolo Garza “Infrequent Weighted Itemset Mining using Frequent Pattern Growth”, IEEE Transactions on Knowledge and Data Engineering, pp. 1- 14, 2013.

AUTHOR(S) PROFILE

Sonia Jadhav, received the B.Tech degree in Computer Science and Engineering from Kurukshetra Institute of Technology & Management, Kurukshetra , Haryana, India in the year of 2011. She is currently doing her M.E. degree in Computer Science and Engineering in Bhivarabai Sawant Institute of Technology & Research, Wagholi Pune, India. Her area of interests includes Data Mining, Networking.

G. M. Bhandari is currently working as head of the Computer Science and Engineering department in Bhivarabai Sawant Institute of Technology & Research, Wagholi, and Pune. India. She obtained her M.tech Degree from College of Engineering, Pune. She is having an experience of 14 years and published many research papers in various international journals and conferences. Her areas of interests includes Data Mining, Networking, Image Processing, Cloud Computing, Algorithm Analysis and Soft Computing