A Survey on Infrequent Weighted Mining

Rizwana Begum S.M, Pratibha.S. Yalagi

Abstract— **Pattern mining has become an essential function of data mining. Mining infrequent and frequent pattern from a dataset is the most essential part of data mining. Most of the research study has been done on frequent itemset mining but much less study has been given to mining Infrequent itemsets even though it has obtained its usage in many different real life applications such as Market basket analysis, Risk analysis, Fraud Detection, Biological data analysis. In a Mark Basket analysis infrequent itemset mining can be used to increase profits of items purchased by customers. The infrequent Itemset mining is the process of mining itemsets whose support is less than or equal to maximum support threshold. Discovering infrequent patterns efficiently from large datasets and an important pattern from the discovered patterns is the challenging problem in the field of infrequent Itemset mining. This paper provides a broad summary of the different papers that gives insight into various algorithms designed for mining infrequent itemsets which can help in future research work in the field of itemset mining. This paper also gives comparative study on Various infrequent itemset mining algorithms.**

Index Terms— **Data Mining, Frequent Itemset Mining, Hadoop, Infrequent Itemset Mining, MapReduce, Pattern Mining.**

I. INTRODUCTION

Data Mining is the way of discovering interesting and important knowledge from large amounts of data. This knowledge can be utilized to increase profits of the items or to reduce costs of items or both. Itemset mining is a data mining technique commonly used for determining the valuable correlation among data. The first attention was focused on finding frequent itemsets. An itemset is called frequent if its support threshold is greater than or equal to a user specified minimum support (ms) threshold otherwise the itemset is called infrequent itemset. Much of the time Infrequent itemsets are considered to be unimportant and are destroyed using the support count, but Some infrequent itemsets may also indicate the occurrence of important occasional events or unpredictable situations in the database. Consider the scenario in a Medical data analysis if people are affected by rare disease then it can be intercepted by finding solution of such unusual diseases. Here, study shows that don't occure such unusual disease to repeatedly. To discover such unusual condition, the required support of an itemset must be decided, so that, if an itemset turns out to have a significantly lower support than the required support then it is stated as an unimportant infrequent itemset otherwise it is important infrequent itemset. In the safety area, the usual action is very frequent, whereas unusual action is less frequent. Suppose a database where the action of human beings in receptive locations such as a Hospital or the airport is recorded, if we record those actions, it is possible to discover that usual actions that can be represented by frequent patterns and unusual actions can be represented by infrequent patterns [2]. Infrequent Itemset Mining has obtained its usage in many different real life applications such as Market basket analysis [3], Fraud Detection, Risk Analysis, Biological data analysis [4].

The study below illustrates details of various techniques & algorithms used for finding infrequent itemsets from transactional datasets:

A. Minimal infrequent Itemset mining

In [5] David J. Haglin and Anna M. Manning designed a novel algorithm of **MIN**imal **I**nfrequent i**T**emsets(MINIT), for Mining minimal τ infrequent or minimal τ concurrent itemsets. This algorithm was first designed especially for discovering minimal infrequent itemsets. At first step, support of each item in the transaction dataset is estimated according to this ranking of the items is done & then producing a list of items in scaling order according to their support. Minimal τ infrequent itemsets are uncovered according to rank order of an item, MINIT is called repeatedly on the maintained support set of the dataset it only takes into account only those items that have a high rank than a rank of current items, consequently every candidate of minimal infrequent items is checked across the original dataset. A technique that can be applied to examine only high ranking items in the repetition is to support a "liveness" vector for representing which items stay feasible at each level of the repetition.

B. Frequent Pattern Growth

In [6] Luca Cagliero and Paolo Garza focused on considering the weights assigned to each of the items in the weighted transactional dataset in order to find infrequent itemsets. The IWI-support count is stated as an aggregation of weights of the items belongs to Items in the transaction dataset. The weights of items are obtained from the weights associated with items in each transaction in a dataset by employing a certain cost function. This study concentrates on IWI- support-min count and IWI support-max count only. Consider I is an itemset, T is a weighted transactional data set, t_q is the set of items in tq€T.So, the task of estimating weight of Itemset in a given weighted transaction data is classified into two steps. Firstly, the significance of an itemset I associated with a weighted transaction tq€T is defined as an aggregation of its item weights in tq. Secondly, the weight of itemset I with regard to the whole data set T is computed by merging the itemset significance weights associated with each transaction in the weighted transaction dataset. Author proposes two IWI mining algorithm that is IWI miner and Minimal IWI miner for mining infrequent itemset efficiently.

C. Apriori_Multilevel Minimum Support Model Algorithm

In [7] Xiangjun Dong1, 2, Zhiyun Zheng3, Zhendong Niu1, Qiuting Jia2 proposed a Multilevel Minimum Support Model for discovering infrequent itemsets.

Infrequent itemset can be discovered using a single minimum support, but would result in such a difficult problem: if the support count is very high, less number of frequent itemsets can be generated and if the support count is very low, more number of frequent itemsets can be generated. In order to solve the problem, study uses multiple level minimum supports (MLMS) model, i.e., Study uses different minimum supports to itemsets with different lengths. An Apriori _MLMS algorithm is used to find frequent itemsets as well as infrequent itemsets simultaneously based on MLMS model.

D. Pattern Growth Paradigm and Residual Trees

Ashish Gupta e.al in [8] proposed a novel algorithm IFP min which is based on the pattern growth paradigm to uncover minimally infrequent itemsets. An infrequent Itemset is called minimally infrequent Itemset if it has all its subset which is frequent. The Study designed a novel algorithm IFP min based on pattern growth paradigm for mining minimally infrequent itemsets. This work makes use of new idea called residual trees using a variant of the Frequent pattern tree (FP tree) structure termed as an inverse Frequent Pattern tree (IFP tree). Optimization of apriori is carried out to find out minimally infrequent itemsets. Finally, it uses this algorithm with MLMS model to discover frequent Itemset. Computationally complexity of Algorithms based on Pattern growth is low on dense datasets.

E. Mine Weighted Infrequent Itemset based on MapReduce

In [9] T Ramakrishnudu presented an algorithm based on MapReduce model to find infrequent weighted itemsets. Initially algorithm Scan the input dataset and Divide it into a number of blocks and attach one block to each node. The Mapper functions at each node Scans each transaction of the input data subset and generate all possible subsets of the transaction called local candidate itemsets & then Calculate IWI-Support for all local candidate itemsets and generate intermediate output that is stored in HDFS (Hadoop Distributed File System). HDFS is a part of the Apache Hadoop main project [10]. This intermediate output of mappers will be given as input to Reducer. The Reducer functions accept itemset, IWI-Support as input & Calculate IWI-support of the global candidate Itemset & check if IWI-Support(I) < User defined minimum threshold value, then assign itemset I to output list Else discard Itemset I. Finally, Reducer function gives Output the itemset I & IWI-Support. The mapper function accepts one input called input split, and the reducer function accepts two different inputs called intermediate output and weighted minimum support given by the user [11] [12].

F. Positive and Negative Association rule

In [13] X. Wu, C. Zhang and S. Zhang this work have been focused on a new technique for mining both positive and negative association rules efficiently from databases. This research emphasis on discovering the associations between frequent itemsets. This work devises a novel technique different from previous research efforts on association mining for efficiently mining both positive and negative association rules within databases. It is necessary to mine infrequent itemset because Some infrequent itemsets are of importance which can give more benefits than some frequent itemsets. They had also deliberated a method for reducing the search space, and had used the rising level of the conditional probability relative to the prior probability to determine the confidence of positive as well as negative association rules.

G. Rare Association Rules generation

In [4] Laszlo et.al presented formation of rare association rules for discovering of infrequent itemsets from datasets. They introduced a technique for pruning rare association rules that remain hidden for traditional frequent itemset mining algorithms. When this method is compared with another method the presented method finds better than other, but rare associations that are local regularities in the data are obtained. These rules can be called as “mRI rules”. Apriori computes the support of minimal rare itemsets (mRIs). Minimal rare itemsets are the itemsets who has no proper subsets that are infrequent. Instead of taking out the mRIs, they are retained. Furthermore, it is presented that the mRIs form a generator set of infrequent itemsets, i.e. all infrequent itemsets can be restored from the set of mRIs, which have two benefits. At the beginning, they are highly descriptive in the case that they have a predecessor which is a producer itemset at the time by combining the resultant to give ways for a closed itemset. Next, the number of these rules is minimal, that is the mRi rules include a dense illustration of all, mostly from the least rare itemsets confident associations can be taken.

II. COMPARATIVE STUDY OF VARIOUS INFREQUENT ITEMSET MINING ALGORITHMS

In this section, Comparison of infrequent itemset mining algorithms is done. This comparison is based on different algorithms, techniques used in the algorithms, types of Itemset generated by algorithm, performance of algorithms based on number of transactions. In Figure 1 shows that the T40I10D100K is sparse dataset. Since Apriori min is a test based and candidate-generation algorithm, it gets halted when all the candidate itemsets are infrequent. Due to this, it avoids the complete traversal for all possible lengths of the database. However, both MINIT & IFP min is based on the recursive elimination procedure, in order to report the MIIs they have to complete their full run. This results in higher computational times of these methods. On analysis, and comparing the MIIs generated by IFP min, Apriori min and MINIT algorithms, the itemsets having zero support count in the transaction database are not reported by the MINIT algorithm, thereby leading to its insufficiency. Based on the experimental results, it is noticed that for large, dense datasets, it’s superior to use the IFP min algorithm. For small, dense, datasets, IFP min should be used at high support thresholds and MINIT should be used at low support thresholds. For sparse datasets, Apriori min should be used for reporting MIIs. In IFP min algorithm, each candidate MII itemset is checked for set membership in a residual database, whereas in MINIT the candidates are verified by measuring the support from the whole database. Due to reduced validation space, IFP min performs superior than MINIT.

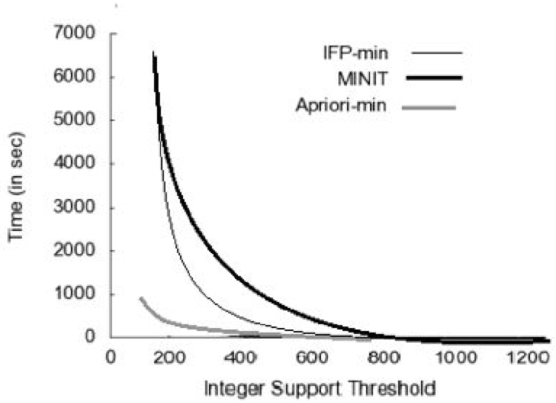


Figure 1. T40I10D100K Dataset

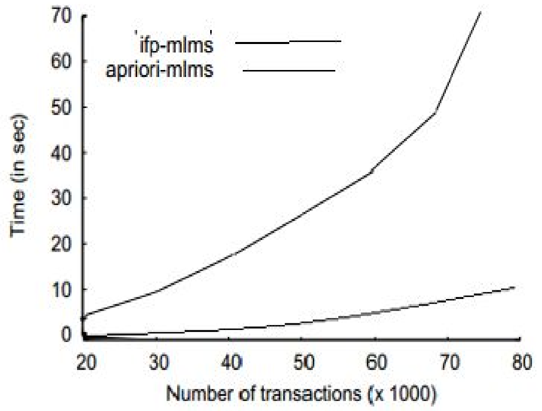


Figure 2. The IFP_MLMS and Apriori_MLMS algorithms on the Anonymous Microsoft Web Dataset

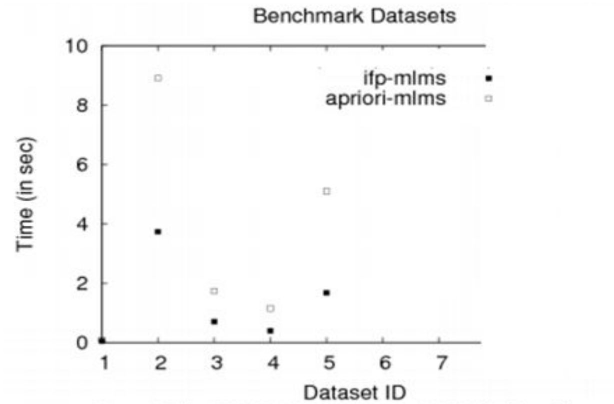
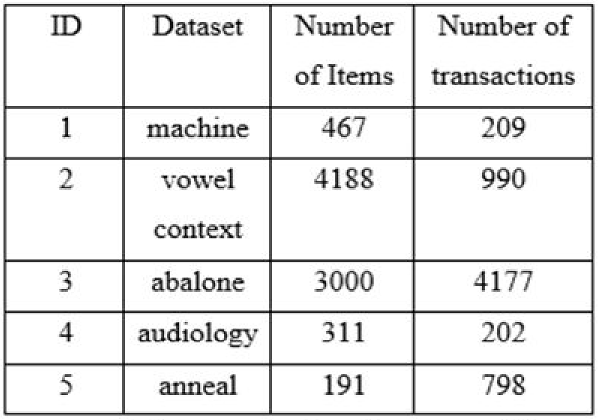


Figure 3.The IFP_MLMS and Apriori_MLMS Algorithms on data sets given in Table 2.

As shown in Figure 2, the number of transactions is varied in the Anonymous Microsoft Web dataset. It is noticed that the execution time for both Apriori_MLMS and IFP_MLMS algorithms increases with the number of transactions. However, the execution time for the Apriori MLMS algorithm is much higher than IFP_MLMS and at 80,000 transactions, Apriori MLMS has crashed due to insufficiency of memory. Since for large datasets Apriori MLMS has crashed to give results, for comparing IFP MLMS with Apriori_MLMS smaller datasets can be obtained from http://archive.ics.uci.edu/ml/datasets/. The properties of these datasets are shown in Table 2. For each such dataset, time required for IFP MLMS and Apriori MLMS are marked in Figure 3. IFP MLMS algorithm performs better than Apriori MLMS. Experiment result shows that, the running time of both IFP MLMS and Apriori MLMS algorithms to be independent of support thresholds for the MLMS model.



ID	Dataset	Number of Items	Number of transactions
1	machine	467	209
2	vowel context	4188	990
3	abalone	3000	4177
4	audiology	311	202
5	anneal	191	798

Table 2. Details of smaller Datasets[8]

As shown in below TABLE1.Comparison of infrequent itemset mining algorithms. This comparison is based on techniques used in the algorithms, types of itemset generated by algorithm, performance & drawback of algorithms. In MINimal Infrequent iTemsets(MINIT) algorithm Minimal infrequent itemset mining technique[5]is used & it is observed that it takes less time to search but running time of this algorithm is high. Frequent Pattern Growth [6] technique is used in Infrequent Weighted itemset miner, MIWI Miner to mine infrequent weighted itemsets. MIWI Miner is faster when dealing with dense datasets, but execution time of this algorithm increases with dataset size. Multilevel Minimum Support Model [7] is used in the apriori algorithm for finding frequent and infrequent Itemset in an appropriate degree, but the user has to give multiple min.support for itemsets of different lengths. Pattern Growth Paradigm and Residual Trees [8] technique is used in IFP min in order to mine Minimal infrequent itemset.It achieves better performance but not scalable for mining maximal frequent itemsets. Mine Weighted Infrequent Itemset based on MapReduce [9] designed algorithm based on the MapReduce model for mining weighted infrequent itemsets.Main advantage of it is that it can provide lots of computation power and storage space.

III. Conclusion

The present work focused on comparative study of Various infrequent itemset mining algorithms. Infrequent itemset mining can be used in various applications like Market basket analysis, Fraud Detection, Biological data analysis, Risk analysis. In a Mark Basket analysis infrequent itemset mining can be used to increase profits of items purchased by customers. Comparative analysis indicates that IFP min is faster than MINIT when min. support is high. IFP min and MIWI work superior than MINIT for every support. When min. support threshold is low MIWI works better than IFP and MINIT. Execution time of MIWI grows when IWI support min is medium. Execution time of IWI miner rises when IWI support min is high. Comparative study of IFP_MLMS with Apriori_MLMS shows that execution time of both algorithm increases when the number of transactions increases. The growth rate of execution time of Apriori_MLMS is higher than IFP_MLMS. Apriori MLMS fails to give results for large datasets so IFP_MLMS performs better than Apriori_MLMS.

IV. References

[1] Shipra Khare, Prof. Vivek Jain,” A Review on Infrequent Weighted Itemset Mining using Frequent Pattern Growth”, Shipra Khare et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (2), 2014, 1642-1647

[2] Mehdi Adda, Lei Wu,"Pattern Detection with Rare Itemset Mining”, International Journal On Soft Computing, Artificial Intelligence and Applications, vol.1, No.1, August 2012.

[3] R. Agrawal, T. Imielinski, and Swami, “Mining Association Rules between Sets of Items in Large Databases,” Proc. ACM SIGMOD Int’l Conf. Management of Data (SIGMOD ’93), pp. 207-216, 1993.

[4] G. Cong, A.K.H. Tung, X. Xu, F. Pan, and J. Yang, “Farmer: Finding Interesting Rule Groups in Microarray Datasets,” Proc. ACM SIGMOD Int’l Conf. Management of Data (SIGMOD ’04), 2004.

[5] David J. Haglin and Anna M. Manning,” On Minimal Infrequent Itemset Mining ”.

[6] Luca Cagliero and Paolo Garza,” Infrequent Weighted Itemset Mining Using Frequent Pattern Growth”, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 26, NO. 4, APRIL 2014.

[7] Dong, Z Zheng, Z Niu and Q Jiam” Mining infrequent itemset based on multiple level minimum supports”, 2nd Int. Conf. on Innovative Computing, Information Control, 2007.

[8] 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

[9] T Ramakrishnudu, R B V Subramanyam,” Mining Interesting Infrequent Itemsets from Very Large Data based on MapReduce Framework”, I.J. Intelligent Systems and Applications, 2015, 07, 44-49

[10] Mohammadhossein B and Madhi Niamanesh, “ScaniBino: An effective MapReduce-based association rule mining method”, in proceedings of the sixteenth International Conference on Electronic commerce, USA, August 2014, pp.1 -8.

[11] Jeffery Dean and Sanjay Ghemawat “MapReduce: simplified data processing on large clusters”, 6th Symposium on Operating Systems Design and Implementation, October 2004, pp.107-113.

[12] Jeffery Dean and Sanjay Ghemawat “MapReduce: simplified data processing on large clusters”, Communications of the ACM, Vol. 51, No.1, 2008, pp. 107-113.

[13] X. Wu, C. Zhang and S. Zhang,” Efficient mining of both positive and negative association rules”, ACM Trans. On Information Systems, vol.22 (3), 2004, pp 381 –405.

[14] X. Dong, S. Wang, and H. Song, “2-level Support based Approach for Mining Positive & Negative Association Rules”, Computer Engineering, 31(10) 2005, pp. 16-18.

[15] Chris Cornelis, Peng Yan, Xing Zhang, Guoqing Chen: “mining positive and negative association rules from large databases”, in IEEE conference on Cybernetics and Intelligent Systems, Bangkok, June 2006, pp.1-6.

TABLE1.Comparison of infrequent Itemset mining algorithms

Sr. No.	Techniques	Algorithm	Types of Itemset	Performance	Drawback
1	**Minimal infrequent itemset mining [5]**	**MIN**imal **I**nfrequent i**T**emsets(MINIT) algorithm	minimal τ-infrequent or minimal τ-occur rent itemsets	Search complexity is inferior and achieves better performance	Computational complexity and running time is high.
2	**Frequent Pattern Growth [6]**	Infrequent Weighted Itemset miner, MIWI Miner	infrequent weighted itemsets	MIWI Miner is faster when dealing with dense datasets.	With the data set size execution time scales linearly roughly
3	**Multilevel Minimum Support Model [7]**	**Apriori**_Multilevel Minimum Support Model	Frequent & infrequent itemsets	Discover both frequent and infrequent itemsets in an appropriate degree	User has to give different Min. support for items of different lengths
4	**Pattern-Growth Paradigm and Residual Trees [8]**	Infrequent Pattern Miner(IFP) miner	Minimal infrequent itemset	Computation complexity is low and achieves better performance.	Better scalability is not accomplished for mining maximal frequent itemsets.
5	**Mine Weighted Infrequent Itemset based on MapReduce [9]**	An Algorithm based on MapReduce	weighted infrequent Itemsets	Can provide lots of computation power and storage space	Weighting function is not so good

Rizwana Begum S.M is pursuing her Master of Engineering in Computer Science & Engineering in Walchand Institute of Technology, Solapur, India. Her area of interest includes Data Mining. (e-mail: sayyadriz@gmail.com).

Pratibha S. Yalagi is currently working as Assistant Professor of Information Technology department in Walchand Institute of Technology, Solapur, India. (e-mail: pratibhayalagi@gmail.com).