Efficient Utility Based Infrequent Weighted Item-Set Mining

R.Priyanka[1]
[1]Department of Computer Science and Engineering,
Kumaraguru College of Technology,
Coimbatore, Tamil Nadu, India.
Priyankarajan123@gmail.com

S.P. Siddique Ibrahim[2]
[2]Department of Computer Science and Engineering,
Kumaraguru College of Technology,
Coimbatore, Tamil Nadu, India.
Siddiqueibrahim.sp.cse@kct.ac.in

Abstract— Association Rule Mining (ARM) is one of the most popular data mining techniques. Most of the past work is based on frequent item-set. In current years, the concentration of researchers has been focused on infrequent item-set mining. The infrequent item-set mining problem is discovering item-sets whose frequency of the data is less than or equal to maximum threshold. This paper addresses the mining of infrequent item-set. To address this issue, the IWI-support measure is defined as a weighted frequency of occurrence of an item set in the analyzed data. This Infrequent weighted item set mining discovers frequent item sets from transactional databases using only items occurrence frequency and not considering items utility. But in many real world situations, utility of item sets based upon user's perspective such as cost, profit or revenue is of significant importance. In our proposed system we are proposing the High Utility based Infrequent Weighted Item set mining (HUIWIM). High Utility based Infrequent Weighted Item set mining (HUIWIM) to find high utility Infrequent weighted item set based on minimum threshold values and user preferences. The proposed system is used for efficiently and effectively mine high utility infrequent weighted item set from databases and it can improve the performance of the system compared to the existing system.

Index Terms— Data Mining, Frequent item -set, infrequent item set, FP- growth Algorithm, HUIWIM

1 INTRODUCTION

Data Mining is defined as "Extraction interesting patterns or knowledge from the huge amount of data". Association rule mining (ARM) is one of the most widely used techniques in data mining and knowledge discovery and has tremendous applications like business, science and other domains. A group of items in a transaction database is called Item Sets. Item Set mining is an exploratory data mining. technique widely used for discovering correlation among data. Item-set mining focused on discovering frequent item-set

Item-set is frequent if its support satisfies given minimum support threshold. Frequent item-set find application in many real life contexts. For example digital camera, then a memory card and then buying PC if it occurs frequently in shopping history, then it is called frequent pattern. Market basket analysis [1] is one of the frequent item-set mining applications. A weight is associated with each item and characterizes its importance within each transaction.

Most of the past work is based on frequent Item Set. In recent years, the concentration of researcher is focused on infrequent Item Set mining problem, i.e., discovering Item Sets whose frequency of occurrence of analyzed data is less than or equal to threshold value. In [7], [8] used an algorithm for finding minimal infrequent Item Set, they proposed that infrequent Item Sets does not contain any infrequent subset. (i) Fraud detection (ii) Statistical disclosure risk assessment for census data are real life application of infrequent Item Set discovery. In [4],[5],[6] are used to derive frequent weighted Item Set mining process.

Consider an example, the data set reported in Table 1. It includes five transactions, each one composed of five distinct items weighted by the corresponding degree of interest (e.g., item System1 has weight 0 in Tied 1,and 100 in Tied 4). For instance, Tied 1 means that, at a point of time, item jelly is fully utilized (weight 100), butter and milk have used intermediary (weights 57 and 71 respectively), and item bread is temporarily idle (weight 0).

This paper addresses the discovery of utility based infrequent and weighted Item Sets. Here considers both the individual profit of each item in a database and the bought quantity of each one in a transaction simultaneously. . Here utility threshold value is given based on the user's interest. Infrequent Weighted Item-set having utility value greater than the minimum utility threshold is generated from different time periods.

TABLE 1
EXAMPLE OF WEIGHTED TRANSACTIONAL DATA SET

Tied	System utilization			
1	Sys1, 0	Sys2,100	Sys3,57	Sys4,71
2	Sys1, 0	Sys2,43	Sys3,29	Sys4,71
3	Sys1, 43	Sys2,0	Sys3,43	Sys4,43
4	Sys1, 100	Sys2,0	Sys3,0	Sys4,100
5	Sys1, 86	Sys2,71	Sys3,0	Sys4,71

2 RELATED WORK

Author in [1] has proposed a new algorithm called MINIT (Minimal Infrequent Item-set), which is a used for mining minimal infrequent item-set (MIIs) . This is the first algorithm for finding rare item-set. Item-set that satisfy a maximum support threshold and does not contain any infrequent subset, from transactional data set. It is based on SUDA2 algorithm. Author in [2] has proposed a new measure w-support, which is used to find weight of item-set, and weight of transaction does not require preassigned weight. These weights are completely based on internal structure of the database. HITS model and algorithm are used to derive the weights of transactions from a database with only binary attributes. Based on these weights, a new measure w-support is defined to give significance of item-set, and it differs from the traditional support in taking the quality of transactions into consideration. An apriori-like algorithm is proposed to extract association rules whose w-support and w-confidence are above some give thresholds.

Probabilistic frequent item-set mining in uncertain transactional database. Based on world semantics in [3], [4] author introduces new probabilistic formulations of frequent item-sets. To address this issue, probabilistic models have been constructed and integrated in Apriori-based [3] or projection-based [5] algorithm. In a probabilistic model, an item-set is said to be frequent if the probability that item-set happens in at least min support is higher than that of given threshold. In addition to probabilistic model, framework is presented which has a capacity to solve the Probabilistic Frequent Item-set Mining (PFIM) problem powerfully.

Based on interest/intensity of the item within the transaction Wei Wang in [6] has proposed by allowing weight to be associated with each item within the transaction. In turn, to associate a weight parameter with each item in a resulting association rules. Then it is called as weighted association rule (WAR). This method produces a higher quality results than previous known method on quantitative association rules. In weighted settings author in [7] deal with the problem of finding significant binary relationship in transactional dataset. In weighted association rule mining problem each item is allowed to have a weight. The aim is to focus on mining significant relationship relating items with significant weights rather than insignificant relationship. A new algorithm called WARM (Weighted Association Rule Mining) is developed. WARM algorithm is both scalable and efficient in discovering significant relationship in weighted settings

Author in [8] present SUDA2 a recursive algorithm for finding Minimal Sample Uniques (MSUs). SUDA2 uses a new method for demonstrating the search space for MSUs and observe about the properties of MSUs to prune and traverse this space. It has a ability to identify the boundaries of the search space for MSUs with an execution time which is several orders of magnitude faster than that of SUDA. SUDA2 is a good candidate for parallelism as a search can be divided up according to rank; it will produce efficient load balancing.

Ashish Gupta, Akshay Mittal, Arnab Bhattacharya [9] propose a new algorithm based on pattern-growth paradigm for finding minimally infrequent item-set. They introduce a new concept called residual tree. To mine a multiple level minimum support item-sets, different length of the item-set by using residual tree. For mining minimally infrequent item-sets (MIIs) author [7] introduces a new algorithm called IFP min. Here Apriori algorithm is proposed to find MIIs. Extension of the algorithm is designed for finding frequent item-set in the multislevel minimum support (MLMS) model.

Luca Cagliero and Paolo Garza [10] deal with the issues of discovering rare and weighted item-sets, i.e., the infrequent item-set mining problem. Finding rare data correlations is more interesting than mining frequent ones. The IWI-support measure is defined as a weighted frequency of occurrence of an item-set in the analyzed data. Occurrence weights are derived from the weights associated with items in each transaction by applying a given cost function. They mainly focus on IWI- support-min measure and IWI-support-max measure.

Author in [17] introduces a new algorithm High Utility Interesting Pattern Mining with a strong frequency affinity. HUP mining extracts important knowledge from databases, it requires long calculations and multiple database scans. In this framework, new measure called frequency affinity is introduced among the items in a HUP. High utility interesting pattern (HUIP) with a strong frequency affinity is defined using this measure. A new tree structure, UTFA, and a novel algorithm, HUIPM, are proposed for the single-pass mining of all HUIPs from a database. Therefore, it significantly reduces the overall runtime of the existing HUP mining algorithms. This algorithm is scalable and can handle a large number of distinct items and transactions.

3 EXISTING SYSTEM

Infrequent weighted Item Set mining discovers infrequent and weighted Item Sets, from transactional data sets. Here weighted frequency of occurrence of an Item Set in the analyzed data is defined using IWI-support measures. Two different IWI-support measures, i.e.(i) The IWI-support- min measure, which based on a minimum cost function, i.e., the occurrence of an Item Set in a given transaction is weighted by the weight of its least interesting item, (ii) The IWI-support-max measure, which based on a maximum cost function, i.e., the occurrence of an Item Set in a given transaction is weighted by the weight of the most interesting item. Minimum and maximum are the most commonly used cost functions, when dealing with optimization problems.

Here present two new algorithms, namely Infrequent Weighted Item Set Miner (IWI Miner) and Minimal Infrequent Weighted Item Set Miner (MIWI Miner), which performs IWI and MIWI mining driven by IWI-support thresholds. IWI Miner and MIWI Miner are FP-Growth-like mining algorithms. (i) Early FP-tree node pruning driven by the maximum IWI-support constraint, i.e., early discarding of part of the search space thanks to a novel item pruning strategy, and (ii) cost function-independence, i.e. they work in the same way, regardless of which constraint (either IWI-support-min or IWI-support-max) is applied, (iii) early stopping of the recursive FP-tree search in MIWI Miner to avoid extracting non-minimal IWIs.

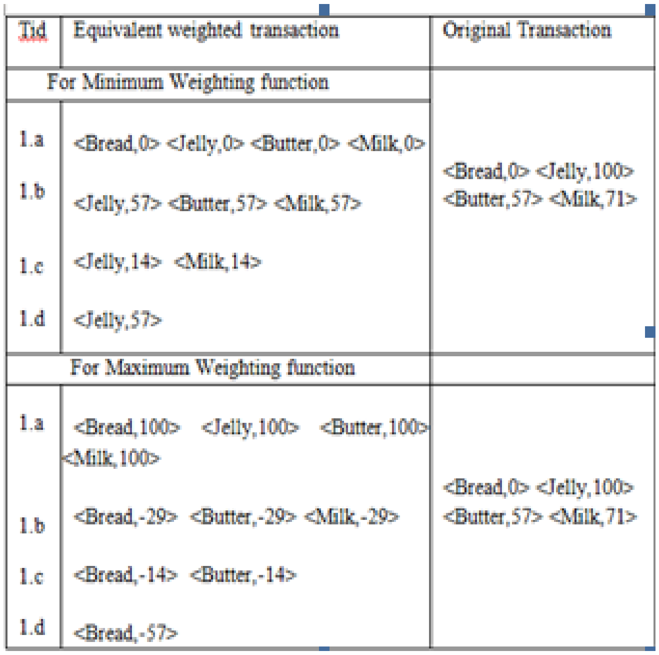
In IWI- miner algorithm, first weighted transactional dataset, and maximum threshold is given as input. Scan each weighted

- *R.Priyanakae is currently pursuing masters of engineering degree program in computer science &engineering in Kumaraguru College of Technology, India, PH-9790590322. E-mail: priyankarajan123@gmail.com*
- *S.P. Siddiqueibrahim is currently working as Assistant Professor in computer science & engineering in Kumaraguru College of Technology, India, PH-01123456789. E-mail: siddiqueibrahim.sp.cse@kct.ac.in* (This *information is optional; change it according to your need.)*

transaction and count the IWI-support of each item. For each weighted transaction the equivalent transaction set is generated reported in Table 2. FP-tree is created using equivalent transactions. IWI-miner relies on projection based approach, items belonging to the header table associated with the input. Generate new Item Set 'I' by combining each item with current prefix, i.e. I= prefix U {i}. If new Item Set 'I' is infrequent, then it is stored in the output set, and then FP-tree projected with respect to I is generated. Then IWI-Mining procedure is recursively applied on projected tree to mine all infrequent items. IWI Miner adopts a different pruning scheme. Select the items that will be a part of any infrequent Item -set, and they are pruned. MIWI Miner algorithm is same as IWI Miner algorithm. MIWI Miner focuses on generating only minimal infrequent patterns. As soon as an infrequent Item Set occurs the recursive extraction in the MIWI Mining procedure is stopped.

TABLE 2

Equivalent Weighted Transaction Associated with the Transaction with Tied 1 in the Example Data Set.



Tid	Equivalent weighted transaction	Original Transaction
For Minimum Weighting function		<Bread,0> <Jelly,100> <Butter,57> <Milk,71>
1.a	<Bread,0> <Jelly,0> <Butter,0> <Milk,0>	
1.b	<Jelly,57> <Butter,57> <Milk,57>	
1.c	<Jelly,14> <Milk,14>	
1.d	<Jelly,57>	
For Maximum Weighting function		
1.a	<Bread,100> <Jelly,100> <Butter,100> <Milk,100>	<Bread,0> <Jelly,100> <Butter,57> <Milk,71>
1.b	<Bread,-29> <Butter,-29> <Milk,-29>	
1.c	<Bread,-14> <Butter,-14>	
1.d	<Bread,-57>	

4 PROPOSED SYSTEM

Existing system discovers infrequent item sets by using weights for differentiate between relevant items and not within each transaction. It analysis only items occurrence frequency and not considering items utility. In our proposed system, we are analyzing the utility parameters of the infrequent weighted item sets. We are considering both the individual profit of each item in a database and the bought quantity of each one in a transaction simultaneously. This system consists of two main phases. In the first phase, Infrequent Weighted Item Set are generated from the transactional database by considering those item sets which have support value less than the maximum support threshold. In the second phase, by inputting the utility threshold value according to the Infrequent Weighted Item Set having utility value greater than the minimum utility threshold is generated from different time periods.

1 Aims at discovering infrequent item sets having high utility

2 This system provides efficient infrequent item sets mining along with the analyzing of utility parameter.

3 Here discover product combinations which are composed of items with low frequency and high profit. Hence frequency is not sufficient to answer a product combination whether it is highly profitable or whether it has a strong impact

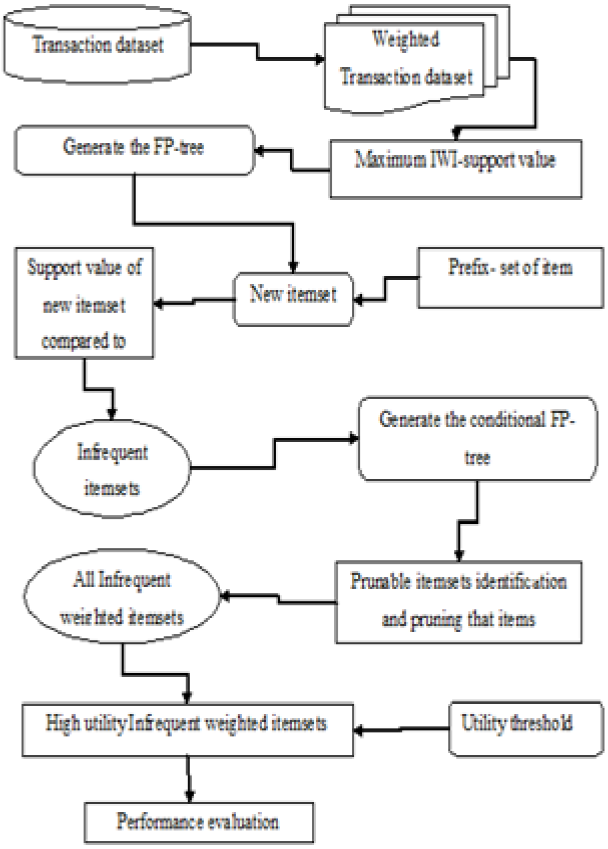

Transaction dataset
Weighted Transaction dataset
Generate the FP-tree
Maximum IWI-support value
Support value of new itemset compared to
New itemset
Prefix- set of item
Infrequent itemsets
Generate the conditional FP-tree
All Infrequent weighted itemsets
Prunable itemsets identification and pruning that items
High utility Infrequent weighted itemsets
Utility threshold
Performance evaluation

Fig1. Proposed System Architecture

5 IMPLEMENTATION

For experimentation, System utilization dataset has been used. The screen shots show the comparisons of IWI-miner and HUWI-Mining algorithm

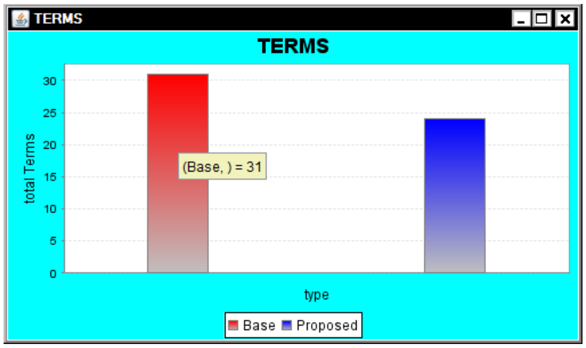


Fig 1. Based on type and total terms to find performance

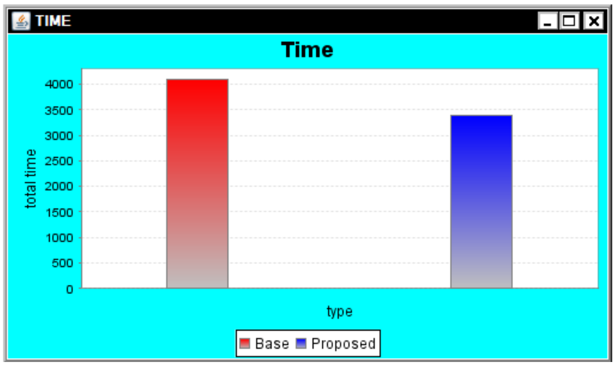


Fig 2. . Based on type and total time to find performance

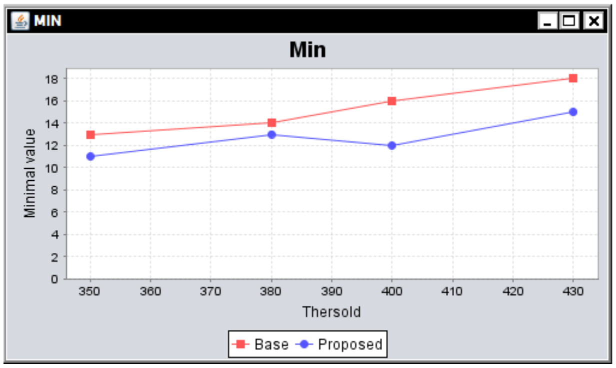


Fig3. Based on threshold and minimal value to find accurate infrequent items

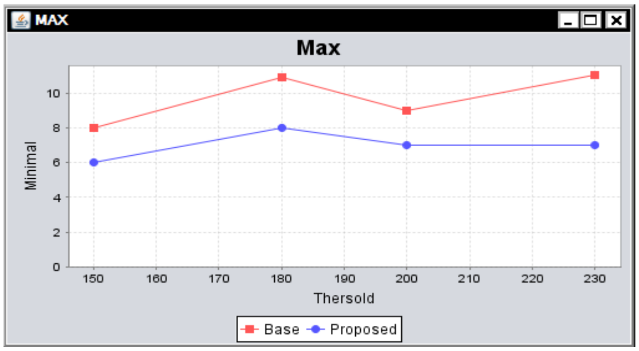


Fig4. . Based on threshold and maximal value to find accurate infrequent items

6 CONCLUSION

In this paper, methods for mining text documents along with the use of side-information are presented. Side-information or meta-information is present in many forms of databases. It can be used to improve the clustering process. In order to design the advance clustering method, iterative partitioning technique and a probability estimation process are combined. It computes the importance of different kinds of side-information. For designing the clustering and classification algorithms a general approach is used**.** COATES Algorithm proves to be very effective. Effective feature selection method is used to extract the features in text documents. Besides using Gini index, correlation based feature selection and symmetric uncertainty are used. Thus it improves the accuracy of text clustering in less time complexity.

REFERENCES

[1] D. J. Haglin and A.M. Manning, “On Minimal Infrequent Item-set Mining,” Proc. Int’l Conf. Data Mining (DMIN ’07), pp. 141-147, 2007

[2] K. Sun and F. Bai, “Mining Weighted Association Rules Without Preassigned Weights,” IEEE Trans. Knowledge and Data Eng.,vol. 20, no. 4, pp. 489-495, Apr. 2008.

[3] C.-K. Chui, B. Kao, and E. Hung, “Mining Frequent Item-sets from Uncertain Data,” Proc. 11th Pacific-Asia Conf. Advances in Knowledge Discovery and Data Mining (PAKDD ’07), pp. 47-58, 2007.

[4] T. Bernecker, H.-P. Kriegel, M. Renz, F. Verhein, and A. Zuefle, “Probabilistic Frequent Itemset Mining in Uncertain Databases, ” Proc. 15th ACM SIGKDD Int’l Conf. Knowledge Discovery and Data Mining (KDD ’09), pp. 119-128, 2009.

[5] C.K.-S. Leung, C.L. Carmichael, and B. Hao, “Efficient Mining of Frequent Patterns from Uncertain Data,” Proc. Seventh IEEE Int’l Conf. Data Mining Workshops (ICDMW ’07), pp. 489-494, 2007.

[6] W. Wang, J. Yang, and P.S. Yu, “Efficient Mining of Weighted Association Rules (WAR),” Proc. Sixth ACM SIGKDD Int’l Conf. Knowledge Discovery and data Mining (KDD ’00), pp. 270-274, 2000.

[7] F. Tao, F. Murtagh, and M. Farid, “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, 2003.

[8] A. Manning and D. Haglin, “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

.

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

[10] Luca Cagliero and Paolo Garza," Infrequent Weighted Item set Mining Using Frequent Pattern Growth" IEEE Transactions On Knowledge And Data Engineering, Volume 26, No.4, April 2014

[11] X. Dong, Z. Zheng, Z. Niu, and Q. Jia, "Mining Infrequent Item-sets Based on Multiple Level Minimum Supports,"Proc. Second Int'l Conf. Innovative Computing, Information and Control (ICICIC '07), pp. 528-531, 2007

[12] J. Han, J. Pei, and Y. Yin, "Mining Frequent Patterns without Candidate Generation," Proc. ACM SIGMOD Int'l Conf. Management of Data, pp. 1-12, 2000.

[13] 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.

[14] Anis Suhailis Abdul Kadir, Azuraliza Abu Bakar, Abdul Razak Hamdan, "Frequent Absence and Presence Item-set for Negative Association Rule Mining," 11th International Conference On Intelligent System Design and Applications, 2011.

[15] T. Bernecker, H.-P. Kriegel, M. Renz, F. Verhein, and A. Zuefle, "Probabilistic Frequent Item-set Mining in Uncertain Databases,"Proc. 15th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD '09), pp. 119-128, 2009.

[16] Mehdi Adda, Lei Wu, Sharon White(2012), Yi Feng " Pattern detection with rare item-set mining" International Journal on Soft Computing, Artificial Intelligence and Applications (IJSCAI), Vol.1, No.1, August 2012.

[17]IdhebaMohamad Ali O. Swesi, Azuraliza Abu Bakar, AnisSuhailis Abdul Kadir," Mining Positive and Negative Association Rules from Interesting Frequent and Infrequent Item-sets", 9th International Conference on Fuzzy Systems and Knowledge Discovery, 2012.

[18] Chowdhury Farhan Ahmed, Syed Khairuzzaman Tanbeer, Byeong-Soo Jeong, Ho-Jin Choi, "A framework for mining interesting high utility patterns with a strongfrequency affinity", Information Science, 2011.

[19] Adinarayanareddy B , O Srinivasa Rao, MHM Krishna Prasad, "An Improved UP-Growth High Utility Itemset Mining", International Journal of Computer Applications (0975 – 8887) Volume 58– No.2, November 2011.