ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 4, April 2015

Efficient Utility Based Infrequent Weighted Itemset Mining From Transactional Weighted Datasets

J.Jaya[1], S.V.Hemalatha[2]

PG Scholar, Department of CSE, Kalaignar Karunanidhi Institute of Technology, Coimbatore, India[1]

Assistant Professor, Department of CSE, Kalaignar Karunanidhi Institute of Technology, Coimbatore, India[2]

Abstract: Traditional data mining techniques have focused mainly on detecting the statistical correlations between the items that are more frequent in the transaction databases. Also termed as frequent itemset mining, these techniques were based on the rationale that itemsets which appear more frequently must be of more importance to the user from the business perspective. In recent years, the research community has also been focused on the infrequent itemset mining problem, i.e., discovering itemsets whose frequency of occurrence in the analyzed data is less than or equal to a maximum threshold. This work addresses the discovery of infrequent and weighted itemsets, i.e., the infrequent weighted itemsets, from transactional weighted data sets. To address this issue, the IWI-support measure is defined as a weighted, frequency of occurrence of an itemset in the analyzed data. In particular, we focus our attention on two different IWI-support measures: (i) The IWI-support- min measure, (ii) The IWI-support-max measure. Furthermore, two algorithms that perform IWI and Minimal IWI mining efficiently. Here, we throw light upon an emerging area called Utility Mining which not only considers the weighted frequency of the itemsets but also considers the utility associated with the itemsets. The term utility refers to the importance or the usefulness of the appearance of the itemset in transactions quantified in terms like profit, sales or any other user preferences. To address this issue, in our system we are proposing the Utility based Infrequent Weighted Itemset mining (UIWIM) to find high utility Infrequent weighted itemset based on minimum threshold values and user preferences

Keywords: Association rule, Utility mining, clustering.

I. INTRODUCTION

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. Data mining is processed with the great deal of consideration in the information construction and in society recently, because of the extensive certainty of huge amounts of data and the future necessitate for figuring such data into practical information and associate. Data mining finds its application mainly on Market basket analysis, Risk analysis, Fraud Detection, DNA data analysis, Web Mining etc.Data mining helps users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Data Mining tasks that try to find interesting patterns from databases, such as association rules, correlations, sequences, episodes, classifiers and clusters.

Association rule mining depicts the implicit relationship among the data attributes. The extraction of interesting correlations, frequent patterns, associations or casual structures among itemsets in the transaction databases is the main aim of Association rule mining. Association rule mining extracts interesting correlation and relation between large volumes of transactions. The identification of sets of items, products, symptoms and characteristics, which often occur together in the given database is a main thing to be performed before framing the rule Itemset mining was focused on discovering frequent itemset, i.e., patterns whose observed frequency of occurrence in the source data (the support) is above a given threshold. Itemset below the threshold value is referred as Infrequent itemset. Considerably less thought has been noticed to mining of infrequent itemsets, even though it has obtained major usage in mining of negative association rules from infrequent itemsets, statistical disclosure risk measurement whereas exceptional patterns in anonymous sample data can direct to statistical disclosure. Then infrequent itemsets is adapted to fraud detection whereas uncommon patterns in financial or tax data might imply unusual action associated with fraudulent behavior and then applied in the field of bioinformatics where unusual patterns in microarray data could imply genetic disorders.

Generally, the primary issues in infrequent patterns mining are identification of appropriate infrequent patterns and efficiently discovering such patterns in large data sets. It is necessary to discover infrequent and weighted itemsets, i.e., the infrequent weighted itemsets, from transactional weighted data sets. To address this issue, the IWI-support measure is defined as a weighted frequency of occurrence of an itemset in the analyzed data. Occurrence weights are derived from the weights associated with items in each transaction by applying a given cost function. We center our attention on two different IWI-support measures: (i) The IWI-support- min measure, which relies on a

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 4, April 2015

minimum cost function, i.e., the occurrence of an itemset in a given transaction is weighted by the weight of its least interesting item, (ii) The IWI-support-max measure, which relies 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. When dealing with optimization problems, minimum and maximum are the most commonly used cost functions. Hence, they are believed appropriate for driving the selection of a worthwhile subset of infrequent weighted data correlations.

The significance of a weighted transaction, i.e., a set of weighted items, is commonly evaluated in terms of the corresponding item weights. Furthermore, the main itemset quality measures (e.g., the support) have also been tailored to weighted data and used for driving the frequent weighted itemset mining process. Particularly, the following problems have been addressed:

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

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

In view of this, utility mining emerges as an important topic in data mining field. Mining high utility itemsets from databases refers to finding the itemsets with high profits.

Utility of an itemset is defined as the product of its external utility and its internal utility.An itemset is called a high utility itemset if its utility is no less than a user-specified minimum utility threshold. Mining high utility itemsets from databases is an important task has a wide range of applications such as website click stream analysis, business promotion in chain hypermarkets, cross-marketing in retail stores, online e-commerce management, mobile commerce environment planning and even finding important patterns in biomedical applications. For example, the sales manager may not be interested in frequent itemsets that do not generate significant profit. Recently, one of the most challenging data mining tasks is the mining of high utility itemsets efficiently.

Identification of the itemsets with high utilities is called as Utility Mining. The utility can be measured in terms of cost, profit or other expressions of user preferences. For example, a computer system may be more profitable than a telephone in terms of profit. Utility mining model define the utility of itemset.

The utility is a measure of how useful or profitable an itemset X is. The utility of an itemset X, i.e., u(X), is the sum of the utilities of itemset X in all the transactions containing X. An itemset X is called a high utility itemset if and only if u(X) >= min_utility, where min_utility is a user defined minimum utility threshold. The main objective of high-utility itemset mining is to find all those itemsets having utility greater or equal to user-defined minimum utility threshold.

II. RELATED WORK

In[1]RAgarwal introduces Frequent itemset mining which is widely used data mining technique. Here, the rules are framed based on the itemset mined which is said to be frequent . The main problem with this is items in a transaction are treated equally.

In[2]Feng Tao et.al presents Weighted Association Rule Mining for frequent itemset mining. In this work the limitation of the conventional Association Rule Mining model is avoided specifically its inability for treating units differently. The presented method uses weights which can be incorporated in the mining process to resolve this difficulty. . *Transaction weight* is a type of itemset weight. It is a value attached to each of the transactions. Usually the higher a transaction weight, the more it contributes to the mining result. However weights are to be priorly assigned which is difficult in real life cases.

In[3]Ke Sun and FengshanBai presented novel framework of w-support mechanism in association rule mining. Initially, the HITS model and algorithm are utilized to obtain the weights of transactions from a database record with simply binary attributes. By derived from these weights, a novel assessment of w-support is described to provide the consequence of item sets. However the presented method differs from the conventional support in taking the quality of transactions into account. Then, the w-confidence and w-support of association rules are described in similarity to the description of confidence and support.Then an Apriori-like algorithm is presented to extract association rules whereas w-confidence and w-support are resulted above fixed thresholds in nature. The analyzed data set is represented by means of Bipartite graph in order to automate item weight assignment.

In[4] Jiawei Han et.al presented novel frequent pattern tree (FP-tree) structure, which is an widened prefix-treeconstruction for storing compressed, critical information about frequent patterns, and expands an effective FP-treebased mining system, FP-growth, for mining the absolute set of frequent patterns by pattern fragment growth.

Effectiveness of mining is attained with three methods:1) a huge database is compressed into a largely reduced. 2)the presented FP-tree-based mining approves a pattern fragment growth process to eliminate the costly generation of a huge number of candidate sets. 3)Finally a partition-based method known as divide-and-conquer system is used to divide the mining job into a set of minor tasks for mining detained patterns in conditional databases, where the search space is reduced appropriately.

In[5]X.wu Efficient mining of both positive and negative association rules . They focus 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

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 4, April 2015

association analysis. Some infrequent itemsets are of 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.

In[6]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.

In[7]U.Yun research on weighted interesting pattern (WIP)mining, a new measure called weight (w)-confidence was proposedto mine correlated patterns with a strong weight affinity. The weight of a pattern P is the ratio of the sum of all its items weight values to the length of P. The w-confidence measure is the ratio of the minimum weight of an item to the maximumweight of an item inside the pattern.

In [8]Utility-based data mining is a new research area interested in all types of utility factors in data mining processes and targeted at incorporating utility considerations in both predictive and descriptive data mining tasks. High utility item set mining is a research area of utility-based descriptive data mining, aimed at finding item sets that contribute most to the total utility. A specialized form of high utility item set mining is utility-frequent item set mining, which – in addition to subjectively defined utility – also takes into account item set frequencies. This paper presents novel efficient algorithms UP-Growth and High Utility Item Set which finds all utility-frequent item sets within the given utility and support constraints threshold. And it is based on efficient methods for frequent item set mining.We start with the definition of a set of terms that leads to the form al definition of utility mining problem. Although DGU and DGN strategies are efficiently reduce the number of candidates in Phase 1(i.e., global UP - Tree). But they cannot be applied during the construct ion of the local UP - Tree (Phase 2). Instead use, DLU strategy (Discarding local unpromising items) to discarding utilities of low utility items from path utilities of the paths and DLN strategy (Discarding local node utilities) to discarding item utilities of descendant nodes during the local UP-Tree construction. Even though, still the algorithm facing some performance issues in phase-2.To overcome this, maximum transaction weight utilizations (MTWU) are computed from all the items and considering multiple of min_sup as a user specified threshold value as shown in algorithm. By this modification, performance will increase compare with existing UP-Tree construction also improves the performance of UP-growth algorithm. An improved utility pattern growth is abbreviated as High Utility Item miner algorithm. Experimental evaluation on datasets shows that, in contrast with High Utility Item Set, and also the performances are evaluated through the factors of time and space complexities**.**

III. IMPLEMENTATION

A.Weighted Transactional Data Set construction

An itemset I is a set of data items.. The support (or occurrence frequency) of an itemset is the number of transactions containing I in T. An itemset I is infrequent if its support is less than or equal to aitemset is said to be minimal if none of its subsets is infrequent. Given a transactional data set T and a maximum support threshold, the infrequent (minimal) itemset mining problem entails discovering all infrequent (minimal) itemsets from T.

IMPLEMENTATION

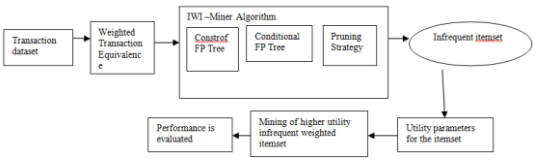

Figure 1:Implementation Detail

Unfortunately, using the traditional support measure for driving the itemset mining process entails treating items and transactions equally, even if they do not have the same relevance in the analyzed data set. To treat items differently within each transaction we introduce the concept of weighted item as a pair $<i_k, w_{kq}>$, where i_k I is an item contained in tq T, while w_{kq} is the weight associated with i_k that characterizes its local interest/intensity in tq. Concepts of weighted transaction and weighted transactional data set are defined accordingly as sets of weighted items and weighted transactions, respectively.

Note that, in general, weights could be either positive, null, or negative numbers. Itemsets mined from weighted transactional data sets are called weighted itemsets. Their expression is similar to the one used for traditional itemsets, i.e., a weighted itemset is a subset of the data items occurring in a weighted transactional data set. The problem of mining itemsets by considering weights associated with each item is known as the weighted itemset mining problem. For the sake of simplicity, by convenient abuse of notation weighted itemsets will be denoted by itemsets whenever it is clear from the context. For the same reason, a generic weighted data set and transaction are denoted by T and tq, respectively,. It is a weighted transactional data set T composed of 6 transactions, each one including four weighted items.

DOI 10.17148/IJARCCE.2015.4482

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 4, April 2015

Since, for instance, the weight of item a in tid 1 (0) is significantly lower than the ones of b (100) and d (71) then a, b, and d should be treated differently during the mining process.

Procedure for Maximum Weighting Function

- Step 1:Wref = lowest among weights in the original Transaction iterative :
- Step 2:Assign the Wref value to each item.
- Step 3:NewWref=next lowest weight in the original transaction-(sum of the previous Wref value)
- Step4:The process is continued until S is empty.

Id	Original Transaction	id	Equivalent Weighted Transaction
1	<a,0><b,100> <c,57><d,71>	1.a	<a,100><b,100><c,100><d,100>
		1.b	<a,-29><c,-29><d,-29>
		1.c	<a,-14><c,-14>
		1.d	<a,-57>

Similarly for minimum Weighting Function

B.Infrequent Weighted Itemset Miner Algorithm

Given a weighted transactional data set and a maximum IWI-support (IWI-support-min or IWI-support-max) threshold ξ, the Infrequent Weighted Itemset Miner algorithm extracts all IWIs whose IWI-support satisfies ξ .Since the IWI Miner mining steps are the same by enforcing either IWI-support-min or IWI-support-max thresholds, we will not distinguish between the two IWI support measure types.

1)FP Tree

IWI Miner is a FP-growth-like mining algorithm that performs projection-based itemset mining. Hence, it performs the main FP-growth mining steps: (a) FP-tree creation and (b) recursive itemset mining from the FPtree 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:(i) A novel pruning strategy for pruning part of the search space early and (ii)a slightly modified FP Tree structure,which allows storing the IWI-support value associated with each node.

2)Pruning Strategy

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 ξ. In particular, since the IWI-support value of an itemset is at least equal to the one associated with the leaf node of each of its covered paths, then the IWI-support value stored in each leaf node is a lower bound IWI-support estimate for allitemsets covering the same paths. Hence, an item (i.e., its associated nodes) is pruned if it appears only in treepaths from the root to a leaf node characterized by IWI-support value greater than ξ.

Steps

- Initially root node is null.
- Insert each transaction into FP Tree
- Start with new path for new prefix.
- Mean time increment the value in the header table for each item
- Items belonging to the header table associated with the input FP-tree are *iteratively* considered.
- Initially, each item is combined with the *current prefix* to generate a new itemset I .
- If I is infrequent, then it is stored in the output IWI set F .
- Then, the FP-tree projected with respect to I is generated and the IWIMining procedure is *recursively* applied on the projected tree to mine all infrequent extensions of I .
- Unlike traditional FP-Growth-like algorithms , IWI Miner adopts a different *pruning* strategy

Algorithm 1 IWI-Miner(T,ξ)

Input:T,a weighted transactional dataset
Input:ξ,a maximum IWI-support thersold
Output:F,the set of IWIs satisfying ξ

1. Initially FP-Tree is constructed
2. for all weighted transaction
3. Calculate Equivalent transaction
4. create and insert into FP tree
5. IWIMining(Tree,ξ,null)

Algorithm 2 IWIMining(Tree,ξ,prefix)

Input: Tree,a FP –tree
Input:ξ,a maximum IWI-support thershold
Input: prefix,the set of items
Output: F,the set of IWIs extending prefix

1. Header table is created that holds for all items i in tree
2. A new item set I is generated with prefix and support of item i
3. The itemset is compared with threshold & store if I - Infrequent item
4. Construct I as conditional pattern FP tree
5. The infrequent items is selected from the set
6. finally apply recursive mining

C.Minimal Infrequent Weighted Itemset Miner Algorithm

Algorithms for discovering minimal infrequent itemsets, i.e., infrequent itemsets that do not contain any infrequent subset .Given a weighted transactional data set and a maximum IWI-support (IWI-support-min or IWI-support-max) threshold, the Minimal Infrequent Weighted Itemset Miner algorithm extracts all the MIWIs that satisfy thersold. The pseudo code of the MIWI Miner algorithm is similar to the one of IWI Miner. 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. In fact, whenever an infrequent itemset I is discovered, all its extensions are not minimal

D.Utility based Mining

Utility u(ip,Tq), is the quantitative measure of utility for item ip in transaction Tq, it is defined by

$u(i_p,T_q)=l(i_p,T_q)*p(i_p)$

DOI 10.17148/IJARCCE.2015.4482

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 4, April 2015

The utility of an itemset X in transaction Tq denoted as u(X,Tq), is defined by $u(X,T_q)= \Sigma u(i_p,T_{q)}$

The transaction utility of transaction Tq denoted as tu(Tq) describes the total profit of that transaction and it isdefined by $tu(T_q)=\Sigma u(i_p,T_q)$

Steps:

1. TU of each transaction is computed.
2. TWU of each single item is also accumulated.
3. Discarding global unpromising items.
4. Unpromising items are removed from the transaction and utilities are eliminated from the TU of thetransaction.
5. The promising items remains in the transaction.

TID	A	b	C
1	0	2	5
2	4	6	0
3	0	3	6

Table 2:Transaction

Item	Profit
a	2
b	6
c	3

Table 3: Item &Profit

The utility of item a in transaction T2 is given by
u(a,t2)= 4*2 =8
The utility of an itemset X is given as,
u(ab,T2) = (4*2)+(6*6)=8+36=44

The utility of itemset x in the entire transaction is given as
U(X)=Σu(i,t)
u(ab,t1)+u(ab,t2)+u(ab,t3)=(0*2)+(2*6)+(4*2)+(6*6)+(0*2)+(3*6)=74

E. Performance evaluation

It is necessary to analyze performance of algorithms . In particular, we should analyze:

1. the impact of the IWI-support thresholds on both the number of mined patterns and the algorithm execution time
2. Comparison between the weighted mining and utility weighted mining of itemsets

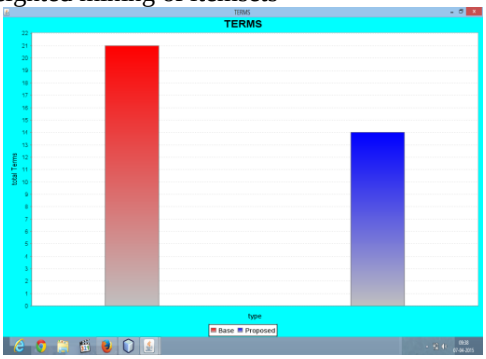

Figure 2:Comparison of no.of terms between weighted itemset and utility based mining

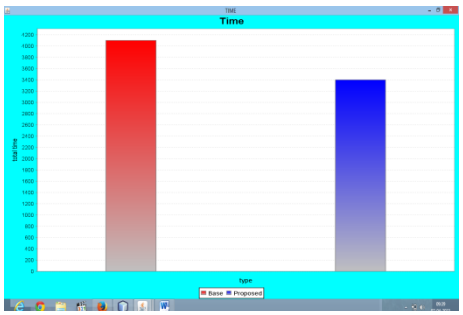

Figure 3:Comparison of Execution time

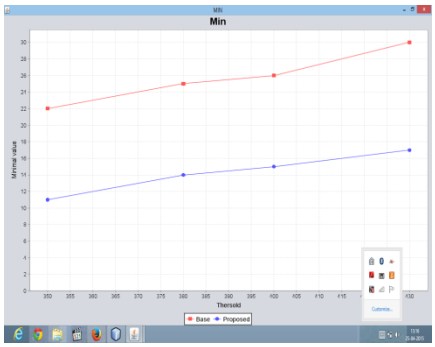

Figure 4:Comparison of different Thersolds

IV. CONCLUSION

Weights associated with the item is used instead of occurrence of item in the transaction. FP growth algorithm along with pruning techniques are efficiently used. One of the latest data mining research areas is Utility Mining which emphasis on all types of utility factors and incorporates utility concepts in data mining tasks. The utility-based descriptive data mining which aims at discovering item sets having high total utility is termed as Highutility itemset mining. High Utility item sets may contain frequent as well as rare itemsets.The utility parameters of the infrequent weighted itemsets is considered for mining purpose.

REFERENCES

[1]. 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.

[2]. 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 DataMining (KDD '03), pp. 661-666, 2003.

[3]. K. Sun and F. Bai, "Mining Weighted Association Rules WithoutPreassigned Weights," IEEE Trans. Knowledge and Data Eng., vol. 20, no. 4, pp. 489-495, Apr. 2008.

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

[5]. 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

[6]. D.J. Haglin and A.M. Manning, "On Minimal Infrequent Itemset Mining," Proc. Int'l Conf. Data Mining (DMIN '07), pp. 141-147,2007.

[7]. U. Yun, Efficient mining of weighted interesting patterns with a strong weight and/or support affinity, Information Sciences 177 (2007) 3477–3499

[8]. A framework for mining interesting high utility patterns with a strong frequency affinity by ChowdhuryFarhan Ahmed a, Syed KhairuzzamanTanbeer a, Byeong-SooJeong a, Ho-Jin Choi b.