An Efficient Mining Approach of Infrequent Weighted Itemset

Ms. Sonia Jadhav, Prof. G. M. Bhandari

***Abstract*— Mining frequent items in data mining is valuable for retrieve the related items present in the dataset. Using input dataset the weighting function is used to calculate minimum and maximum support value is calculated. Using the minimum support value the infrequent weighted itemset support value is calculated. Then the summation is calculated for all the systems in alone. Then combine the two systems and summate the values which are minimum. Finally three systems are combined and summate the value which is minimum among the three. Find the threshold value for the dataset and filter the systems combination. If the summation value is greater than the threshold means then the combination of systems are not considered. Otherwise it is considered for the future result. Then find the equal weighted transaction dataset from transaction dataset. And find the infrequent weighted itemset minimum support value. And also find the threshold value for the equal weighted itemset dataset. And obtain the fulfilled system combinations. Then infrequent weighted itemset miner is used to find the common systems that present in the two results. Using that the infrequent weighted itemset mining is calculated.**

***Index Terms*— association rule, weighted itemset, infrequent itemset mining, weight, Correlation, utility itemset.**

I. INTRODUCTION

Records is a numbers, text, contents, or facts, that can be processed by a machines. The patterns, links or the bond among all this records can provide information. Information can be converted into acquaintance about historical patterns and future development. Data Mining is the process of finding relationship or patterns among dozens of fields in large relational databases. However in other words, It is to extracting information or patterns from data in huge databases. Data mining is the procedure for discovering data from different viewpoints and summarizing it into precious information. This information can be used to improve costs and profits of data information or both. Association rule learning (Dependency modeling) Searches for dealings between variables. For example a supermarket might collect all the data on customer purchasing behavior. By means of association rule learning, the supermarket can resolve which products are frequently bought collectively and exploit this information for marketing purposes. This is sometimes called to as market basket analysis.

The main objective of the Association rule learning is extraction of interesting association, frequent patterns, associations or casual structures among sets of items in the transaction databases or other data repositories. Association rule learning primarily emphasis on extracts interesting correlation and relation between large volumes of transactions. Itemset mining is an exploratory data mining technique mostly used for discovering valuable correlations among items or information. Infrequent item sets are produced from very big or huge data sets by applying some rules or association rule learning algorithms like Apriori method, that take larger computing time to compute all the frequent item sets. Mining of frequent item sets is a central step in many association analysis techniques. The occurrence of frequent item is expressed in terms of the support count.

The first attempt to perform itemset mining [2] which focusing on discovering frequent itemsets i.e. those itemset whose frequency of occurrence in the source data(support) is greater than a given threshold. In this paper we focusing on the infrequent itemset mining problem i.e. discovering of itemset whose frequency of occurrence is less than or equal to a maximum threshold. For illustration, in [3],[4] use algorithms for finding the minimal infrequent itemsets, i.e. infrequent itemset that does not contain any subset of infrequent itemsets. However, significantly less attention has been paid to mining of infrequent weighted item sets, but it has attain important usage in mining of negative association rules from infrequent weighted 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. This paper deal with the discovery of infrequent and weighted item sets i.e. IWI from transactional weighted data sets. To tackle this issue, the IWI-support measure is defined as a weighted frequency of occurrence of an itemset in the investigated data. Happening of weights is derived from the associated weights with items in each transaction by applying given cost function.

Manuscript received Aug, 2015.

***Sonia Jadhav**, P.G. Scholar Department of CSE, BSIOTR, Wagholi, Pune, Pune, India,*

***Prof. G.M. Bhandari,** Head of Department , Department of CSE, BSIOTR, Wagholi, Pune ,Pune ,India.*

ISSN: 2278 – 7798

II. LITERATURE REVIEW

A. Mining Weighted Association Rule

Weighted association rules (WARs) mining are made because importance of the items is different. Negative association rules (NARs) play vital roles in decision-making. But according to the authors the misleading rules occur and some rules are uninteresting when discovering positive and negative weighted association rules (PNWARs) simultaneously. So another factor is added to remove the tedious rules. They propose the support-confidence framework with a sliding interest measure which can avoid generating misleading rules. An attention measure was defined and added to the mining algorithm for association rules in the model. The attention measure was set according to the demand of clients. The experiment demonstrates that the algorithm discovers motivating weighted negative association rules from large database and deletes the dissimilar rules [5].

B. Support and Confidence framework

Firstly try to pushing items weights into the itemset mining process. For originating apriori based itemset mining process the work produces the anti-monotonicity of the weighted support constraint. In many cases the weights are not preassigned this would be happen in the [7] and [10] only. To tackle this problem, we examine the transactional data set was presented as a bipartite hub authority graph and estimated by means of the well-known indexing strategy i.e. HITS , 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. Our approach is distinct from the above mentioned approach since it focuses on mining infrequent itemsets from weighted data instead of frequent ones.

C. Minimal Infrequent itemset using Pattern-Growth

Itemset mining has been an active area of research due to its successful application in various data mining scenarios including finding association rules. Though most of the earlier work has been on finding frequent itemsets, infrequent Itemset mining has demonstrated its value in web mining, bioinformatics and another fields. In this paper, we proposed a new algorithm based on the pattern-growth paradigm to find minimally infrequent itemsets. A minimally infrequent weighted itemset which has no subset which is also infrequent. We also introduce the new concept of residual trees. We also added exploit the residual trees to mine itemsets at multiple level which are having minimum support of different thresholds, and are used for finding itemsets which are frequent for different lengths of the itemset. Finally, we examine the performance of our algorithm with respect to different parameters and show through experiments that it out performs the competing ones.

In this paper, we leverage the pattern-growth paradigm to propose an algorithm IFP min for mining minimally infrequent itemsets. For some datasets, the infrequent sets of itemsets can be exponentially huge. To exposure an infrequent itemset which has an infrequent proper subset is redundant, since the earlier can be deduced from the latter. Hence, it is crucial to report only the minimally infrequent item sets [6].

D. Frequent and Infrequent Itemsets Generation and Association Rule Mining

Association rule mining is a technique that finds all association rules which satisfying the minimum support and minimum confidence in a given transaction database. A common strategy adopted by many association rule mining algorithms is to decompose the problem into two major subtasks:

1) Frequent and Infrequent Itemsets Generation- The itemsets whose objective is to find all the itemsets that satisfy the minimum support threshold are called frequent itemsets. In the contrast they are called infrequent itemsets.

2) Rule Generation- The objective is to extract all the high confidence rules from the frequent itemsets found in the previous step. These rules are called strong rules. The computational requirements for frequent and infrequent itemset generation are generally more expensive than those of rule generation.

III. EXISTING WORK

The discovery of infrequent and weighted itemsets, i.e., the Infrequent Weighted Item sets from transactional weighted datasets. The IWI-support measure is defined as a weighted frequency of occurrence of an itemset in the analyzed data. The IWI support-min measure, which relies on a minimum cost function. That is 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 relies on a maximum cost function. Equivalent weighted transaction dataset is use for the process. It is generated from the transaction dataset. IWI Miner is used for find the common systems that present in the transaction dataset and the equivalent weighted transaction dataset. IWI Mining is used for get the frequent items from the transaction dataset. That is the occurrence of an itemset in a given transaction is weighted by the weight of the most interesting item[1].

A. Limitation:

a. Too many scans are needed for itemsets by using FP-growth algorithm.

b. More candidate item sets are generated because of too many scans.

c. Accuracy is not achieved in the final utility item set.

d. Higher processing time is needed for getting Infrequent weighted itemset.

IV. SYSTEM ARCHITECTURE

A. Organizing the Log Files

The Log Files are collected from the data catalogs. Firstly the patterns are generated, such as each user is initialized with their own id. The quantities which determines the number of times user accessed the websites. After that for each log , the profit table is initialized. However the transaction utility (TU) hereby called as log utility(LU) is estimated by multiplying the quantity and log Profit value.

ISSN: 2278 – 7798

B. Transaction-Weighted Downward Closure

In this firstly we compute the minimum weighted utility. Compute the transaction utility of a transaction Td. After that compute the Transaction-weighted utility of an itemset X which is the sum of the transaction utilities of all the transactions containing X, refer to as TWU(X). Estimate the high transaction weighted utility itemset. Transaction weighted utility itemset which is not less than min util. Evaluate the Transaction Weighted Downward Closure by downward closure property which can be done by applying the transaction weighted utility.

C. UP Catalog

UP catalog maintain the information of transactions and high utility itemsets. Two approaches are applied to minimize the overrated utilities stored in the nodes of global UP Catalog. From this throw away Global Unpromising items while constructing the Catalog.

D. UP Growth

The conditional patterns are generated by tracing the paths in the original Catalog. The item utility which is minimum evaluated by minimum utility threshold. To discover the promising items which are local in the catalog . Then apply Discarding Local Unpromising Items (DLU) to reduce path utilities of the paths while constructing a Local UP-Catalog . The path utility of an item is estimated. This is conditional Catalogs (also called local Catalogs). After that path will be reorganized. This reorganized path construction process is done by using DLN strategies.

E. Final strategy

In final strategy, we compute retrieved path utility. The doubtful items which are local and their estimated Node Utilities from the paths and path utilities of conditional pattern all are extracted. The local Node utilities are decreased for the nodes of local UP-Catalog by estimated utilities of descendant Nodes Reorganized Path utility (DNU). After that Support count is estimated utility itemset. The steps are repeated certainly. After finding all PHUIs, the final step is to identify high utility itemsets and their utilities from the set of PHUIs by scanning original database once again.

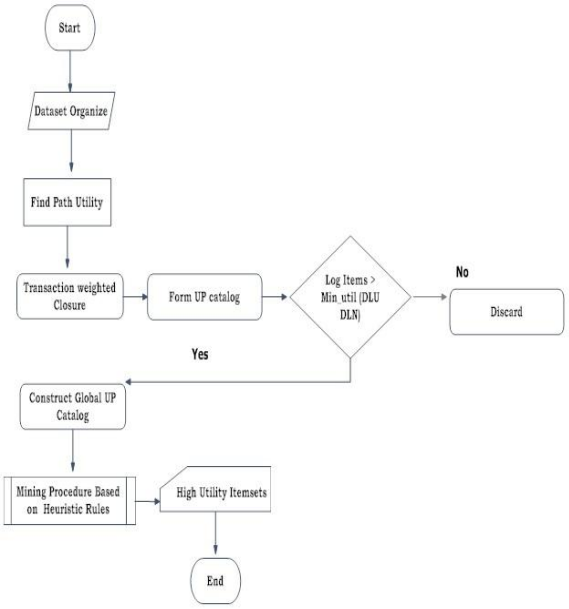


Fig 1: System Architecture

1. Mathematical Model Design

S= {W, H,I, Tree, I, Td, IWI, F, Φ}

Where,

S- Proposed System

W- Input to the System i.e. a weighted Transactional Dataset.

H- Header table for UP-Tree

i-Number of item set.

Generate a new item set by Joining Prefix and I with IWI support value of item i.

Tree – Projection Pattern Tree

Insert Td into global UP-Tree

Apply DGU and DGN strategies on global UP-Tree

Re-construct the UP-tree

I-New item.

Td- Transactional Database

Here, Apply strategy DLN and insert paths into Td

F-Frequent Itemset

F-Infrequent Item sets i.e (F= f1, f2,f3,…..)

IWI- Infrequent Weighted Itemset

IWI-Infrequent Weighted Item sets extending prefix i.e. (IF- if1,if2,if3,…)

Φ- Null value if any.

2. UP Growth

Input: A UP-Tree, a header table H for UP Tree, item set X, Transactional Database D, User Defined Threshold value.

Output: Infrequent Weighted Item sets

Begin

1) Scan Database for Transactions Td->D

2) Determine transaction utility of Td and TWU of itemset(X).

3) Compute min_support

if(TWU (X) <=min _sup) then remove items from transaction database.

else

ISSN: 2278 – 7798

4) insert into header table H and to keep the items in the descending order.
5) Repeat step 4 & 5 until end of the D.
6) Insert Td into global UP-Tree.
7) Apply DGU and DGN strategies on global UP-Tree.
8) Re-construct the UP-tree.
9) For each item a_i in H
10) do
11) Generate a HUI = X U ai
12) Estimate utility of Y is set as ai's utility value in H.
13) put local promising items in CBP into H.
14) Apply strategy DLU to reduce path utilities of the paths.
15) Apply strategy DLN and insert paths into Td.
16) if Td! = null then call for loop
end for
End

V. RESULT

A. *Performance Evaluation*

By the Performance Evaluation the result is achieved on transactional datasets which improve the performance of utility itemset.

TABLE 1: Transactional Dataset

Tid	CPU usage readings			
1	(a,0)	(b,100)	(c,57)	(d,71)
2	(a,0)	(b,43)	(c,29)	(d,71)
3	(a,43)	(b,0)	(c,43)	(d,43)
4	(a,100)	(b,0)	(c,43)	(d,100)
5	(a,86)	(b,71)	(c,0)	(d,71)
6	(a,57)	(b,71)	(c,0)	(d,71)

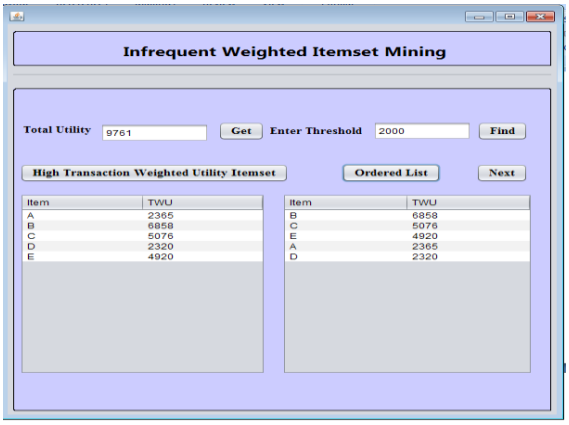
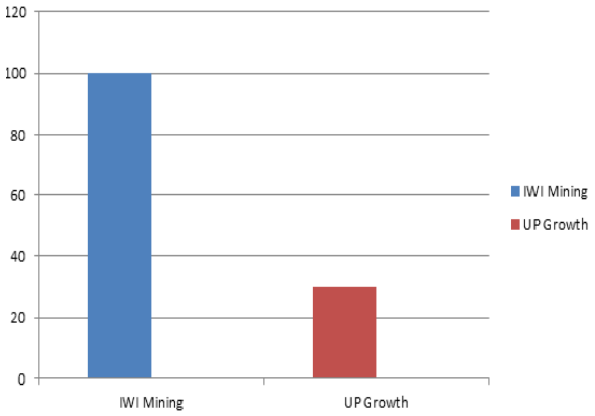


FIG 2: SHOWS HTWU IN DESCENDING ORDER BY USING SELECTED UTILITY ITEMSETS.

B. *Scalabilty Analysis*



VI. CONCLUSION

Existing approach generates the huge number of candidate itemset which degraded the mining performance consequently. The mining performance becomes worse when database contains lots of transaction or under a low minimum utility threshold. So implemented approach in proposed system named UP-Growth (Utility Pattern Growth) algorithm, for discovering high utility item sets and maintaining important information related to utility patterns within databases. UP growth algorithm generates fewer candidates and identify high utility itemsets and their utilities form the set of candidates. UP Growth algorithms that accomplish IWI(Infrequent Weighted Itemset Miner) and MIWI(Minimal Infrequent Weighted Itemset Miner Algorithm) mining efficiently. Experiments show that implemented UP-Growth outperforms the state-of-the-art algorithm substantially and has a good scalability for large database. In particular, UP-Growth is over 10,000 times faster than existing algorithms when database contains lots of long transactions.

REFERENCES

[1]. Infrequent itemset mining Using Frequent Pattern growth , vol. 26 , No. 4, 2014.
[2]. Agrawal R, Imielinski T, and 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.

ISSN: 2278 – 7798

[3]. A. Manning and D. Haglin, "A New Algorithm for Finding Minimal Sample Uniques for Use in Statistical Disclosure Assessment," *Proc. IEEE Fifth Int 1 Conf. Data Mining (ICDM '05),* pp. 290-297, 2005

[4]. A. M. Manning, D. J. Haglin, and J. A. Keane, A Recursive Search Algorithm for Statistical Disclosure Assessment, *Data Mining and Knowledge Discovery, vol.16, no. 2,* pp. 165-196, http://mavdisk.mnsu.edu/haglin, 2008.

[5]. 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.

[6]. A. Gupta, A. Mittal, and A. Bhattacharya, Minimally Infrequent Itemset Mining Using Pattern-Growth Paradigm and Residual Trees, *Proc. Intl Conf. Management of Data (COMAD),*pp. 57-68, 2011.

[7]. F. Tao, F. Murtagh, and M. Farid, "Weighted Association Rule Mining Using Weighted Support and Significance Framework," *Proc. nineth ACM SIGKDD Intl Conf. Knowledge Discovery and Data Mining (KDD 03),* pp. 661-666, 2003.

[8]. 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.

[9]. X. Dong, Z. Zheng, Z. Niu, and Q. Jia, Mining Infrequent Itemsets Based on Multiple Level Minimum Supports, *Proc. Second Intl Conf. Innovative Computing, Information and Control (ICICIC 07),* pp. 528-531, 2007.

[10]. W. Wang, J. Yang, and P.S. Yu, "Efficient Mining of Weighted Association Rules (WAR)," *Proc. Sixth ACM SIGKDD Intl Conf. Knowledge Discovery and data Mining (KDD 00),* pp. 270-274, 2000.

[11]. A. Frank and A. Asuncion, "UCI Machine Learning Repository," http://archive.ics.uci.edu/ml, 2010.

Sonia Jadhav received B. Tech. degree in Computer Science & Engineering from Computer Department of Kurukshetra Institute of Technology & Management, Kurukshetra , from Kurukshetra University, Haryana, India in the year of 2011. She is currently doing M.E. degree in Computer Science and Engineering from 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.

ISSN: 2278 – 7798