Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 4, Issue. 6, June 2015, pg.877 – 881

RESEARCH ARTICLE

Discovery of Infrequent Weighted Itemset with High Utility

Kalyani Tukaram Bhandwalkar[1], Mansi Bhonsle[2]

[1]PG Scholar, Department of Computer Engineering, GHRCOEM, Ahmednagar, University of Pune, India

[2]Asst. Professor, Department of Computer Engineering, GHRCOEM, Pune, University of Pune, India

[1] kalyanibhandwalkar@gmail.com; [2] mansi.bhonsle@gmail.com

Abstract— *Knowledge discovery has been an interesting area of research due to its various applications. Traditionally, frequent pattern mining plays an important role. Generally, infrequent items within the dataset are ignored. Infrequent Itemset mining is a variation of frequent itemset mining where rarely occurring patterns are discovered. This paper tackles the issue of discovering rare and weighted itemsets, i.e., the infrequent weighted itemset (IWI) mining problem. Also high utility itemsets are discovered using UP Growth algorithm. Proposed system not only considers the frequency of the itemsets but also considers the utility associated with the itemsets.*

Keywords—*Data Mining, Frequent Itemset Mining, Infrequent Itemset Mining, Utility Mining, knowledge discovery*

I. INTRODUCTION

"We are living in information age" is a popular saying. Large amount of data is stored into the computer. The explosive growth of data has led to the birth of data mining. Data mining also known as data or knowledge discovery is the process of identifying data from different views and getting the useful information[4] . There is need of the tools which automatically extract knowledge from data. The term knowledge discovery deals with the automatic discovery of information or knowledge within the collection of large databases. Data Mining is technique discovers interesting patterns among the large amount of data. It makes the analysis of data from many different dimensions. It may categories it and identifies the relationships between various data. Data mining is used for automatic discovery of the hidden valuable patterns and relationships from huge volume of data stored in databases for better business decisions making. Mining useful patterns hidden in a database plays an important role in several data mining tasks.

1.1 Frequent Itemset Mining:

Frequent itemsets, as name suggest, are the sets of items often occurring frequently in transactional dataset.eg computer and antivirus software are frequently brought by customer together. It leads to discovery of the association rules from the datasets. Frequent itemsets are appearing with frequency more than a user-specified threshold.

1.2 Infrequent Itemset Mining:

Infrequent itemsets are the sets of items often occurring rarely in transactional dataset.eg computer and laptop software are rarely brought by customer together. Patterns that are rarely found in database are considered to be uninteresting and are ignored by many traditional techniques but they can be used for discovery of the unusual

behavior from the datasets, identifying relatively rare diseases, finding associations between infrequently purchased items in the market basket domain and predicting equipment failure like if {Fire = Yes} is frequent, but {Fire = Yes, Alarm = On} is infrequent, then the alarm system probably is faulting [1]. In medical field, researchers spend more time on studying an abnormal case rather than checking the millions of records of healthy people the study of finding a better treatment approach for a special disease.

1.3 Utility Mining:

To improve the quality of ARM technique there is need to consider the utility of items. Utility Mining is helpful in finding the utility which is a measure to find the usefulness or profitability of an itemset[16].The utility of item means its contribution towards the business profit or quantity of the item sold, etc. Hence Utility mining focuses on identifying the itemsets with high utilities. Many practical situations shows that rare itemsets have high utilities provide very useful insights to the user. For example, in a market analysis, customers purchase washing machine rarely as compared to washing powder, shampoo, soap. But this transactions yield more profit.

II. LITERATURE REVIEW

Apriori algorithm was proposed by R. Agrawal et al in [28] which is used to obtain frequent itemsets from the database. MINimal Infrequent Itemsets (MINIT) is the first algorithm designed specifically for mining minimal infrequent itemsets [2]. MINIT computes both minimal and non-minimal (unweighted) infrequent itemset mining from unweighted data which is based on SUDA2 algorithm and also proved that the minimal infrequent itemset problem is NP-complete problem.

[10] has purposed a way to find out weights of items and weights of transactions without pre-assigned weights in the database. The w-support gives the significance of item sets based on the weights, which differs from the traditional support in taking the quality of transactions into consideration. An Apriori-like algorithm then extracts association rules whose w-confidence and w-support are above some given thresholds specified by user.

ARANIM algorithm mines rare and non-present itemsets in [7]. Non present items are used to detect what is missing in defective process. Also author proposed an approach based on rare patterns to detect suspicious uses and behaviors in the context of a Web application.

The author proposed algorithm in [1] finds the infrequent patterns and non-present patterns but it does not consider any pruning strategy. IFP_min algorithm in [3] uses concept of residual tree to reduce computation time which recursively mines the minimally infrequent itemsets (MIIs) by dividing the IFP-tree into two sub-trees called projected tree and residual tree[3].

Podpecan et al suggested Fast Utility-Frequent Mining (FUFM) algorithm for finding all utility-frequent itemsets within the given utility and support constraints in [12]. TPHURI i.e. Transaction Profitability using HURI algorithm used to discover profitable transactions consisting of high utility rare items [16].

III. PROPOSED SYSTEM

3.1 Problem Statement:

To develop a system for discovering high utility infrequent itemsets from transactional database efficiently. To find the infrequent items in a transactional database is a main task. It results the itemsets whose occurrences is less than threshold value. For that purpose we are proposing IWI algorithm which finds infrequent weighted itemsets and UP Growth algorithm which finds high utility itemsets. Our proposed approach is to enhance the efficiency of mining infrequent itemsets in a transactional dataset, proposing a new methodology, which outcomes the better results than existing methodology.

3.2 System Architecture

In this paper the discovery of infrequent and weighted itemsets from transactional weighted data sets is proposed. The weighted itemset mining problem means mining of itemsets by considering weights associated with each item. A weight is associated with each data item and characterizes its local significance within each transaction. The significance of a weighted transaction, i.e., a set of weighted items, is commonly evaluated in terms of the corresponding item weights [10]. The discovery of interesting infrequent patterns in large datasets is one of the key issues related to the infrequent itemset mining. To solve this problem itemsets with high utility are mined.

Here, IWI ,FPGrowth- like algorithms proposed for discovering infrequent itemsets by using weights for differentiating between relevant items and not within each transaction [4]. IWI-support measure is a weighted frequency of occurrence of an itemset in the given dataset. Cost function calculates the occurrence weights which is derived from the weights associated with items in each transaction by applying a given cost function. This converts the transactional dataset into weighted dataset.

Weight consideration for each distinct item in a transaction in independent manner adds effectiveness for finding infrequent itemset mining.

A) Transactional Dataset: A transactional dataset is a collection of transactions where each transaction is a record of items[16].

B) Weighted Transactional Dataset: The weighted transactional dataset stores weighted values of each item in the transaction. Let $I=\{i_1,i_2,\ldots,i_m\}$ is a set of items then weighted transactional dataset Tw is a set of weighted transactions, where each weighted transaction t^w_q is a set of weighted items $< i_k,w_k^q>$ where w_q is respected weight of item $i_k[4]$..

C) IWI Miner: The IWI Miner mines the infrequent weighted itemset using either IWI support min or IWI-support-max thresholds. IWI Miner adopts an FP- tree like structure [4]. The Infrequent Weighted Itemset Miner algorithm (IWI miner) extracts all infrequent itemsets whose IWI-support satisfies the condition. The pseudo code of IWI miner is same as stated in [4]. IWI Miner performs the main FP-growth mining steps:

(a) FP-tree creation

(b) Recursive itemset mining from the FP tree index. Basic difference between FP-Growth and IWI Miner is that IWI Miner discovers infrequent weighted item sets instead of frequent ones.

D) UP-Tree: To avoid scanning database repeatedly, a compact tree structure called UP-Tree is used .It maintains the information of transactions and high utility itemsets. Each node of UP-Tree contains the item name, count of item, node utility value, name of parent node, link and a set of child nodes. Header table is used to the traversal of UP-Tree [33].

3.3 Algorithms

A. UP-Growth (Utility Pattern Growth) algorithm:

It is proposed for discovering high utility itemsets. UP-Tree is a compact tree structure used in this algorithm. It is proposed to store the necessary information of the transaction database related to the utility patterns. The pseudo code of UP Growth algorithm is same as stated in [32].High utility itemsets are then generated with only two scans of the database.

The proposed method consists of three parts:

1) Construction of UP-Tree by removing TWU(X) <= min sup

2) Generation of the potential high utility itemsets (PHUI) from the UP-Tree by UP-Growth,

3) Identification of high utility itemsets from the set of potential high utility itemsets.

IV. RESULT

Results of work done are as shown in some of the following output screen. Transactional dataset is input for the system.

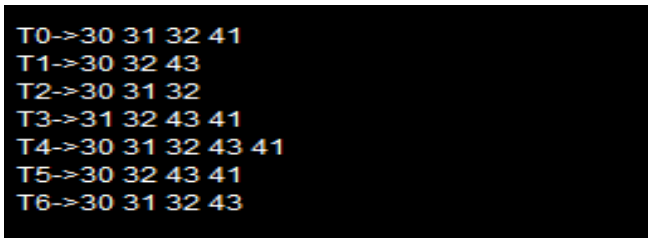

Fig 1 Transactional Dataset

The weighted transactional dataset stores weighted values of each item in the transaction.

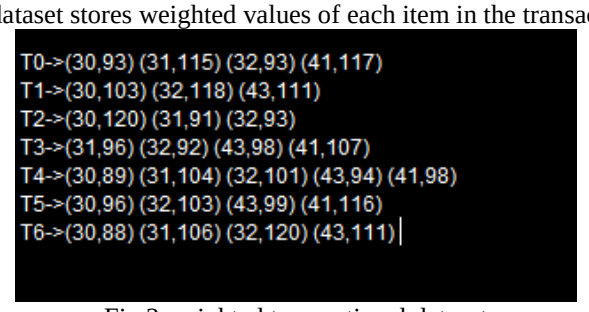

Fig 2 weighted transactional dataset


```
30 = 4
41 = 5
31 = 2
32 = 1
43 = 2
```

Fig 3 Profit Table

```
T0->(30,93) (31,115) (32,93) (41,117) =1280
T1->(30,103) (32,118) (43,111) =752
T2->(30,120) (31,91) (32,93) =755
T3->(31,96) (32,92) (43,98) (41,107) =1015
T4->(30,89) (31,104) (32,101) (43,94) (41,98) =1343
T5->(30,96) (32,103) (43,99) (41,116) =1265
T6->(30,88) (31,106) (32,120) (43,111) =906
```

Fig 4 Transactional utility

```
Transactions -> Itemsets =        RTU
T0->(30,93) (31,115) (41,117) =1187
T1->(30,103) (43,111) =634
T2->(30,120) (31,91) =662
T3->(31,96) (43,98) (41,107) =923
T4->(30,89) (31,104) (43,94) (41,98) =1242
T5->(30,96) (43,99) (41,116) =1162
T6->(30,88) (31,106) (43,111) =786
```

Fig 5 Recognized transaction and RTU values

V. CONCLUSION

Most of the past work has been done on computing frequent item sets than infrequent itemset though it has number of application in various fields like medical, intrusion detection, fraud detection etc. The discovery of interesting infrequent patterns in large datasets is one of the key issues related to the infrequent itemset mining. This paper addresses the discovery of infrequent and weighted itemsets with high utility from transactional weighted data sets using UP growth algorithm. Algorithm scans the database only limited number of times. The proposed system algorithm considers the frequency of the itemsets as well as the utility associated with the itemsets. System algorithm improves the accuracy as compare to existing algorithms for infrequent itemsets.

REFERENCES

[1] Sujatha Kamepalli,Rajasekhara Rao Kurra and Sundara Krishna Y. K.,"Apriori Based: Mining Infrequent and Non-Present Item Sets from Transactional Data Bases," *International Journal of Electrical & Computer Science* IJECS-IJENS Vol:14 No:03

[2] Haglin D. and Manning A. (2007)." On Minimal Infrequent Itemset Mining". In 2007 *International Conference on Data Mining (DMIN'07)*, Las Vegas, June 25-28, 2007, pp. 141- 147.

[3] 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.

[4] Luca Cagliero and Paolo Garza "Infrequent Weighted Itemset Mining using Frequent Pattern Growth", *IEEE Transactions on Knowledge and Data Engineering*, pp. 1- 14, 2013.

[5] SakthiNathiarasan, Kalaiyarasi, Manikandan , "Literature Review on Infrequent Itemset Mining Algorithms" *International Journal of Advanced Research in Computer and Communication Engineering* Vol. 3, Issue 8, August 2014

[6] INFREQUENT PATTERNS, CHAPTER- 5 Introduction to data mining, Pang-Ning Tan, Michael Steinbach, Vipin Kumar Pearson Education, book.

[7] 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.

[8] L.Szathmary,A.Napoli, P.Valtchev, Towards rare itemset mining ,in: *Proceedings of the 19th IEEE Interational Conference on Tools with Artificial Intelligence* , 2007, Volume-1, pp.305-312

[9] JyothiPillai, O.P.Vyas, "Overview of Itemset Utility Mining and its Applications" *International Journal of Computer Applications* (0975 – 8887) Volume 5– No.11, August 2010

[10] K.Sun and F.Bai,"Mining Weighted Association Rules Without Preassigned Weights", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 20, No. 4, April 2008, pages 489-495

[11] B.Barber, H.J Hamilton , Extracting share frequency itemsets with infrequent subsets, Data Mining and Knowledge Discovery 7(2) (2003)153-185.

[12] V. Podpecan , N. Lavrac, I. Kononenkom, A fast algorithm for mining utility-frequent itemsets,in workshop on Constraint-Based Mining and Learning at ECML/PKDD ,2007, pp. 9-20.

[13] 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.

[14] 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.

[15] 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.

[16] Jyothi Pillai and O.P.Vyas ,"TRANSACTION PROFITABILITY USING HURI ALGORITHM [TPHURI]", *International Journal of Business Information Systems Strategies (IJBISS)* Volume 2, Number 1,November 2013

[17] Barber B., Hamilton, H. J. "Extracting share frequent itemsets with infrequent subsets", Data Mining and Knowledge Discovery, 7(2) (2003), pp 153-185.

[18] J. Yang and J. Logan. A data mining and survey study on diseases associated with paraesophageal hernia. In AMIA Annual Symposium Proceedings, pages 829–833, 2006.

[19] Pang-Ning Tan, Michael Steinbach, Vipin Kumar Introduction to data mining, Pearson Education, book.

[20] Jiawei Han · Hong Cheng · Dong Xin · Xifeng Yan. Frequent pattern mining: current status and future directions.

[21] Ling Zhou, Stephen Yau, "Efficient association rule mining among both frequent and infrequent items" Computers and Mathematics with Applications 54 (2007) 737–749.

[22] Laszlo Szathmary1, Petko Valtchev2, Amedeo Napoli3, and Robert Godin2 2012." Efficient Vertical Mining of Minimal Rare Itemsets" ISBN 978,84{695{5252{0, pp. 269{280, 2012.

[23] Diti Gupta and Abhishek Singh Chauhan "Mining Association Rules from Infrequent Itemsets: A Survey*", International Journal of Innovative Research in Science, Engineering and Technology*, ISSN: 2319-8753, Vol. 2, Issue 10, pp. 5801 – 5808, 2013

[24] Younghee Kim, Wonyoung Kim and Ungmo Kim "Mining Frequent Itemsets with Normalized Weight in Continuous Data Streams", *Journal of Information Processing Systems*, Vol.6, No.1, March 2010

[25] Laszlo Szathmary, Petko Valtchev, Amedeo Napoli, "Finding Minimal Rare Itemsets and Rare Association Rules", Proceedings of the 4th International Conference on Knowledge Science, Engineering and Management (KSEM 2010) 6291 (2010) 16—27.

[26] Jyothi Pillai, O.P.Vyas, "CSHURI – Modified HURI algorithm for Customer Segmentation and Transaction Profitability*", International Journal of Computer Science, Engineering and Information Technology (IJCSEIT),* Vol.2, No.2, April 2012, pp 79-89.

[27] J. Han, J. Pei, and Y. Yin, "Mining frequent patterns without candidate generation," Procedings of ACM *SIGMOD International Conference on Management of Data*, ACM Press, Dallas, Texas, pp. 1-12, May 2000.

[28] R. Agrawal, T. Imielinski, and A. Swami, "Mining association rules between sets of items in large databases," in *Proceedings of the 1993 International Conference on Management of Data (SIGMOD 93)*, May 1993, pp. 207–216.

[29] *en.wikipedia.org/wiki/**Data_mining***

[30] *en.wikipedia.org/wiki/Association_rule_learning*

[31] Frequent Itemset Mining Dataset Repository http://fimi.ua.ac.be/data/

[32] Shoban Babu Sriramoju*,Madan Kumar.Chandran, " UP-Growth Algorithms for Knowledge Discovery from Transactional Databases" International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 2, February 2014

[33] V.S. Tseng, C.W. Wu, B.E. Shie, and P.S. Yu. "UP-Growth: An efficient algorithm for high utility itemset mining," ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2010, Washington, DC, United states.

