Enhancing Data mining Techniques for Graphical Analysis and Representation of High Utility Itemset

P. Payal Swamy[1] Amit Pimpalkar[2]

[1,2]*Department of Computer Science and engineering*
G.H.Raisoni Academy of Engineering & Technology, Nagpur

Abstract -Identifying high utility objects can be defined as invention of itemsets with high utility just like profit. There are many number of approaches for discovering high utility objects but the difficulty with them is, they produce several number of candidate itemsets. Generation of many number of itemsets leads to degrade the performance in terms of execution time and the speed required. If the database contains lots of long transactions or long high utility itemsets then it would be very complicated to deal with them. Here we propose UP++ Growth algorithm for finding the graphical representation and compared the results with the already existing UP Growth and (CHUD) closed$^+$ High Utility Itemset Discovery. Further (DAHU) Derive All High Utility Itemsets will be used to recover all (HUIs) high utility itemsets without accessing the original database. Results will be observed on both real as well as synthetic data

Keywords – *Data mining, Utility mining, high utility mining.*

1. INTRODUCTION

Threshold value is defined as the minimum limit that must be present, mining of frequent itemsets focuses on the threshold value only and finds items in a given database which passes the threshold value or we can say that, the items that occurs above the defined (set) frequency threshold. In this the quantity or profit of the purchased items is not considered. Therefore there is no need for users to find the importance of the items that are present in database .That’s why new technique is introduced called as high utility mining. From transactional database, mining of high utility itemsets is done and this mining defined as discovery of itemsets with high utility like profits.

High utility itemsets have utility greater than user-defined minimum utility threshold if the utility is less than defined one then it is called a low-utility itemsets, threshold value is defined before only. Progress in different technologies has made it possible for retail organization to collect and store massive amount of sales data, referred to as the basket analysis. Earlier defined methods, that were used for utility mining produces large itemsets that will degrade performance consequently this has became a challenging problem to the mining performance. To address this issue, we proposed a new algorithm with a compact data structure which will help efficiently in discovering high utility itemsets from transactional databases.

2. RELATED WORK

Asha B., et.al examined the problem of generation of high number of candidate itemsets, which they tried to overcome using UP-Growth and UP-Growth+ algorithm and there functionality is extended for incremental database.

Rakesh A., et.al[3] examined the problem of discovery of association rule between items in large transaction using the best features Apriori and AprioriTid named hybrid Apriori algorithm, the new hybrid algorithm is compared to the previously known AIS and SETM algorithm, the algorithm outperforms then the previously known algorithm in terms of execution times, the execution time decreases as the number of items in the database increases also the scale up experiments conducted shows that the Hybrid algorithms scales linearly with the number of transactions.

Shankar S., et.al suggested a utility based minig technique, to overcome the problem of frequent itemsets which only reflect the significance of correlation between items but not the semantic significance of items. They proposed a new algorithm namely (FUM) fast utility mining and compared the performance of it with the already existing Umining algorithm. The FUM finds all the utility within a particular threshold and performs much better than the Umining algorithms even the Umining has a drawback that it only prune only few high utility itemsets whereas the FUM generates entire high utility itemsets without missing any.

The problem of formation of huge number of candidate itemsets is reduced by using a new approach namely (IIDS) isolated item discarding strategy and the number of candidate itemsets is reduced in every database scan .In this paper Yein C.et al. [8] used two algorithm FUM and DCG+ for calculating high utility itemsets and these algorithms performs exceptionally better than the other earlier algorithms but the problem of multiple database scan comes along and thus it became a drawback.

Utility mining being one of the emerging approach today, many new algorithms are implemented to improve the efficiency and reduce the candidate itemsets generated in mining high utility itemsets. A new algorithm named (EUP- Growth +) Enhanced utility pattern with set of strategies is used to achieve the above mentioned. In this, a tree data structure named IMUP-Tree is maintained using hashing techniques and pruning of the high utility items is done. The experimental results showed that EUPG+ performs better than UPG+ and UPG algorithm.

Han J.,etal.,[7] developed an algorithm for mining frequent pattern, not including candidate generation. While mining of frequent pattern, high utility of patterns is not considered, the frequent patterns that will be mined may be having high profits or not and thus it is not necessary that they contribute to the overall profit. Algorithm named FP growth don't generate candidate itemsets and thus the efficiency is improved but the drawback with this algorithm is only frequent patterns are considered.

To generate (HTWUIs) high transactional weighted utility itemsets in phase I and to overcome the problem of scanning database multiple number of times and generate HTWUIs, a tree-based algorithm named IHUP, for mining high utility itemsets was proposed by Ahmed C.et al.[2] In this algorithm an IHUP-Tree is constructed to maintain the information of high utility itemsets. Here with the constructed of tree are the global as well as local non-promising itemsets are removed and thus the itemsets with maximum utility remain, the advantage over all this is only one database scan is required.

Widespread studies have been done for finding frequent itemsets, Apriori algorithm , which is the lead the way for efficiently mining association rules from large databases but has many drawbacks such as many time scan of database and candidate generation. The tree-based approaches such as FP-Growth were proposed later to effectively generate frequent patterns without candidate generation as well as number of database scans required is also reduced to only two.

3. Proposed Method

The proposed method working will be as follow –

In this diagram, the transactional database consisting of all the transactions involved. A threshold value will be predefined which is known as the minimum utility threshold, below that value all the items are rejected.

All the items having threshold lower than that are known as the unpromising itemsets and those are removed .The generated items which are promising are then recognized and will be then used for creating the UP tree and performing CHUD algorithm. On the generated UP tree the UP and UP++ algorithm is applied. Association rule mining is used to gather all the items, DAHU (derive all high utility itemsets) is applied for that .

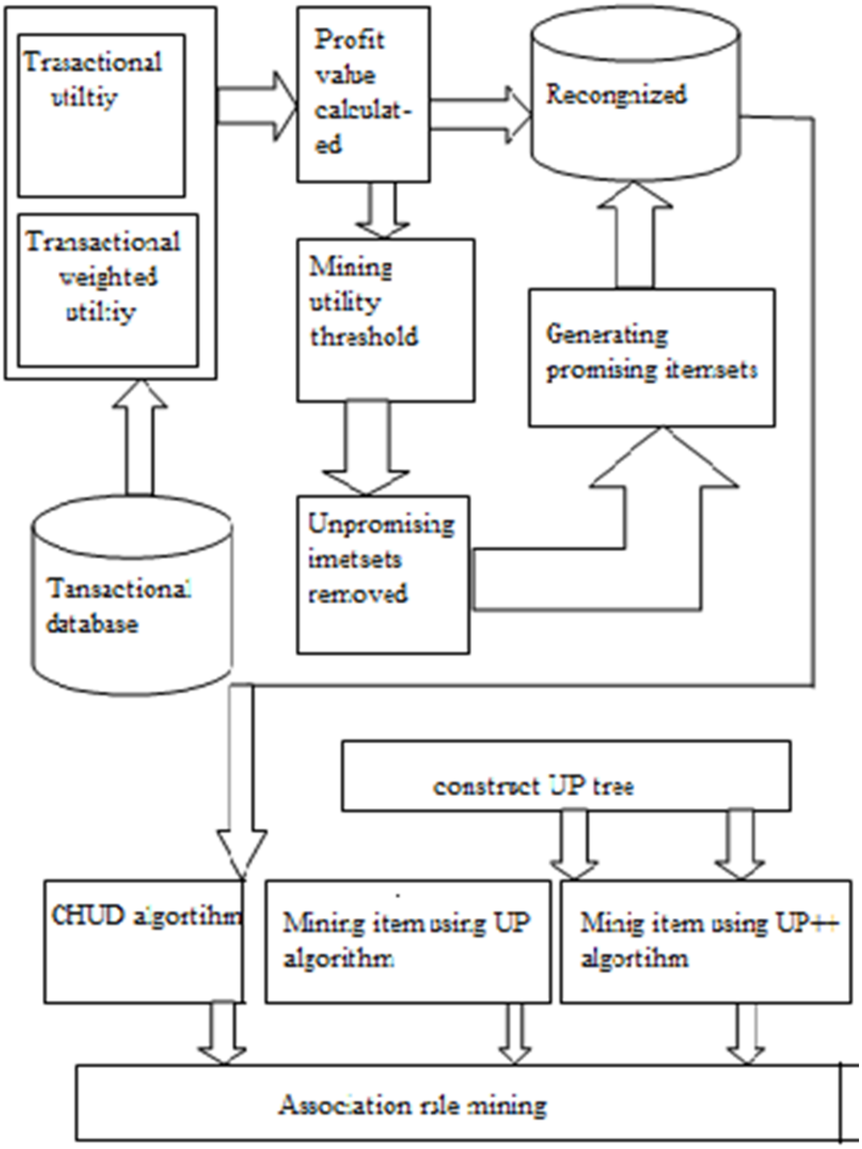

Figure 1:- Showing the program flow.

The architectural diagram consist of transactional database consisting of all the transactions after that the transactional utility and the transactional weighted utility of each item is calculated .From the table TWU and TU profit value will be calculated which help to decide the threshold value .The defined threshold value will help to remove unpromising itemsets, the itemsets having threshold value below the defined one are unpromising and hence rejected. All the promising itemsets are then recognized to construct UP tree mining.

In the proposed strategy first UP tree is constructed

The framework of the UP++ strategy consist of following steps :-

(1) Firstly, construction of UP-Tree is done.
(2) PHUIs are generated from the UP tree
(3) Thirdly, classification of high utility itemsets is done from the set of potential high utility itemsets.

The TU for each item in the transactional database will be calculated along with the TWU for each other item. The TWU table will be arranged in the decreasing order and PHUI will be calculated.

Different strategies will be applied for construction of UP tree so that, the proposed algorithm perform much better than the existing

(1) The TU and TWU for each items is calculate.
(2) Then the UP++ Growth algorithm is applied it is a modified algorithm of UP Growth i.e. UP-Growth++ reduce the execution time by effectively identifying high utility itemsets.

After that to the itemsets that we have we separately apply CHUD Algorithm CHUD is an extension of DCI Closed, one of the currently best methods to mine closed Itemsets CHUD is adapted for mining CHUIs and include several effective strategies for reducing the number of candidates generated in Phase1. Finally, the Main procedure performs Phase2 on these candidates to obtain all CHUIs.

At the end the UP Growth, UP ++ Growth and CHUD algorithm results are seen for the synthetic and real database.

5. CONCLUSION

In this paper, we have proposed an efficient algorithm named UP++ Growth for mining high utility itemsets from transaction. UP Growth and UP-growth++ algorithms both are efficient algorithms for high utility itemsets mining they gives better performance on incremental database. This algorithm works better even if some changes are made to the database they generates candidate itemsets with only two scans of the original database. A tree named UP-Tree is proposed for maintaining the information of high utility itemsets. Besides these we have four strategies that will lead us to decrease the estimated utility value and to enhance the mining performance in utility mining.

In the experiments, both of synthetic and real datasets are used to analyze performance. The mining performance is enhanced as both the search space and the number of candidates are effectively reduced by the proposed strategies. The experimental results show that UP++Growth outperforms the state-of-the-art algorithms substantially, especially when the database contains lots of long transactions. The three algorithms that we are taking will be compared on the basis of their memory usage, utilization and many more factors.

In this paper, we will study graphical representation using different algorithms. However, there are many other compact representations [9], [10] and they have not been included till now, this other representation can be included in the future

REFERENCES

[1] V. S. Tseng, B. E. Shie, C.-W. Wu and P. S. Yu, Fellow," Efficient Algorithms for Mining the Concise and Lossless Representation of High Utility Itemsets", IEEE Transaction on Knowledge and Data Engineering, vol. 27, no. 3, March,2015.

[2] F. A. Chawdhary, S. K. Tanbeer, B.-S. Jeong, and Young-Koo Lee, Member, IEEE "Efficient Tree Structures for High Utility Pattern Mining in Incremental Databases" IEEE Trans. Knowledge and Data Engineering, Vol. 21, Issue 12, pp. 1708-1721, December 2009.

[3] R. Agrawal and R. Srikant, “Fast algorithms for mining association rules”, In Proceedings of the 20th International Conference on Very Large Data Bases, pp. 487-499, 1994.

[4] C. F. Ahmed, S. K. Tanbeer, B.-S. Jeong, and Y.-K. Lee, “Efficient tree structures for high utility pattern mining in incremental databases”, In IEEE Transactions on Knowledge and Data Engineering, Vol. 21, Issue 12, pp. 1708-1721, 2009.

[5] R. Chan, Q. Yang, and Y. Shen, “Mining high utility itemsets”, In Proceedings of Third IEEE Int'l Conference on Data Mining, pp. 19-26, Nov., 2003.

[6] A. Erwin, R. P. Gopalan, and N. R. Achuthan,”Efficient mining of high utility itemsets from large datasets”,In Proceedings of PAKDD, LNAI 5012, pp. 554-561,2008.

[7] J. Han, J. Pei, and Y. Yin, “Mining frequent patterns without candidate generation”, In Proceedings of the ACM-SIGMOD International Conference on Management of Data, pp. 1-12, 2000.

[8] Y.-C. Li, J.-S. Yeh, and C.-C. Chang, “Isolated items discarding strategy for discovering high utility itemsets”, In Data & Knowledge Engineering, Vol. 64, Issue 1, pp. 198-217, Jan., 2008.

[9] Y. Liu, W. Liao, and A. Choudhary, “A fast high utility itemsets mining algorithm”, In Proceedings. of the Utility-Based Data Mining Workshop, 2005.

[10] T. Hamrouni, S. Yahia, and E. M. Nguifo, “Sweeping the disjunctive search space towards mining new exact concise representations of frequent itemsets”, Data Knowledge Engineering, vol. 68, pp. 1091–1111, 2009.

[11] H. Li, J. Li, L. Wong, M. Feng, and Y. Tan, “Relative risk and odds ratio: A data mining perspective,” in Proceedings ACM SIGACT-SIGMOD-SIGART Principles Database Systems pp. 368–377,2005.

[12] J.-F. Boulicaut, A. Bykowski, and C. Rigotti, “Free-sets: A condensed representation of Boolean data for the approximation of frequency queries,” Data Mining Knowledge Discovery, vol. 7, no. 1,pp. 5–22, 2003.

[13] T. Calders and B. Goethals, “Mining all non-derivable frequent itemsets”,in Proceedings International Conference Principles Data Mining Knowledge Discovery, pp. 74–85, 2002.