A survey on high utility item set mining with various techniques

Kirti Santoki
M.E. Student, Dept. of Computer Engineering, BHGCET, Rajkot, India
Priyanka Raval
Assistant Professor, Dept. of Computer Engineering, BHGCET, Rajkot, India

***Abstract*— mining high utility itemset from transaction database refers to discovery of itemset with high utility like profits. Itemset Utility Mining is an extension of Frequent Itemset mining, which discovers itemsets that occur frequently. In High Utility Itemset Mining the goal is to recognize itemsets which have utility values above a given utility threshold. The high utility itemset is the itemset with a utility not less than a user specified minimum support threshold value; else that itemset is treated as a low utility itemset. In this paper, we present a literature survey of the present state of research and the various algorithms and its limitations for high utility itemset mining.**

Keywords—Association rules mining, frequent itemset mining, high utility itemset mining.

I. INTRODUCTION

Data mining is the process of revealing non-trivial, previously unknown and potentially useful information from large databases. Association Rule Mining (ARM) is an important data mining technique that is used to discover the patterns/rules among items in a large database [1]. The goal of ARM is to identify group of items which occur together, for example in a market basket analysis. Mining association rules can be decomposed into two steps: the first is generating frequent itemsets. The second is generating association rules. The main challenge in association rule mining is to identify frequent itemsets. Finding frequent itemset is one of the important steps in association rule mining. Since the solution of second sub-problem is straight forward, most of the researchers had focus on how to generate frequent itemsets.

Frequent itemsets are the itemsets which occur frequently in the transaction database. The objective of Frequent Itemset Mining is to identify all the frequent itemsets in a transaction database. Moreover, items having high and low selling frequencies may have low and high profit values, respectively. For example, some frequently sold items such as bread, milk and pen may have lower profit values as compared to that of infrequently sold higher profit value items such as gold, platinum and diamond. Therefore, finding only traditional frequent patterns in a database cannot fulfill the requirement of finding the most valuable itemsets/customers that contribute to the major part of the total profits in the real world retail database. This gives the motivation to develop a mining model to discover the itemsets/customers contributing to the majority of the profit. A frequent itemset is the itemset having frequency support greater a minimum user specified threshold. [1]

II. HIGH UTILITY ITEMSET MINING

The high utility itemset mining problem is to find all itemsets that have utility larger than a user specified value of minimum utility. The value or profit Associated with every item in a database is called the utility of that itemset. For example computer system is more profitable them telephone in terms of profit.[2]

Utility define as Interestingness, profitability or importance of item. Utility measured in terms of cost profit or other user preference.
Utility of items in transaction database involves following two aspects:

(1) The importance of distinct items, called external utility (e), i.e. unit profit and

(2) The importance of items in transactions, called internal utility (i), i.e. quantity

Utility of Itemset (U) = external utility (e) * internal utility (i).

In many applications like cross-marketing in retail stores, online e-commerce management, website click stream analysis and finding the important pattern in biomedical applications High utility mining are widely used. [2]

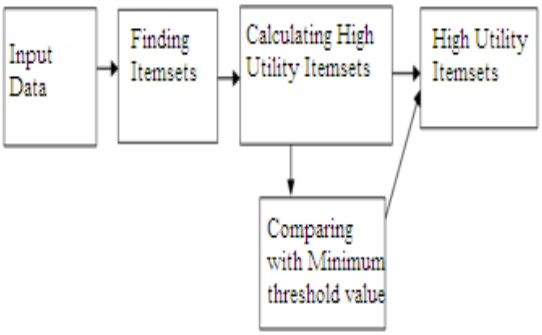

Fig. 1 Flow diagram of HUIM

III. LITERATURE SURVEY

A. Two phase algorithm

This method maintains a Transaction-weighted Downward Closure Property [3]. Thus, only the combinations of high transaction weighted utilization itemsets are added into the candidate set at each level during the level-wise search. Phase I may overestimate some low utility itemsets, but it never underestimates any itemsets. In phase II, only one extra database scan is performed to filter the overestimated itemsets. Two-Phase demands multiple databases scan and generate a huge number of candidate itemsets because of a level-wise method.

B. Compressed Transaction Utility (CTU-Mine)

Erwin et al in [4] observed that the conventional candidate-generate-and-test approach for identifying high utility itemsets is not suitable for dense date sets. Their work proposes a novel algorithm CTU-Mine which mines high utility itemsets using the pattern growth approach. A similar argument is presented by Yu et al. Existing algorithms for high utility mining are column enumeration based adopting an apriori like candidate set generation and test approach and thus are inadequate in datasets with high dimensions

C. Temporal High utility itemset (THUI)

A novel method, namely THUI (Temporal High Utility Itemsets) –Mine was proposed by V.S. Tseng et al in[3] for mining temporal high utility itemsets from data streams efficiently and effectively. The novel contribution of THUI-Mine is that it can effectively identify the temporal high utility itemsets by generating fewer temporal high transaction weighted utilization 2-itemsets such that the execution time can be reduced substantially in mining all high utility itemsets in data streams. In this way, the process of discovering all temporal high utility itemsets under all time windows of data streams can be achieved effectively with limited memory space, less candidate itemsets and CPU I/O time. This meets the critical requirements on time and space efficiency for mining data streams. The experimental results show that THUI-Mine can discover the temporal high utility itemsets with high performance and less candidate itemsets as compared to other algorithms under various experimental conditions. Moreover, it performs scalable in terms of execution time under large databases. Hence, THUI-Mine is promising for mining temporal high utility itemsets in data streams.

D Utility pattern growth (UP-growth)

To address issue of generating a large number of candidates, UP-Growth [5] (V.S Tseng et al., 2010) has recently been proposed and it uses PHU (Potential High Utility) model. For reducing the number of candidate itemsets, the UP-Growth applies four strategies, DGU (Discarding Global Unpromising items), DGN (Decreasing Global Node utilities), DLU (Discarding Local Unpromising items), and DLN (Decreasing Local Node utilities). Besides, it constructs a tree structure, named UP-Tree, with two database scans and conducts mining high utility itemsets. In other words, it demands three database scans for discovering high utility itemsets. In the first database scan, TWU values of each item are accumulated. In the second database scan, items having less TWU values than the user-specified minimum utility threshold are removed from each transaction. In addition, items in transactions are arranged according to TWU descending order and the transactions are inserted into the UP-Tree. In this stage, DGU and DGN are applied for reducing overestimated utilities. After that, high utility itemsets are generated from the UP-Tree with DLU and DLN.

E. High utility itemset miner (HUI-Miner)

Liu & Qu (2012) proposed HUI-Miner algorithm [6]. It is a high utility itemsets with a list data structure, called utility list. It first creates an initial utility list for itemsets of the length 1 for promising items. Then, HUI-Miner constructs recursively a utility list for each itemset of the length k using a pair of utility lists for itemsets of the length k-1. For mining high utility itemsets, each utility list for an itemset keeps the information of TIDs for all of transactions containing the itemset, utility values of the item set

in the transactions, and the sum of utilities of the remaining items that can be included to super itemsets of the itemset in the transactions. The distinct advantage of HUI-Miner is that it avoids the costly candidate generation and utility computation.

F. Faster high utility itemset (FHM)

Philippe Fournier-Viger (2014) proposed FHM algorithm [6]. It extends the Hui-Miner Algorithm. It is a Depth-first search Algorithm. It relies on utility-lists to calculate the exact utility of itemsets. This algorithm integrates a novel strategy named EUCP (Estimated Utility Co-occurrence Pruning) to reduce the number of joins operations when mining high-utility itemsets using the utility list data structure. Estimated Utility Co-Occurrence Structure (EUCS) stores the transaction weighted utility (TWU) of all 2-itemsets. It built during the initial database scans. EUCS represented as a triangular matrix or hash map. The memory footprint of the EUCS structure is small. FHM is up to 6 times faster than HUI-Miner.

G. Efficient high utility itemset (EFIM)

EFIM (Efficient high-utility Item set Mining), which introduces several new ideas to more efficiently discovers high-utility item sets both in terms of execution time and memory [7]. EFIM relies on two upper-bounds named sub-tree utility and local utility to more effectively prune the search space. It also introduces a novel array-based utility counting technique called Fast Utility Counting to calculate these upper-bounds in linear time and space. Transaction merging is obviously desirable. However, a key problem is to implement it efficiently. To find identical transactions in O(n) time,, sort the original database according to a new total order T on transactions. Sorting is achieved in time, and is performed only once. Projected databases generated by EFIM are often very small due to transaction merging.

Table 1.Performance summery of a survey

Sr. no.	Studies	Year	Dataset	Method used	algorithm	Limitation
1	Ying Liu, Wei-keng Liao, Alok Choudhary	2005	Transaction dataset	Level wise approach	Two phase	Multiple scan of database and generate many candidate itemset
2	Alva Erwin, Raj P. Gopalan, N.R. Achuthan	2007	Transaction dataset	Pattern growth approach	Compressed Transaction Utility(CTU-Mine)	Complex for Evaluation due to the Tree structure
3	Vincent S. Tseng, Chun-Jung Chu, Tyne Liang	2008	Transaction dataset	Pattern growth	Temporal high utility itemset mining(THUI)	Huge memory requirement and a lot of false candidate itemset
4	Vincent S. Tseng, Cheng-Wei Wu, Bai-En Shie, and Philip S. Yu	2010	Transaction dataset	Pattern growth	Utility pattern growth(UP-growth)	Complex for Evaluation due to the Tree structure
5	Mengchi Liu, Junfeng Qu	2012	Transaction dataset	Level wise approach	High utility itemset miner(HUI-Miner)	Calculating the utility of an itemset joining utility list is very costly.
6	Philippe Fournier-Viger, Cheng-Wei Wu, Souleymane Zida, Vincent S. Tseng	2014	Transaction dataset	Level wise approach	Faster high utility itemset mining(FHM)	Static database, large memory overhead

IV. CONCLUSION

In data mining Association Rule Mining is one of the most important tasks. A large number of efficient algorithms are available for association rule mining, which considers mining of frequent itemsets. But an emerging topic in Data Mining is Utility Mining, which incorporates utility considerations during itemset mining. Utility Mining covers all aspects of economic utility in data mining and helps in detection of itemset having high utility, like profit. High Utility itemset mining is very beneficial in several real-life applications. In this paper survey paper, we provide the various method of high utility itemset mining and comparison of all technique with transaction dataset, which method is used and limitation of each algorithm.

V. REFERENCES

[1] U Kanimozhi, J K Kavitha, D Manjula, Mining High Utility Itemsets – A Recent Survey, International Journal of Scientific Engineering and Technology, Volume No.3 Issue No.11, pp: 1339-1344, 2014.
[2] Maya joshi, Mansi patel, A Survey on High Utility Itemset Mining Using Transaction Databases, International Journal of Computer Science and Information Technologies, Vol. 5 (6) ,7407-7410,2014.
[3] Jyothi Pillai, O.P. Vyas, Overview Of Itemset Mining And its Application, International Journal of Computer Applications (0975 – 8887) ,Volume 5– No.11, August 2010.
[4] Sudip Bhattacharya, Deepty Dubey, High Utility Itemset Mining, International Journal of Emerging Technology and Advanced Engineering , Volume 2, Issue 8, August 2012.
[5] Hua-Fu Li, Hasin-Yug Huang, Yi-Cheng Chen, Yu-Jiun Liu, Suh-Yin Lee, Fast and Memory Efficient Mining of High Utility Itemsets in Data Streams, IEEE International Conference on Data Mining,2008.
[6] Shekhar Patel B Madhushree, A Survey on Discovering High Utility Itemset Mining from Transactional Database, Information and Knowledge Management, Vol.5, No.12, 2015
[7] R.Shyamala Devi, D.Shanthi, A Survey Mining High Utility Item Sets And Frequent Item Sets , International Journal of Innovative Research in Science, Engineering and Technology, Vol. 4, Issue 12, December 2015.
[8] Smita R. Londhe,, Rupali A. Mahajan,, Bhagyashree J. Bhoyar ,"Overview on Methods for Mining High Utility Itemset from Transactional Database", International Journal of Scientific Engineering and Reaserach (IJSER), Volume 1 Issue4, Deccember2013.