Webpage: www.ijaret.org
UGC Approved Journal-43847

Volume 6, Issue III, Mar 2018
eISSN 2320-6802

INTERNATIONAL JOURNAL FOR ADVANCE RESEARCH IN ENGINEERING AND TECHNOLOGY

WINGS TO YOUR THOUGHTS.....

Survey on high utility itemset mining with various techniques

Neha Agrawal[1], Amit Sariya[2]

[1]Alpine Institute of Technology, Ujjain
Gram Chandesra,Ujjain (M.P.) India
agrawal.na@gmail.com
[2]Alpine Institute of Technology, Ujjain
Gram Chandesra,Ujjain (M.P.) India
amit.sariya861@gmail.com

Abstract: *Utility mining is the emerging research topic in the field of data mining. It addresses the limitation of frequent pattern mining which aims to find the frequent itemsets from a transactional database or a relational database. Among utility mining problems, high utility pattern mining with the itemset framework is more challenging than the other categories of utility mining and frequent pattern mining. The utility mining not only considers the frequency but also see the utility associated with the itemsets. The main objective of utility mining is to extract the itemsets with high utilities, by considering user preferences such as profit, quantity and cost. Several researches about utility pattern mining have been proposed. This research paper presents a review of the various approaches and algorithms for high utility pattern mining.*

Keywords: *Data mining, Frequent Patterns, High Utility Pattern Mining, High Utility Itemsets, High Utility mining Algorithms.*

1. INTRODUCTION

Data Mining is defined as a process of extracting implicit, previously unknown and potentially useful information from large databases [1].The information thus obtained is useful in an enterprise's decision making process. The frequent itemsets mining, a type of association rule mining, was developed in 1990s to identify the frequent itemsets (set of items) from transactional or relational data sets. Frequent itemsets are the itemsets which occur frequently in the transaction database. It has been used extensively in commercial marketing . The notion of frequent itemsets was introduced by Agrawal et al [2]. The most important concept of frequent itemsets mining is "inimal support" Among all items bought by a customer in one transaction there can be many subsets of items (itemsets), i.e. many possible combinations of individual items. If an itemset is repeatedly purchased with the frequency not less than the minimal support, then it is marked as a frequent itemset [3].

In High Utility Itemset Mining the goal is to recognize itemsets that have utility values above a given utility threshold. The utility value of an itemset is the measurement of the importance of that itemset in the users perspective.For e.g. if a sales analyst involved in some retail research needs to find out which itemsets in the stores earn the maximum sales revenue for the stores he or she will define the utility of any itemset as the monetary profit that the store earns by selling each unit of that itemset. Here note that the sales analyst is not interested in the number of transactions that contain the itemset but he or she is only concerned about the revenue generated collectively by all the transactions containing the itemset. In practice the utility value of an itemset can be profit, popularity, page-rank, measure of some aesthetic aspect such as beauty or design or some other measures of user's preference.

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).

2. LITERATURE REVIEW

2.1 Two phase algorithm

This algorithm runs in two phases .This method maintains a Transaction-weighted Downward Closure Property [4]. In Phase I, we define transaction-weighted utilization and propose a model ─ transaction-weighted utilization mining. The transaction-weighted utilization of an itemset X, denoted as twu(X), is the sum of the transaction utilities of all the transactions containing X.This model maintains a Transaction-weighted Downward Closure Property. 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.

2.2 Compressed Transaction Utility (CTU-Mine)

CTU-Mine, This algorithm is suitable for dense dataset with long pattern. It Use pattern growth algorithm and also eliminates the expensive second phase of scanning the database. Limitation wa s Complex for evaluation due to the tree structure. It was stated in paper CTU-Mine: An Efficient High Utility Item set Mining Algorithm Using the Pattern Growth Approach [5] by Alva Erwin, Raj P. Gopalan, and N.R.Achuthan in 2007.

2.3 Transaction Weighted Utility (TWU)

Erwin et al., [7] proposed a transaction weighted utility

(TWU) algorithm which is based on compact utility pattern tree data structures. This work implements the parallel projection scheme to utilize the disk storage. This algorithm first identifies the TWU items from transaction database and the compressed utility pattern tree is constructed for mining complete set of high utility patterns. In this algorithm parallel projection is used to create subdivision for subsequently mining. This algorithm has anti-monotone property which is used to discover the pruning space. In this work the task of high utility itemset mining discovers all the utility which has utility higher than the user specified utility.

2.4 Fast Utility Mining (FUM)

Shankar et al., [8] proposed a fast utility mining (FUM) algorithm that finds all high utility itemset within the given utility constraint threshold. It is faster and simpler than the original UMining algorithm. This algorithm efficiently handles the duplicate itemsets. It checks whether a transaction defined by an itemset purchased in it, repeats its occurrence in a later transaction. If a later transaction also contains same itemset purchased in any of the previous transactions, then that transaction is ignored from processing and duplicate itemset are removed. This reduces the execution time of the algorithm further more. This algorithm provides absolute accuracy and proves to be extremely efficient in finding every possible high utility itemset from the transactions in the database. This algorithm executes transaction datasets exceptionally faster when more itemset are identified as high utility itemset and when the number of distinct items in the database increases.

2.5 Utility pattern growth (UP-growth)

To address issue of generating a large number of candidates, V.S Tseng et al. in 2010 proposed an a 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.

2.6 Utility Pattern Growth Plus (UP-Growth+)

Tseng et al., [9] proposed an efficient algorithm called as utility pattern growth plus (UP-Growth+) which is an improved version of utility pattern growth (UP-Growth) mining algorithm. In this work the information of high utility itemset is maintained in a special data structure named utility pattern tree (UP-Tree) and the candidate itemsets are generated with one scans of the database. The four strategies, applied in this algorithm are discarding global unpromising items (DGU), decreasing global node utilities (DGN), discarding local unpromising items (DLU), and decreasing local node utilities (DLN). By these strategies, the estimated utilities of candidates are well reduced, by discarding the utilities of the items which are impossible to be high utility or not involved in the search space. The proposed strategies not only decrease the estimated utilities of the potential high utility itemsets but also reduce the number of candidates. This algorithm outperforms substantially in terms of execution time, especially when the database contains lots of long transactions. However the operation time and search space of high-utility itemset mining can increase the high computation cost.

2.7 High Utility Itemset Miner (HUI-Miner)

Liu et al., [10] proposed a high utility itemset miner (HUI-Miner) for high utility itemset mining. This algorithm uses a novel structure called utility-list which is used to store both the utility information about an itemset and the heuristic information for pruning the search space. This algorithm first creates an initial utility list for itemsets of the length 1 for promising items. This algorithm constructs recursively a utility list for each itemset of the length k using a pair of utility lists for itemset of the length k-1 for mining high utility itemset, each utility list for an itemset keeps the information of indicates transaction 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 itemset of the itemset in the transactions. This algorithm first estimate the utilities of the itemsets and generate the candidate itemsets and then by scanning the database compute the exact utilities of the itemset to generate the high utility itemset. This algorithm mines the high utility itemset without generation of the candidates and the algorithm outperforms in terms of both running time and memory consumption.

2.8 Direct Discovery of High Utility Pattern (D2HUP)

Junqiang Liu et al., [11] proposed an algorithm direct discovery high utility pattern (D2HUP**)** which gains the combination of high utility pattern miner and utility pattern. This algorithm mines utility itemset in share framework.The direct discovery of high utility patterns, which is an integration of the depth-first search of the reverse set enumeration tree. This algorithm addresses the scalability and efficiency issues occurred in the existing systems as it directly extracts the high utility patterns from large transactional databases. This algorithm is based on the powerful pruning approaches. The look ahead strategy tries to find the patterns in recursive enumeration and it utilizes the singleton and closure property to enhance the efficiency of dense data. The linear data structure as chain of accurate utility list is used to show the original information of utility in the unrefined data. This work helps to discover the root causes of prior algorithm which employs to maintain data structure information of original utility.

INTERNATIONAL JOURNAL FOR ADVANCE RESEARCH IN ENGINEERING AND TECHNOLOGY

WINGS TO YOUR THOUGHTS.....

3. CONCLUSION

High utility pattern mining is an important development of data mining technology which addresses the limitation of frequent pattern mining by considering the user's expectation. A large number of algorithms have been proposed for utility pattern mining having their own advantages and disadvantages. This research paper presents an overview of various existing high utility itemset mining algorithms. The reviewed algorithms effectively mine high utility itemsets by using various data structures and constraint techniques. However there can be other algorithms to improve the performance and search space of high utility itemsets.

References

[1] A. Erwin, R.P. Gopalan, and N.R. Achuthan,"Efficient Mining of High Utility Itemsets from Large Datasets", T. Washio et al. (Eds.):PAKDD2008, LNAI 5012, pp. 554–561, 2008. © Springer-Verlag

[2] R. Agrawal , T. Imielinski, A. Swami, 1993," mining association rules between sets of items in large databases", in: proceedings of the ACM.

[3] 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.

[4] 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.

[5] T.B. Ho, D. Cheung, and H. Liu (Eds.): PAKDD 2005, LNAI 3518, pp. 689 – 695, 2005. © Springer-Verlag Berlin Heidelberg 2005.

[6] Utility Sudip Bhattacharya, Deepty Dubey," High Itemset Mining", International Journal of Emerging Technology and Advanced Engineering , Volume 2, Issue 8, August 2012.

[7] Erwin A., Gopalan R. P and. Achuthan N. R., "Efficient mining of high utility itemsets from large datasets," In Proceeding of the Pacific-Asia Conference on Knowledge Discovery and Data Mining, pp. 554–561, 2008.

[8] Shankar S., Purusothoman T.P, Jayanthi S., Babu N., "A fast agorithm for mining high utility itemsets" ,In roceedings of IEEE International Advance Computing Conference (IACC), Patiala, India, pp.1459-1464, 2009.

[9] Tseng V. S., Shie B.-E., Wu C.-W., and Yu P. S., "Efficient algorithms for mining high utility itemsets from ransactional databases," IEEE Transactions on Knowledge and Data Engineering, Vol. 25, No. 8, pp. 1772–1786, 2013.

[10] Liu M., Qu J., "Mining high utility itemsets without Candidate generation," Conference on Information and Knowledge Management. Association for Computing Machinery, pp. 55–64, 2012.

[11] Junqiang Liu., Ke Wang., Benjamin., Fung C.M.,"Mining High Utility Patterns in One Phase without Generating Candidates", IEEE Transactions on Knowledge and Data Engineering, Vol. 28, No. 5, pp.1–14, 2016.