EFFICIENT MINING TECHNIQUES FOR TRANSACTION HANDLING IN HIGH UTILITY ITEM SETS

Name: Miss.Saroj Fugare
ME Computer,
Computer Engineering Department
Mail-id: fugare.saroj@gmail.com

Guide: Dr.Prof.B.L.Gunjal
ME Computer,
Computer Engineering Department
Mail-id: Hello_baisa@yahoo.com

Abstract

Now a day's users are curious while purchasing any kind of product in market. Sometime time peoples may have confusion regarding product reliability among large set of products which has different brands, Sometimes, they decides specific criteria to purchase a product such as, product quality, quantity, other customers reviews etc. User may also define a threshold value for required product. In the research of data mining "HUI" i.e. high utility mining seems to be efficient for mining min_util itemeset from dataset with user specified threshold. However, HUI is very tedious and inefficient method in which min_util is set too low then utility are generated and the other hand if min_util is set too high then no utility will be found. Top-k high utility mining can be a better solution for such kind of problem due to its efficient scalability on huge datasets. There are two types of technique available namely, TKU and TKO for discovering top-k utility. In the real-world application, some of changes occurred in transactional database in case of insertion or deletion operation. Therefore, a technique is required to maintain and update the high utility itemsets in transactional database with the small number of itemset rescanning.

Keywords— Utility mining, high utility itemset, frequent itemset mining, top-k pattern mining

I.INTRODUCTION

Due to the ability of intelligent data analysis research in data mining increases day by day. It steps towards discovering knowledge which is crucial task in marking area. Discovering or extracting product information or their related information while purchasing any product is major concern at user end. Therefore, it is also necessary for product manager to preserve appropriate identity of their selling products in the market by which they can preserve CRM i.e. customer relationship management. HUI is mainly work for such kind of area. HUI is high utility mining outputs the products as per the occurrence of product transaction and quantity. Utility mining depicts item utility i.e. profit which then indicates the importance of an item. And number of occurrence of each item represents the transaction i.e. quantity. Itemset profit represents its importance such as, quantity, value and other related information as per user specification. Therefore, high utility mining is applicable in market analysis, steaming analysis and in mobile computing. However, HUI have downward closure property. Downward closure property is one in which each subset of frequent item set is also frequent known as apriority. In this case we don't have to find no. of item set. Pruning search is technique that reduces the size of decision tree by reducing small sections of tree which provides little power to classify instances. Problem with pruning search for high utility item sets may difficult because there may be chances that superset of low utility can be high utility. To overcome such problem TWU i.e. transaction weighted utilization model is suggested in [2]. It improves the mining performance. High transaction-weighted utilization itemset (HTWUI) is called as itemset only if TWU is lower than min_util and it representing upper bound on its utility. TWU contains two phases such as phase 1 in which complete set of HTWUIs are found and the other is phase 2 in which all high utility itemsets are obtained by calculating approximate utilities of HTWUIs with one database scan. Many algorithms used for HUI mining are resulting in inefficiency because user can choose approximate utility threshold which may produce very high or very small output. For exact value of min_util user have to guess correct threshold value and again have to re-execute algorithm till he satisfied with output result. However, it is inconvenient as well as time consuming procedure. An idea of depicting Top-k high utility itemset can be a promising solution in which user can specifies the number of itemset he required it is defined by k variable rather than specifying threshold value. It can work worth as user can specify number of itemset required by him instead of setting threshold value without knowing characteristics of database.

Using Top-k HUI mining user can get top-k sets of products which also increases profit of product company. There are certain kinds of challenges in top-k UI mining such as, top-k frequent pattern mining that rely on antimonotonicity to eliminate the search time cannot directly applied on top-k high utility itemset mining . Second thing is TWU model is widely used in utility mining

and TWU is difficult to adapt with top-k HUI mining due unexplored utility itemset in phase 1. Third one is search space can be efficiently removed as min_util threshold is provided in advance whereas in top-k min_util is provided in advanced. Therefore it is mandatory to design such system that can increment min_util threshold as high as possible and quickly as possible. Last challenge is to raise min_util_{Border} threshold without missing any top-k HUI result. TKO (mining Top-k utility itemset in one phase) and TKU (mining Top-k utility itemset) algorithm have complete set of top-k HUI's in the database without required to specifying min_util threshold value. TKU algorithm can maintain tree structure of transaction itemset utilities named as up-tree [25]. To increase border min. utility threshold five strategies can be used named as, PE, NU, MD, MC and SE in TKU algorithm whereas, in TKO, RCU, RUZ and EPB can be used for pruning or removing search space. We are going to represent the performance of TKU and TKO algorithm for utility mining algorithms UP-Growths [25] and HUI-Miner [14] tuned with optimal minimum utility threshold. Basically, utility mining approach is proposed to measure utility values of invested items for discovering HUI from static datasets. But in case of real world application there is dynamic updates may happen if any insertion or deletion operation take place. Therefore, transactional updates in real world datasets must be required.

II. RELATED WORK

In this section we are going to discussed related work about top-k high mining utility. There are certain kind of algorithms and techniques used for discovering itemset for large dataset.
They are explain as following:

A. Utility Mining

ARM i.e. association rule mining is the most popular technique is existed in data mining to discover frequent co-relation, co-occcurences, patterns from the large dataset. Whereas, utility mining used to disover all itemset having value above to user – specified threshold value. It tends to find all itemsets with high utility itemsets. Two phase algorithm is existed in [2] which identifies highly effectively exclude the candidate itemset and then make the calculation of utility easier. By doing such kind of task this algorithm reduces memory and computation cost. To reduce search space TWU concept is introduced by author to prune the search space. TWU represents the upper bound of it's utility and if it is lowern than the min_util then it is known as high transaction-weighted utilization itemset (HTWUI)[2]. Two-phase algorithm consist of two phases first one is set high potential itemsets are generated whereas in second phase, exact calculation of found candidate is done to find high utility itemsets. In single phase algorithm discovers the high utility itemsets in single phase and it does not produced any candidate. In[3], two algorithms of one-phase utility mining are discussed such as, d^2HUP and HUI miner. d^2HUP algorithm , horizontal database is converted into tree-based structure,this process is called as CAUL[3]. Due to this process pattern growth stratergy can directly finds high utility itemset from large dataset. In[4], HUI-miner cansider vertical database whch can be transformed into list of utility. HUI miner directly computes database in RAM without scanning original copy of database. These all techniques are used many applications but they are not applicable for top-k high utility itemset searching which required to set approximate threshold value.

B. Pattern Mining

Lots of study have been conctucted onmining top-k results form large dataset. Pattern mining is the way identify efficient top-k frequent results. MTK is memory constraint Top-k mining algorithm. It recognises the concept of e δ-Stair search[5]. MTK applies hash pruning algorithm to reduce unneccesary candidates by using hash table sxtructure. MTK/MTK_closed algorithm are arranged in level wise search and it is based on number candidates that will be generated and tested in every dataset scan. This approach is not flexible in real and synthetic data as it is only striking compromise between memory consumption and execution efficiency. Another is seq-BOMA approach is the combination between seq-Exminer and build once mine anytime. It having many beneficial features for the real application[6]. Seq-exminer algorithm combines top-k frequent patterns linearly. The concept of composite pattern detection is introduced in[7] to overcome the problem of enormous pattern combinations while extracting frequent patterns with combination of single path. CRM is the combination reducimg method used in top-k pattern mining and CRMN is used for N-number of itemsets. It performs outstanding in terms of memory consumption scalibility etc. But this technique cannot work on weight based mining, stream data mining and other applications like,web,bio etc. FP-tree is build for pattern-growth search.

C. Top-k High Utility Pattern Mining

Mining top-k high utility itemsets from large database is the task of extracting items having high profit. In[8], UP-Growth and UP-Growth+ algorithms is disccussed by authors Bai-En Shie ; Cheng-Wei Wu ; Philip S. Yu. This algorithm sets effective methodolgy for pruning candidate itemsets. In this algorithm itemsets are arranged in tree format therefore, it is known as, UP-Tree. It can generate, PHUI efficiently is two scans. Up-Growth outperforms the state-of-the art algorithm when there is need of minimum utility threshold. In[9], TKU algorithms is proposed to obtain top-k high utility itemsets. In this technique both profit and quantities are considered for mining high top-k high utility. T-HUDS is an efficient mining algorithm over large data stream suggested by Zihayat in[10]. In this the major challenge of devising several stratergy for initializing and raising minimum utility threshold is solved. Raising threshold with Exact and Pre-calculated utilities

for Top-k high utility pattern mining (REPT) in disscuessed by H. Ryang and U. Yun in[11]. REPT is proposed to incerease minimum utility threshold in top-k high utility pattern mining. In REPT algorithm four stratergies are implemented known as, PUD, RIU,RSD and SEP. This stratergies helps to reduce search space increasing min. threshold value in mining procedure. There three steps are conducted in REPT algorithm first is to scan original databse twice to build global tree, then generate high utility candidate patterns and finally, to identify top-k high utility patterns from the candidates.

D. TKU

TKU is Top-k utility mining algorithm. This algorithm does not required specify min_util for extraction of high utility itemset from large dataset. TKU gains the compact tree based structure, it is known as UP-tree[13] to preserve the information of transactions and utility itemsets. TKU derives some useful properties of TWU. TKU also consist of two phases such as, in first phase PKHUIs i.e. potential top-k high utility itemsets are generated and in another phase top-k itemsets with high utility are extracted from generated PKHUI. Other than this, four stratergies are developed to minimize the estimated utility value which affects to improve the performance of utility mining.

TKO is mining top-k utility in one phase algorithm. It uses list based approach known as utility-list to preserve utility information of itemsets. To increases the effectiveness of raise border threshold five stratergies are implemented as, PE,NU,MD,MC and SE. They raise border minimum utility threshold. Also stratergies named as, RUZ,TUc and EPB are merged together for reducing search space in TKO.

Stratergy 1. RUC is Raising the threshold byy Utility of Candidates. This stratergy is concerned with any kind of one phase algorithm which have itemset with their utility. It refers TopK-CI-List format for maintaining TopK HUIs.

Stratergy 2. RUZ is Reducing estimated utility values by using Z-element. This stratrgy can be applied during candidate itemset extraction in TopK-HUI search procedure. It also helps to reduce search space.

Stratergy 3. EPB is Exploring most promising Branches Frist. EPB generates the candidate itemset having highest utility first. Itemset with higher utility is found then TKO can raise $min\_util_{Border.}$ There space pruning can also takes place automatically.

In TKU_{Base} three steps are executed such as, 1. Constructing UP-Tree 2. From UP-tree generate potential top-k high utility itemsets. And 3. To discovered top-k HUI from the set of PKHUIs.

E. Frequent Itemset Mining

Mining of frequent itemset is discussed in [14]. Frequnt pattern mining gives the most frquent itemsets in dataset therefore it can increase the processing speed of discovering itemset from large scale datasets. Inrequent mining procedure association of rule mining stratergy is involved. Algorithms such as, apriori and aprioriTid is used to discover significant association rules from large transaction database. Previously, algorithms known as, AIs and SETM are developed for frquent pattern mining which increases the search gap and hence faces size problems. Moreover , factor opf three problems ranged more than oredr of mangnitude then it can be large problem. FP-treee is suggested in[15], to discover frequent patterns in large database. FP-Tree is frequent pattern tree, usually it is used to store compressed and crucial information of frequent patterns. Using this information pattern growth method is developed. FP-Tree is smaller than original database. Developed pattern growth method avoids high cost candidate generation by successively merging 1-itemset in the conditional FP-tree.

In this the procedure of mining utility is not restricted for candidate generation and test but only for the growth of frequent patterns i.e. fragment.

III. CONCLUSION

In this review paper we have studied various approaches of utility mining. Utility mining is concept of discovering itemsets having beyond value than the user specified threshold value. In this review process, we have identified some problems with existing system such as, previously, user have to specify threshold value to get required itemset. However, it is inconvenient task to specify approximate thresholod value because minimum threshold value may output more utility or it may happen that maximum threshold value returns zero utility. Other problem is noticed in reducing search space as pruning search for high utility itemsets is difficult because there may be chances that superset of low utility can be high utility. Also existing system only provides high utility itemset mining which may creates confusion at user side about product itemsets. From this overall literature review we analyze that top-k utility mining can be interesting investigation area to discover top-k itemset with pruned search space and also there is a need of such system that can handle transactions in database with small number of rescan.

VI.REFERENCES

1 Vincent S. Tseng, Senior Member, Cheng-Wei Wu, Philippe Fournier-Viger, Philip S. Yu, Fellow, IEEE," Efficient Algorithms for Mining Top-K High Utility Itemsets", IEEE tansaction on knowledge and data engineering.

2 Y. Liu, W. Liao, and A. Choudhary, ―A Fast High Utility Itemsets Mining Algorithm,‖ in Proc. of the Utility-Based Data Mining Workshop, pp. 90-99, 2005.

3 M. Liu and J. Qu, ―Mining High Utility Itemsets without Candidate Generation,‖ in Proc. of ACM Int'l Conf. on Information and Knowledge Management, pp. 55-64, 2012.

4 J. Liu, K. Wang and B. Fung, ―Direct Discovery of High Utility Itemsets without Candidate Generation,‖ in Proc. of IEEE Int'l Conf. on Data Mining, pp. 984-989 , 2012.

5 K. Chuang, J. Huang and M. Chen, ―Mining Top-K Frequent Patterns in the Presence of the Memory Constraint,‖ The VLDB Journal, Vol. 17, pp. 1321-1344, 2008.

6 R. Chan, Q. Yang and Y. Shen, ―Mining High-utility Itemsets,‖ in Proc. of IEEE Int'l Conf. on Data Mining, pp. 19-26,2003.

7 Pyun and U. Yun, ―Mining Top-K Frequent Patterns with Combination Reducing Techniques, ―Applied Intelligence, Vol. 41(1), pp. 76-98, 2014.

8 T. Quang, S. Oyanagi, and K. Yamazaki, ― ExMiner: An Efficient Algorithm for Mining Top-K Frequent Patterns,‖ in Proc. of Int'l Conf. on Advanced Data Mining and Applications,pp. 436 –447, 2006.

9 Vincent S. Tseng, Bai-En Shie, Cheng-Wei Wu, and Philip S. Yu, Fellow, IEEE," Efficient Algorithms for Mining High Utility Itemsets from Transactional Databases"

10 C. Wu, B. Shie, V. S. Tseng and P. S. Yu, ―Mining Top-K High Utility Itemsets,‖ in Proc. ofthe ACM SIGKDDInt'l Conf. onKnowledge Discovery and Data Mining, pp. 78– 86, 2012.

11 M. Zihayat and A. An, ―Mining Top-K High Utility Itemsets over Data Streams,‖ Information Sciences, Vol.285 (20), pp.138–161, 2014.

12 Ryang and U. Yun, ―Top-K High Utility Pattern Mining with Effective Threshold Raising Strategies,‖ Knowledge-Based Systems, Vol. 76, pp. 109-126, 2015.

13 V. S. Tseng, C. Wu, B. Shie, and P. S. Yu, ―UP-Growth: An Efficient Algorithm for High Utility Itemset Mining,‖ in Proc. of the ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining, pp. 253–262, 2010.

14 R. Agrawal and R. Srikant, ―Fast Algorithms for Mining Association Rules,‖ in Proc. of Int'l Conf. on Very Large Data Bases, pp. 487-499, 1994.

15 J. Han, J. Pei and Y. Yin, ―Mining Frequent Patterns without Candidate Generation,‖ in Proc. of ACMSIGMOD Int'l Conf. on Management of Data, pp. 1-12, 2000.