Survey on Efficient Algorithms for Mining Top-K High Utility Itemsets

Namrata Vardhaman
M.Tech Scholar
Department of CSE
AMCEC, Bengaluru, India

Prof. Doddegowda B J
Associate Professor
Department of CSE
AMCEC, Bengaluru, India
bjdgowda10@gmail.com

***Abstract*—High utility itemsets (HUIs) mining is a booming topic in data mining. HUIsdeals with discovering all the itemsets that has a utility meeting specified by user a minimum utility threshold called min_util. Setting min_util appropriately is a difficult task for the users. Generally finding an appropriate min_util threshold by trial and error method is a tedious task for the users. If the min_util is set too low, too many HUIs will be generated, which may cause mining process to be inefficient. On the other hand, if min_util is set too high, it is likely that no HUIs will be found. In this survey paper we address the above issues by proposing a new methodology for top-k high utility itemset mining, where k is the desired number of HUIs to be mined. Two efficient algorithms TKO (mining Top-K utility itemsets in One phase) and TKU (mining Top-K Utility itemsets) are proposed for mining such itemsets without the setting the minimum utility threshold min_util. A structural comparison of the two algorithms with discussions on their advantages and limitations is provided in this paper. Empirical evaluations on both realand synthetic datasets prove that the performance of the proposed efficient algorithms is close to that of optimal case of state-of-the-art utility mining algorithms.**

Keywords*— *Utility mining, high utility itemset mining, top-K pattern mining, top-K high utility itemset mining.

I. INTRODUCTION

FREQUENT item set mining (FIM) is a fundamental research topic in datamining. The traditional FIM discovers a large amount of frequent but low-value item sets and lose the information on the valuable item sets that has low selling frequencies. Hence, it cannot satisfy the requirement of users who desire to discover item sets with high utilities such as high profits. To address these issues, utility mining emerges as an valuable topic in data mining and has grabbed attention in recent years. In utility mining, each item is associated with a utility called profit and an occurrence count i.e quantity in each transaction. The utility of an item set represents its importance, which can be measured in terms of weight, value, quantity or other information depending on the user specification. An item set is called high utility item set (HUI) when its utility is no less than a minimum utility threshold min_util specified by the user. HUI mining is essential to many applications such as streaming analysis market analysis mobilecomputing and biomedicine. Efficient mining HUIs in databases is not an easy task because the downward closure property used in FIM does not hold for the utility of item sets. Pruning the search space for High Utility Itemset mining is difficult because a superset of a low utility item set can be high utility. To deal with this issue, the concept of transaction weighted utilization (TWU) model [13] was introduced to facilitate the performance of the mining task. In this model, an item set is called high transaction-weighted utilization item set (HTWUI) if its TWU is no less than min_util threshold, where the TWU of an item set represents an upper bound on its utility.Therefore, a HUI must be a HTWUI and all the HUIs must be included in the complete set of HTWUIs. TWU model-based algorithm consists of two phases. In the first phase, called phase I, the complete set of HTWUIs are found. In the second phase, called phase II, all HUIs are obtained by calculating the exact utilities of HTWUIs withone database scan. Many studies are carried out in HUI mining, it is difficult for users to choose an appropriate minimum utility threshold in practice. Depending on the threshold, the output size can be small or large. The choice of the threshold greatly influences the performance of the algorithms. If the threshold is set too low, too many High Utility Itemsets will be presented to theusers and it is difficult for the users to comprehend the results. A large number of HUIs also causes the mining algorithms to become inefficient or even run out of memory, because the more HUIs the algorithms generate, the more resources are consumed. On the other hand, if the threshold is set too high, no HUI will be found. To find an appropriate value for the min_util threshold, users need to try different thresholds by trail and error methods and re-executing the algorithms over and over until satisfying results are achieved. This process is both inconvenient and time-consuming.TKO utilizes minimal node for further decreasing over-estimated utilities of item sets. Even though it spends memory and time to check and store minimal node utilities, they are more effective especially when there are many longer transactions in databases. In contrast, UP-Growth performs better only when min_util is UP-Growth outperforms UP-Growth although they have tradeoffs on memory usage. The reason is that UP-Growth+ utilizes minimal node utilities for further decreasing overestimated utilities of item sets. Even though time and memory is used to check and store minimal node utilities, they are more effective especially when there are many longer transactions in databases. In contrast, UP-Growth performs better only when min_util is small. This is because

when number of candidates of the two algorithms is similar, UP-Growth+ carries more computations and is thus slower. Finally, high utility item sets are efficiently identified from the set of PHUIs which is much smaller than HTWUIs generated by IHUP. By the reasons mentioned above, the proposed algorithms UP-Growth and UP-Growth+ achieve better performance than IHUP algorithm.til is small. This is because when number of candidates of the two algorithms is similar, UP-Growth+ carries more computations and is thus slower. Finally, high utility item sets are efficiently identified from the set of PHUIs which is much smaller than HTWUIs generated by IHUP. By the reasons mentioned above, the proposed algorithms UP-Growth and UP-Growth +achieve better performance than IHUP algorithm.

II. RELATED WORK

This section introduces related works about top-k high utility itemset mining, including high utility itemset mining, top-k frequent pattern mining and top-k high utility itemset mining.

A. High Utility Itemset Mining

High utility itemset mining has received lots of attention and many algorithms have been proposed, such as IHUP [2], Two-Phase [13], UP-Growth,IIDS [17], HUI-Miner [14] and d^{2}HUP [15]. These algorithms can be generally categorized into two types: two-phase and one-phase algorithms. In two-phase algorithms consists of two phases. In the first phase, a set of candidates that are potential high utility itemsets are generated. In the second phase, the exact utility of each candidate found in the first phase to identify high utility itemsets is calculated. Two-Phase, IHUP, IIDS and UP-Growth are two-phase based algorithms.

The main feature of one-phase algorithms is that they discover high utility itemsets using only one phase and produce no candidates. d^{2}HUP and HUI-Miner are one-phase algorithms. Above studies may perform well in some applications, they are not developed for top-k high utility itemset mining and still suffer from the subtle problem of setting appropriate thresholds.

B. Top-k Pattern Mining

Multiple studies have been proposed to mine different kinds of top-k patterns, such as top-k frequent itemsets [3], [19], top-k frequent closed itemsets [3], top-k closed sequential patterns, top-k association rules [6], top-k sequential rules [5], top-k correlation patternsand top-k cosine similarity interesting pairs. Each top-k pattern mining algorithm is distinguished based on the type of patterns discovered, as well as the data structures and search strategies that are employed. For example, some algorithms [5], [6] use a rule expansion strategy for finding patterns, while others rely on a pattern-growth search using structures such as FP-Tree. The choice of data structures and search strategy affect the efficiency of a top-k pattern mining algorithm in terms of both memory and execution time. However, the above algorithms discover top-k patterns in a traditional method instead of the utility method. As a result, they may miss patterns yielding high utility.

C. Top-k High Utility Pattern Mining

The task of top-k high utility pattern mining was introduced by Chan et al. [4]. But the definition of high utility itemset used in their study is different from the one used in this work. Chan et al.'s study has considered utilities of various items, but quantitative values of items in transactions were not taken into consideration. The top-k high utility itemset mining is defined by considering both quantities and profits of items. This work has inspired a few studies for mining top-k high utility patterns. Zihayat and An [21] have proposed an efficient algorithm T-HUDS formining top-k HUIs over data streams.Recently, Ryang and Yun extended [20] to propose the REPT algorithm for top-k HUI mining.

III EXISTING SYSTEM APPROACH

FREQUENT item set mining (FIM) is a fundamental research topic in data mining. The traditional FIM discovers a large amount of frequent but low-value item sets and lose the information on the valuable item sets that has low selling frequencies. Hence, it cannot satisfy the requirement of users who desire to discover item sets with high utilities such as high profits. To address these issues, utility mining emerges as an valuable topic in data mining and has grabbed attention in recent years. In utility mining, each item is associated with a utility called profit and an occurrence count i.e quantity in each transaction. The utility of an item set represents its importance, which can be measured in terms of weight, value, quantity or other information depending on the user specification. An item set is called high utility item set (HUI) when its utility is no less than a minimum utility threshold min_util specified by the user. HUI mining is essential to many applications such as streaming analysis market analysis mobile computing and biomedicine.

Disadvantages Of Existing System:-

1. Efficiently mining High Utility Itemsets (HUIs) in databases is not an easy task because the downward closure property used in FIM does not hold for the utility of item sets.

2. In other words, pruning search space for High Utility Itemset mining is difficult because a superset of a low utility item set can be high utility.

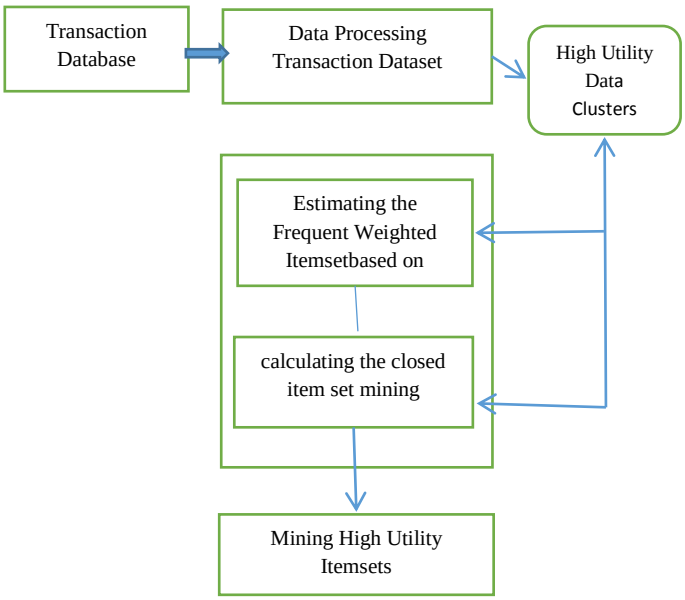
IV PROPOSED SYSTEM

The concept of transaction weighted utilization (TWU) model was introduced to facilitate the performance of the mining task. In this proposed methodolgy, an item set is called high transaction-weighted utilization item set (HTWUI) if its TWU is not less than minimum utility threshold min_util, where the TWU of an item set represents an upper bound on its utility. Hence, a High Utility Itemset (HUI) must be a high transaction-weighted utilization item set (HTWUI) and all the HUIs must be included in the complete set of HTWUIs. A classical TWU model-based algorithm consists of 2 phases. In the first phase, called phase I, the complete set of HTWUIs are found. In the phase II, all HUIs are obtained by calculating the exact utilities of HTWUIs with one database scan.

Advantages Of Proposed System

1. Two efficient algorithms called TKO (mining Top-K utility item sets in one phase) and TKU (mining Top-K Utility itemsets) are proposed for mining the complete set of top-k HUIs in the databases without specifying the minimum utility threshold min_util.
2. The construction of the UP-Tree and pruning the unpromising items in the transactions, the number of nodes that would be maintained in the memory could be reduced and the mining algorithm could achieve better efficiency and performance.

System Architecture



V CONCLUSION

In this paper, we have studied the problem of top-k high utility item sets mining, where k is the desired number of high utility item sets that are to be mined. Two efficient algorithms TKU (mining Top-K Utility item sets) and TKO (mining Top-K utility item sets in One phase) are proposed for mining such item sets without setting minimum utility thresholds min_util. TKU is the first two-phase algorithm for miningTop-k high utility item sets, has five strategies PE, NU, MD, MC and SE to effectively raise the border minimum utility thresholds and further prunethe search space. On the other hand, TKO is the first one-phase algorithm developed for top-k HUI mining, which integrates the strategies RUC, RUZ and EPB to improve its performance. Empirical evaluations on different types of real and synthetic datasets show that the proposed algorithms have good scalability on large datasets and the performance of the proposed algorithms is close to the optimal case of the state-of-threat two-phase and one-phase utility mining algorithms.

REFERENCES

[1] R. Agrawal and R. Srikant, "Fast algorithms for mining associationrules," in Proc. Int. Conf. Very Large Data Bases, 1994, pp. 487–499.
[2] C. Ahmed, S. Tanbeer, B. Jeong, and Y. Lee, "Efficient tree structuresfor high-utility pattern mining in incremental databases,"IEEE Trans. Knowl. Data Eng., vol. 21, no. 12, pp. 1708–1721, Dec.2009.
[3] K. Chuang, J. Huang, and M. Chen, "Mining top-k frequent patternsin the presence of the memory constraint," VLDB J., vol. 17,pp. 1321–1344, 2008.
[4] R. Chan, Q. Yang, and Y. Shen, "Mining high-utility item sets," inProc. IEEE Int. Conf. Data Mining, 2003, pp. 19–26.
[5] P. Fournier-Viger and V. S. Tseng, "Mining top-k sequentialrules," in Proc. Int. Conf. Adv. Data Mining Appl., 2011, pp. 180–194.
[6] P. Fournier-Viger, C.Wu, and V. S. Tseng, "Mining top-k associationrules," in Proc. Int. Conf. Can. Conf. Adv. Artif. Intell., 2012, pp. 61–73.
[7] P. Fournier-Viger, C. Wu, and V. S. Tseng, "Novel concise representationsof high utility item sets using generator patterns," inProc. Int. Conf. Adv. Data Mining Appl. Lecture Notes Comput. Sci.,2014, vol. 8933, pp. 30–43.
[8] J. Han, J. Pei, and Y. Yin, "Mining frequent patterns without candidategeneration," in Proc. ACM SIGMOD Int. Conf. Manag. Data,2000, pp. 1–12.
[9] J. Han, J. Wang, Y. Lu, and P. Tzvetkov, "Mining top-k frequentclosed patterns without minimum support," in Proc. IEEE Int.Conf. Data Mining, 2002, pp. 211–218.
[10] S. Krishnamurthy, "Pruning strategies for mining high utilityitem sets," Expert Syst. Appl., vol. 42, no. 5, pp. 2371–2381, 2015.
[11] C. Lin, T. Hong, G. Lan, J. Wong, and W. Lin, "Efficient updatingof discovered high-utility item sets for transaction deletion indynamic databases," Adv. Eng. Informat., vol. 29, no. 1, pp. 16–27,2015.
[12] G. Lan, T. Hong, V. S. Tseng, and S. Wang, "Applying the maximumutility measure in high utility sequential pattern mining,"Expert Syst. Appl., vol. 41, no. 11, pp. 5071–5081, 2014.
[13] Y. Liu, W. Liao, and A. Choudhary, "A fast high utility item setsmining algorithm," in Proc. Utility-Based Data Mining Workshop,2005, pp. 90–99.
[14] M. Liu and J. Qu, "Mining high utility item sets without candidategeneration," in Proc. ACM Int. Conf. Inf. Knowl. Manag., 2012,pp. 55–64.
[15] J. Liu, K. Wang, and B. Fung, "Direct discovery of high utilityitem sets without candidate generation," in Proc. IEEE Int. Conf.Data Mining, 2012, pp. 984–989.
[16] Y. Lin, C. Wu, and V. S. Tseng, "Mining high utility item sets in bigdata," in Proc. Int. Conf. Pacific-Asia Conf. Knowl. Discovery DataMining, 2015, pp. 649–661.
[17] Y. Li, J. Yeh, and C. Chang, "Isolated items discarding strategy fordiscovering high-utility item sets," Data Knowl. Eng., vol. 64, no. 1,pp. 198– 217, 2008.
[18] G. Pyun and U. Yun, "Mining top-k frequent patterns with combinationreducing techniques," Appl. Intell., vol. 41, no. 1, pp. 76–98,2014.
[19] T. Quang, S. Oyanagi, and K. Yamazaki, "ExMiner: An efficientalgorithm for mining top-k frequent patterns," in Proc. Int. Conf.Adv. Data Mining Appl., 2006, pp. 436 – 447.
[20] H. Ryang and U. Yun, ―Top-K High Utility Pattern Mining with Effective Threshold Raising Strategies,‖ Knowledge-BasedSystems, Vol. 76, pp. 109-126, 2015.
[21] M. Zihayat And A. An, "Mining top-k High Utility Itemsets over data streams", Inf Sci, Vol 285,no. 1, pp.60-83,2011.