|| Volume 4 || Issue 10 || OCTOBER 2019 || ISO 3297:2007 Certified ISSN (Online) 2456-3293

OPEN ACCESS INTERNATIONAL JOURNAL OF SCIENCE & ENGINEERING

IMPLEMENTATION OF PRE-EVALUATION STRATEGY ON ALGORITHMS FOR MINING TOP K HIGH UTILITY ITEM SETS

Nikhil M Jahagirdar[1],V S Karvande[2]

P.G. Student, Computer Science & Engineering, Everest Educational Society's Group of Institutions, Maharashtra, India[1]

HOD, Computer Science & Engineering, Everest Educational Society's Group of Institutions, Maharashtra, India[2]

Abstract: High utility item sets (HUIs) mining is a rising subject in information mining, which aims to finding all item sets having an utility meeting a client determined least utility edge min util. But, setting min util suitably is a troublesome issue for clients. As a rule, finding a fitting least utility edge by experimentation is a monotonous procedure for clients. In the event that min util is set too low, an excessive number of HUI swill be produced, which may bring about the mining procedure to be exceptionally wasteful. Then again, if min util is set too high, it is likely that no HUIs will be found. We address the above issues by redefining the problem of high utility item sets (HUIs) mining by top-k high utility item sets (top-k HUI) mining, where k is the desired number of HUIs to be mined. Two algorithms named TKU (mining Top-K Utility item sets in two stages) and TKO (mining Top-K utility item sets in one stage) are proposed for mining such item sets without the need to set min util. To improve the performance, we apply pre-evaluation strategy to algorithms.

Keywords: HUI, 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 data mining. However, the traditional FIM may discover a large amount of frequent but low-value item sets and lose the information on valuable item sets having low selling frequencies. Hence, it can’t satisfy the requirement of users who desire to discover item sets with high utilities such as high profits. An item set is said to be frequent if its support is no less than a given minimum support threshold. Hundreds of studies have been conducted on this topic. However, an important limitation of FIM is its assumptions that all items have the same importance to the user (e.g., unit profit or weight) and that items may not appear (quantity) more than once in each transaction. These assumptions often do not hold in real life. For example, in transaction databases, items may have different unit profits, and items in transactions may be associated with different purchase quantities. Besides, in real-life applications, retailers may be more interested in finding item sets that yield a high profit rather than discovering frequent item sets. Utility mining, which refers to the discovery of item sets with utilities higher than a user specified minimum utility threshold, is an important task and has a widerange of applications, especially in e-commerce. But setting an appropriate minimum utility threshold is a difficult problem. If the minimum threshold is set too low, too many high utility item sets will be generated and it takes a long time to compute,while setting the minimum threshold too high would result in too few results. Setting appropriate minimum utility threshold by trial and error is not very efficient. To precisely control the output size and discover the item sets with the highest utilities without setting the thresholds, a promising solution is to redefine the task of mining HUIs as mining top-k high utility item sets (top-k HUIs). The idea is to let the users specify k, i.e., the number of desired item sets, instead of specifying the minimum utility threshold. Setting k is more comfortable than setting the threshold because k represents the number of item sets that the users want to find An efficient algorithm named TKU (Top-K Utility item sets mining) and TKO (mining Top-K utility item sets in one stage) are proposed for mining such item sets without the need to set min util.

Motivation

To analyze customer purchase behavior, top-k HUI mining serves as a promising solution for users who desire to know “What are the top-k sets of products (i.e. item sets) that contribute the highest profits to the company?”. Owners of General stores, medical stores, hotels, and supermarket can also take the advantage to find top k high utility item sets (products) for business purpose. It can also be used in e-commerce.

II LITERATURE SURVEY

Top-k high utility itemset mining refers to the discovery of top-k patterns using a user-specified value k by considering the utility of items in a transactional database. Since existing top-k high utility itemset mining algorithms are based on the patterngrowth method, they search the patterns in two steps. Therefore, the generation of many candidates and additional database scan for calculating exact utilities are unavoidable.

In this paper, we propose a new algorithm, TKUL-Miner, to mine top-khigh utility itemsets efficiently. It utilizes a new utility-list structure which stores necessary information at each node on the search tree for mining the itemsets. The proposed algorithm has a strategy using search order for specific region to raise the border minimum utility threshold rapidly. Moreover, two additional strategies for calculating smaller overestimated utilities are suggested to prune unpromising itemsets effectively. Experimental results on various datasets showed that the TKULMiner outperforms other recent algorithms both in runtime and memory efficiency. [1]

High-utility itemset mining (HUIM) has been gaining popularity in the field ofdata mining. Frequent itemset mining (FIM) used to be the main tool to reveal high-frequency patterns but failed to consider the concept of profit. HUIM, on the other hand, obtains the itemsets and is practical in commercial applications. A main challenge in HUIM is that HUIM should handle the exponential search space for HUIM when the number of distinct items and the size of the database are both too large. The other challenge is that existing HUIM methods overlook the length ofhigh-utility itemsets; hence, a large itemset gets an unreasonable estimated profit as opposed to the actual value. Therefore, several algorithms were proposed to mine high average-utility itemsets. High average-utility itemset mining (HAUIM) is an extension for traditional HUIM, which provides a different measure with HUIM. It discovers utility patterns by considering both their utilities and lengths. To reduce the searching space in HAUIM, average-utility upper-bound (auub), looser upper-bound utility (lub), and a revised tighter upper-bound model (rtub) are proposed to prunethe searching graph in HAUIM. These three upper-bounds for high average-utility itemsets decrease the number of candidate patterns efficiently. However, they stilloverestimate a high average-utility itemset and waste on assessing the unnecessary patterns. Two new tighter upper-bounds, maximum following utility upper-bound(mfuub) and top-k transaction-maximum utility upper-bound (krtmuub), are proposedin this article to further contract the size of candidate pattern set. Experiments conducted on several benchmark datasets show the proposed method outperforms the previous HAUIM algorithms in terms of runtime, number of join operations and scalability. [2]

Data mining uses various algorithms for searching interesting information and hidden patterns from the large database. Traditional frequent itemset mining (FIM)generate large amount of frequent itemset without considering the quantity and profit of item purchased. High utility itemset mining (HUIM) gives advantageous resultsas compared to the frequent itemset mining. HUIM algorithm helps to improve the performance of finding data by considering both quantity and profit of itemset from large database. This paper reviews two types of efficient algorithm named TKU (mining top-k utility itemset) and TKO (mining top-k utility itemsets in One phase) for mining high utility itemset without any need to set minimum utility threshold by using strategy of UP-tree data structure which scans the database twice and enhances the efficiency of mining High utility itemset. It find out transaction utility of each transaction and it also compute TWU of each item. Then it reorganizes the transaction and constructs the Up Tree. [3]

Mining high utility itemsets from a transactional database refers to the discovery of itemsets with high utility like profits. Although a number of relevant approaches have been proposed in recent years, but they incur the problem of producing a large number of candidate itemsets for high utility itemsets. Such a large number of candidate itemsets degrades the mining performance in terms of execution time and space requirement. The situation may become worse when the database contains lots of long transactions or long high utility itemsets. An emerging topic in the field of data mining is utility mining which not only considers the frequency of the itemsets but also considers the utility associated with the itemsets. The main objective of High Utility Itemset Mining is to identify itemsets that have utility values above a given utility threshold. Thus Utility mining plays an important role in many realtime applications and is an important research topic in data mining system to find the itemsets with high profit. In this paper we present the implementation of first module where pre-processing of dataset is done to remove unpromising data from web usage and product base dataset by using TopKRules and also we are proposing a new framework for Top-k high utility web access patterns, where k is the desired number of HUIs to be mined. Two types of efficient algorithms named TKU and TKO are proposed for mining such itemsets. In this paper we present a literature review of the present state of research and the various algorithms for high utility itemset mining. [4]

Mining high utility patterns, the subject of which has attracted many researchers in data mining, is the process of discovering patterns with utility satisfying a minimum predetermined threshold. Many studies have been performed, but finding the suitable minimum utility threshold is problematic, because users cannot predict the appropriate threshold that affects mining performance. To solve this problem, mining the highest utility of k patterns, called top-k high utility, has been proposed. Although many approaches have been proposed, the issue of many candidates and the performance of mining needed to be further studied. In this paper, we propose a top-k high utility mining method that does not produce a candidate with an effective threshold-raising strategy. Instead, the proposed method uses a utility-list data structure with improved threshold-raising strategies combined with an efficient pruning strategy. Experimental results on real and synthesis datasets show that the algorithm presented performs better than current methods. [5]

III PROBLEM STATEMENT

Given a transnational database and the value of k, the desired number of HUIs(High Utility Item Sets), the problem of top-k high utility item sets mining is to find the top k high utility item sets by internally raising the value of min util.

Scope

To analyze customer purchase behavior, top-k HUI mining serves as a promising solution for users who desire to know "What are the top-k sets of products (i.e. item sets) that contribute the highest profits to the company?". Owners of General stores, medical stores, hotels, and supermarket can also take the advantage to find top k high utility item sets (products) for business purpose. It can also be used in e-commerce.

IV SYSTEM DESIGN

The TKU algorithm consists of two phases. In the first phase, the potential top k high utility item sets (PTKHUIs) are found using the UP tree. In the second phase, top-k HUIs are obtained by calculating the exact utilities of PTKHUIs with one database scan. In TKO, directly the top-k HUIs are found and updated repeatedly using the utility list structure. It do not require to scan the database once the utility list is constructed.

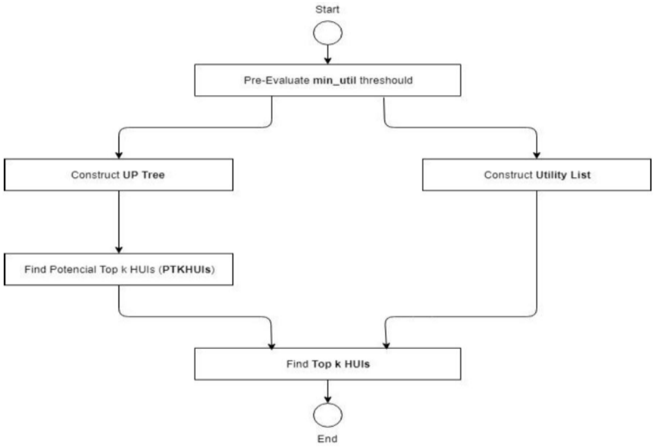

Start
Pre-Evaluate min_util threshould
Construct UP Tree
Construct Utility List
Find Potencial Top k HUIs (PTKHUIs)
Find Top k HUIs
End

Figure 1: Architecture Diagram

Module Descriptions

1 Pre-Evaluation of min util

Though TKU provides a way to mine top-k HUIs, min util is set to 0 before the construction of the UP-Tree. This results in the construction of a full UP-Tree in memory, which degrades the performance of the mining task. If min util could be raised before the construction of the UP-Tree and prune more unpromising items in transactions, the number of nodes maintained in memory could be reduced and the mining algorithm could achieve better performance. Based on this idea, we propose a strategy named PE (Pre-evaluation Step) to raise min util during the first scan of the database.

2 Construction of UP-Tree

A UP-Tree can be constructed by scanning the original database twice. Inthe first scan, the transaction utility of each transaction and TWU of each item are computed. During the second database scan, transactions are reorganized and then inserted into the UP-Tree.

3 Generating PTKHUIs

TKU algorithm uses UP tree to generate the potential top k high utility item sets. Parallelly, it raises the value of minimum utility threshold dynamically during the generation of PTKHUIs.

4 Identifying Top-k HUIs from PTKHUIs

After identifying PTKHUIs, TKU calculates the exact utility of PTKHUI sby scanning the original database once again, to identify the top-k HUIs.

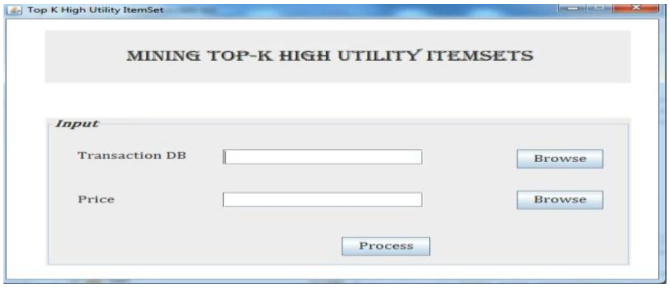
5 Construction of Utility List Structure

In the TKO algorithm, each item (set) is associated with a utility-list. The utility-lists of items are called initial utility-lists, which can be constructed by scanning the database twice. In the first database scan, the TWU and utility values of items are calculated. During the second database scan, items in each transaction are sorted in order of TWU values and the utility-list of each item is constructed.

6 Finding Top-k HUIs

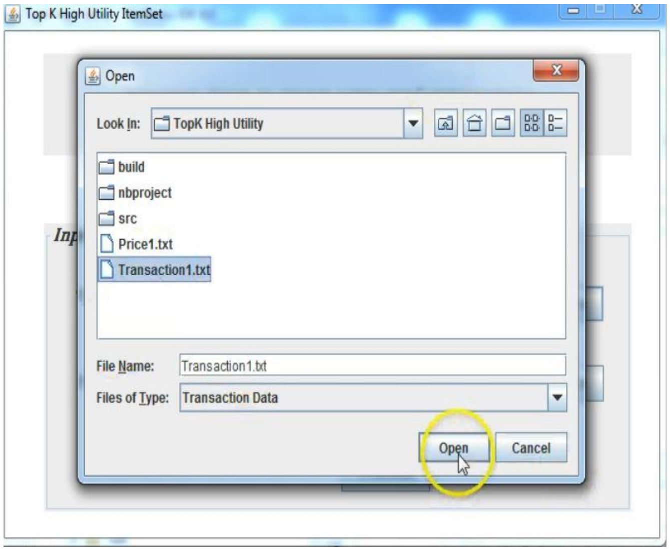
In the TKO algorithm, initially, a list of top-k high utility item sets is gen- erated. Parallelly,it raises the value of minimum utility threshold dynamicallyand updates the list of top-k high utility item sets.

V RESULTS


Top K High Utility ItemSet

MINING TOP-K HIGH UTILITY ITEMSETS

Figure 2: Mining Top-K High Utility Itemstes

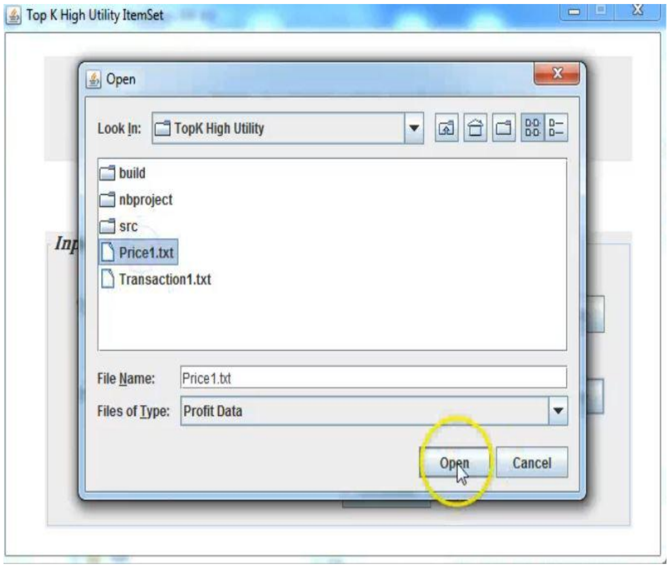


Look In: TopK High Utility

File Name: Transaction1.txt

Files of Type: Transaction Data

Figure 3: Input Transaction Data

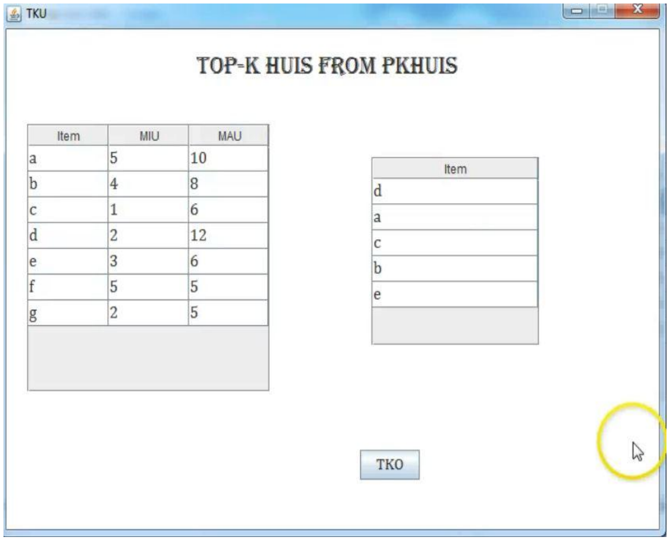


Look In: TopK High Utility

File Name: Price1.txt

Files of Type: Profit Data

Figure 4: Input Profit Data



Item	MIU	MAU
a	5	10
b	4	8
c	1	6
d	2	12
e	3	6
f	5	5
g	2	5

Item
d
a
c
b
e

Figure 5: Top-K Huis from PkHuis

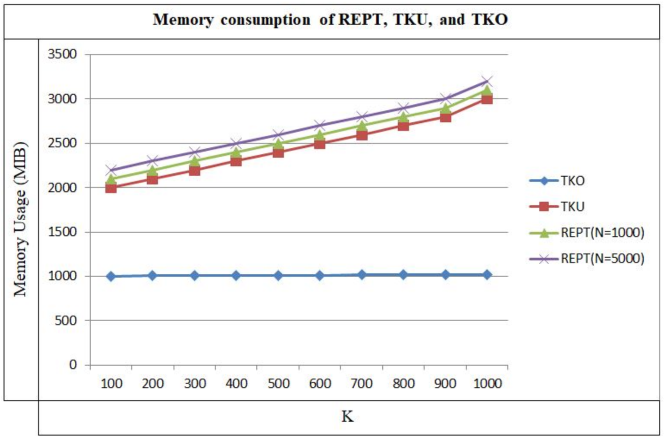


Figure 6: Memory consumption of REPT, TKU, and TKO

VI CONCLUSIONS

In this paper, I have studied the problem of top-k high utility itemsets mining, where k is the desired number of high utility itemsets to be mined. Two efficient algorithms TKU (mining Top-K Utility itemsets) and TKO (mining Top-K utility itemsets in One phase) are proposed for mining such itemsets without setting minimum utility thresholds. TKU is the first two-phase algorithm for mining top-k high utility itemsets, which incorporates five strategies PE, NU, MD, MC and SE to effectively raise the border minimum utility thresholds and further prune the search space. On the other hand, TKO is the first one-phase algorithm developed for top-kHUI mining, which integrates the novel strategies RUC, RUZ and EPB to greatly 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-theart two-phase and one-phase utility mining algorithms.

In this paper, an attempt will be made to improve the performance of algorithms for mining top-k high utility item sets by the use of pre-evaluation strategy. Also we try to raise the value of min util efficiently for better performance.

REFERENCES

[1] Serin Lee, Jong Soo Park, “Top-k High Utility Itemset Mining Based on Utility-List Structures”, IEEE, 2016.

[2] Jimmy Ming-Tai Wu, Jerry Chun-Wei Lin, Matin Pirouz and Philippe Fournier-Viger, “TUB-HAUPM: Tighter Upper Bound for Mining High Average-Utility Patterns”, IEEE Access, 2018.

[3] Snehal D. Ambulkar, Dr. Prashant N. Chatur, “Efficient Algorithms for mining High Utility Itemset”, IEEE International Conference on Recent Trends in Electrical, Electronics and Computing Technologies, 2017.

[4] Ms. Sharda Khode, Dr. Sudhir Mohod, “Mining High Utility Itemsets using TKO and TKU to find Top-k High Utility Web Access Patterns”, IEEE International

Conference on Electronics, Communication and Aerospace Technology, 2017.

[5] Bac Le, Cao Truong, Minh-Thai Tran, "Enhancing Threshold-Raising Strategies for Effective Mining Top-k High Utility Patterns", 4th NAFOSTED Conference on Information and Computer Science, 2017.

[6] Ingle Mayur Rajendra, Shri Chaitanya Vyas, Sanika Sameer Moghe, Deepali Deshmukh, Sachin Sakhare, Prof. Sudhanshu Gonge, "Implementing a Hybridof Efficient Algorithms For Mining Top-K High Utility Itemsets", IEEE, 2018.

[7] Ning Wang, Xiaokui Xiao, Yin Yang, Zhenjie Zhang, Yu Gu, Ge Yu, "PrivSuper: a Superset-First Approach to Frequent Itemset Mining underDifferential Privacy", IEEE 33rd International Conference on Data Engineering, 2017.

[8] Ruixin Yang, Mingyang Xu, Paul Jones, Nagiza Samatova, "Real Time UtilityBased Recommendation for Revenue Optimization via An Adaptive Online Top-K High Utility Itemsets Mining Model", IEEE 13th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD 2017), 2017.

[9]Junfu Yin, Zhigang Zheng, Longbing Cao, Yin Song, Wei Wei, "Efficiently Mining Top-K High Utility Sequential Patterns", IEEE 13th International Conference on Data Mining, 2013.