ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 4, Issue 12, December 2016

A Survey on Algorithm of Top-k High Utility Itemsets Mining over DataStream

Snehal Ghate, Prof. Pankaj Khambre

M. E Student, Dept. of Computer, GHRIET, Wagholi, Pune, India

Assistant Professor, Dept. of Computer, GHRIET, Wagholi, Pune, India

ABSTRACT: Data Mining can be defined as an activity that extracts some new nontrivial information contained in large databases. Traditional data mining techniques have focused largely on detecting the statistical correlations between the items that are more frequent in the transaction databases. Also termed as frequent itemset mining ,these techniques were based on the rationale that itemsets which appear more frequently must be of more importance to the user from the business perspective .Generally speaking, finding an appropriate minimum utility threshold by trial and error is a tedious process for users. If min_util is set too low, too many HUIs will be generated, which may cause the mining process to be very inefficientwe address the above issues by proposing a new framework for top-k high utility itemset mining, where k is the desired number of HUIs to be mined. Two types of efficient algorithms named TKU (mining Top-K Utility itemsets) and TKO (mining Top-K utility itemsets in One phase) are proposed for mining such itemsets without the need to set min_util. The term utility refers to the importance or the usefulness of the appearance of the itemset in transactions quantified in terms like profit , sales or any other user preferences. In High Utility Itemset Mining the objective is to identify itemsets that have utility values above a given utility threshold. In this paper we present a literature review of the present state of research and the various algorithms for high utility itemset mining.

KEYWORDS: Utility mining, high utility itemset mining, top-k pattern mining, top-k high utility itemset mining

I. INTRODUCTION

Data mining is concerned with analysis of large volumesofdata to automatically discover interesting regularities or relationships which in turn leads to better understanding of the underlying processes. The primary goal is to discover hidden patterns,unexpected trends in the data. Data mining activities uses combination of techniques from database technologies, statistics, artificial intelligence and machine earning.

The term is frequently misused to mean any form of large-scale data or information processing. The actual data mining task is the automatic or semi-automatic analysis of large quantities of data to extract previously unknown interesting patterns.Over the last two decades data mining has emerged as a significant research area .This is primary due to the inter-disciplinary nature of the subject and the diverse range of application domains in which data mining based products and Techniques arebeing employed. This includesbioinformatics, genetics,medicine,clinical research,education,retail and marketing research.

Data mining has been considerably used in the analysis of customer transactions in retail research where it is termed as market basket analysis. Market basket analysis has also been used to identify the purchase patterns of the alpha consumer. Alpha consumers are people that play a key role in connecting with the concept behind the inception and design of a product.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 4, Issue 12, December 2016

II. RELATED WORK

Vincent S. Tseng, Cheng-Wei Wu, Philippe Fournier-Viger, and Philip S. Yu [1] proposed a novel framework for mining closed high utility itemsets (CHUIs), which serves as a compact and lossless representation of HUIs. This paper proposed three efficient algorithms named AprioriCH (Apriori-based algorithm for mining High utility closed itemsets), AprioriHC- D (AprioriHC algorithm with Discarding unpromising and isolated items) and CHUD (Closed High Utility Itemset Discovery) to find this representation. To recover all HUIs from the set of CHUIs, authors proposed a method called DAHU (Derive All HighUtility Itemsets) and that to without accessing the original database. Authors claimed that this technique achieves a massive reduction in the number of HUIs. AprioriHC-D and AprioriHC both algorithms can't perform well on dense databases when min_utility is low since they suffer from the problem of a large amount of candidates.

Chowdhury Farhan Ahmed, Syed Khairuzzaman Tanbeer, Byeong-Soo Jeong, and Young-Koo Lee [2] focused on incremental and interactive data mining because this provides the ability to use previous data structures. They proposed three tree structures and claimed that these structures efficiently perform incremental and interactive HUP (High Utility Pattern) inning. His educes the calculations when a minimum threshold is changed or a database is updated. One of the tree structures, incremental HUP Lexicographic Tree (IHUPL-tree), is arranged according to an item's lexicographic order. It can capture the incremental data without any restructuring operation. His second tree structure is the IHUP Transaction Frequency tree (IHUPTF-Tree). This is simple and easy to construct and handle. In this tree the items are arranged according to their transaction frequency. It does not require any restructuring operation even when the data base is incrementally updated. Authors have achieved the less memory consumption. The mining time is reduced by designing the IHUP-Transaction Weighted Utilization Tree (IHUPTWU-Tree). This tree is based on the TWU value of items in descending order.

Vincent S. Tseng, Bai-En Shie,Cheng-Wei Wu, and Philip S.Yu[3] proposed two algorithms utility pattern growth (UP-Growth) and UP-Growth+, for mining high utility itemsets it et of effective strategies for pruning candidate itemsets. A tree-based data structure named utility pattern tree (UP-Tree) s maintained for the information of high utility itemsets such hat candidate itemsets are generated with only two scans of database. The authors then proposed two efficient algorithms UP-Growth and UP-Growth+ for mining high utility itemsets from transaction databases. It is found that the runtime is improved especially when databases contain lots of long transactions. HUCtree to solve the problems of large number of candidates and multiple times of database scanning. This is followed by the algorithms to HUI mine, but these algorithms discover HTWUIs in a pattern-growth approach. Still, the problem of huge memory usage for constructing and visiting conditional trees is unavoidable.

Chun-Jung Chu, Vincent S. Tseng, Tyne Liang [4] considered the database where the utility values for theitems could be negative. They have proposed the method HUINIV (High Utility Itemsets with Negative Item Values)-Mine and claimed that this method can effectively identify high utility itemsets by generating fewer high transaction-weighted utilization itemsets such that the execution time can be reduced substantially in mining the high utility itemsets. They also claimed that memory requirement less and there is less CPU I/O. This HUINIV-Mine algorithm is based on principle of the Two-Phase algorithm and augments negative item value.

Hua-Fu Li, Hsin-Yun Huang, Suh-Yin Lee [5] propose two efficient one pass algorithms MHUI-BIT and MHUITID for mining high utility itemsets from data streams within a transaction sensitive sliding window. To improve the efficiency of mining high utility itemsets two effective representations of a lexicographical tree-based summary data structure and itemset information were developed.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 4, Issue 12, December 2016

III. PROPOSED ALGORITHM

A. *DESCRIPTION OF THE PROPOSED ALGORITHM:*

We propose ideas to raise min_utilBorder during the Phase I of TKUBase. Fig. 1 gives the resulting pseudo code of TKUBase. Each time a candidate itemset X is found by the UP-Growth search procedure, the TKUBase algorithm checks whether it's estimated utility value ESTU(X) is no less than min_utilBorder. If ESTU(X) is less than min_utilBorder, X and all its concatenations are not top-k HUIs (Property 3). Besides, Tubas checks whether MAU(X) is no less than min_utilBorder (Line5). If MAU(X) is smaller than min_utilBorder, X is not a top-k HUI (Lemma 6). Otherwise, X is considered a candidate for Phase II and it is outputted with min {ESTU(X), MAU(X)} according to Property. If X is a valid PKHUI and MIU(X) _ min_utilBorder, MIU(X) can be used to raise min_utilBorder by the proposed strategy.

IV. PSEUDO CODE

ALGORITHM: TKU_{Base}

Input: (1) A database D;
(2) The number of desired HUIs k;
Output: (1) The complete set of PKHUIs C;

```
01.  Set min_util_Border ← 0; TopK-MIU-List ← Ø; C← Ø;
02.  Construct a UP-Tree by scanning D twice;
03.  //Apply a UP-Growth search procedure to generate PKHUIs;
04.  For each PKHUI generated with estimated utility ESTU(X) do
05.  { If (ESTU(X) ≥ min_util_Border and MAU(X) ≥ min_util_Border)
06.    { Output X and min{ESTU(X), MAU(X); C ← C ∪ X;
07.      If(MIU(X) ≥ min_util_Border )
08.      { //Raise min_util_Border by the strategy MC;
09.        min_util_Border ← MC(MIU(X), TopK-MIU-Lit);
10.      }
11.    }
12.  }
```

V. PROBLEM STATEMENT

Generally speaking, finding an appropriate minimum utility threshold by trial and error is a tedious process for users. If min_util is set too low, too many HUIs will be generated, which may cause the mining process to be very inefficient. On the other hand, if min_util is set too high, it is likely that no HUIs will be found. We address the above issues by proposing a new framework for top-k high utility itemset mining, where k is the desired number of HUIs to be mined.

VI. CONCLUSION AND FUTURE WORK

In this paper, we have studied the problem of top-k highutility itemsets mining, where k is the desired number ofhigh utility itemsets to be mined. Two efficient algorithmsTKU (mining Top-K Utility itemsets) and TKO (miningTop-K utility itemsets in one phase) are proposed formining such itemsets without setting minimum utility thresholds. TKU is

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 4, Issue 12, December 2016

the first two-phase algorithm for miningtop-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-k HUI mining, which integrates the novel strategies RUC, RUZ and EPB to greatly improve its performance. Empirical evaluations on different types of real and synthetic datasetsshow 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-heart two-phase and one-phase utility mining algorithms.

REFERENCES

[1] Vincent S. Tseng, Cheng-Wei Wu, Philippe Fournier-Viger, and Philip S. Yu, "Efficient Algorithms for Mining the Concise and Lossless Representation of High Utility Itemsets", IEEE Transactions on Knowledge and Data Engineering, Vol. 27, No. 3, 2015.
[2] Chowdhury Farhan Ahmed, Syed Khairuzzaman Tanbeer, Byeong- Soo Jeong, and Young-Koo Lee, "Efficient Tree Structures for High Utility Pattern Mining in Incremental Databases ", IEEE Transact ions on Knowledge and Data Engineering, Vol.21, No 12, December 2009, pp 1708-1721.
[3] Vincent S. Tseng, Bai-En Shie, Cheng-Wei Wu, and Philip S. Yu, "Efficient Algorithms for Mining High UtilityItemsets from Transactional Databases", IEEE Transactions on Knowledge and Data Engineering, Vol.25, No. 8, AUGUST 2013, pp 1772-1786.
[4] Chun-Jung Chu, Vincent S. Tseng, Tyne Liang, "An efficient algorithm for mining high utility itemsets with negative item values in large databases", Elsevier, 2009.doi:10.1016/j.amc.2009.05.066
[5] Hua-Fu Li, Hsin-Yun Huang, Suh-Yin Lee, "Fast and memory efficient mining of high-utility itemsets from data streams: with and without negative item profits", Springer, 2010. DOI 10.1007.
[6] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules," in Proc. Int. Conf. Very Large Data Bases, 1994, pp. 487– 499.
[7] C. Ahmed, S. Tanbeer, B. Jeong, and Y. Lee, "Efficient tree structures for high-utility pattern mining in incremental databases," IEEE Trans. Knowl. Data Eng., vol. 21, no. 12, pp. 1708–1721, Dec.
2009.
[8] K. Chuang, J. Huang, and M. Chen, "Mining top-k frequent patterns in the presence of the memory constraint," VLDB J., vol. 17, pp. 1321–1344, 2008.
[9] R. Chan, Q. Yang, and Y. Shen, "Mining high-utility itemsets," in Proc. IEEE Int. Conf. Data Mining, 2003, pp. 19–26
[10] P. Fournier-Viger and V. S. Tseng, "Mining top-k sequential rules," in Proc. Int. Conf. Adv. Data Mining Appl., 2011, pp. 180–194.
[11] P. Fournier-Viger, C.Wu, and V. S. Tseng, "Mining top-k association rules," in Proc. Int. Conf. Can. Conf. Adv. Artif. Intell., 2012, pp. 61–73.