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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 10, October 2016

A Survey on Efficient Algorithms for Mining High Utility Item sets from Transactional Databases

Shaikh Sahil Noormohamad[1], Helwade Pratik Pramod[2], Shinde Kiran Kashinath[3], Prof. Hirave K. S[4]

Department of Computer Engineering, Hon. Shri. Babanrao Pachpute Vichardhara Trust's Parikrama College of Engineering, Kashti, SPPU India[1,2,3,4]

ABSTRACT: High utility thing sets (HUIs) mining is a rising subject in information mining, which alludes to finding all thing sets having an utility meeting a client determined least utility edge min_util. Notwithstanding, 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 HUIs will 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. In this paper, we address the above issues by proposing another structure for top-k high utility thing set mining, where k is the coveted number of HUIs to be mined. Two sorts of proficient calculations named TKU (mining Top-K Utility thing sets) and TKO (mining Top-K utility thing sets in one stage) are proposed for mining such thing sets without the need to set min_util. We give an auxiliary examination of the two calculations with talks on their preferences and restrictions. Exact assessments on both genuine and manufactured datasets demonstrate that the execution of the proposed calculations is near that of the ideal instance of best in class utility mining calculations.

KEYWORDS: Utility mining, high utility item set mining, top-k pattern mining, top-k high utility item set mining, Data mining, frequent itemset, transactional database.

I. INTRODUCTION

FREQUENT item set mining (FIM) is a fundamental research topic in datamining. 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 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 important topic in data mining and has receivedextensive attention in recent years. In utility mining, each item is associated with a utility (e.g. unit profit) and an occurrence count in each transaction (e.g. quantity). 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) if its utility is no less than a user-specified minimum utility threshold min_util. HUI mining is essential to many applications such as streaming analysis market analysis mobile computing and biomedicine. However, efficiently mining HUIs in databases is not an easy task because the downward closure property usedin FIM does not hold for the utility of item sets. In other words, pruning search space for HUI mining is difficult because a superset of a low utility item set can be high utility. To tackle this problem, 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, 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. A classical 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. Although many studies have been devoted to 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 very small or very large. Besides, the choice of the threshold greatly influences the performance of the algorithms. If the threshold is set too low, too many HUIs will be presented to the

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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 10, October 2016

users 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 resourcesthey consume. On the contrary, 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 guessing and re-executing the algorithms over and over until being satisfied with the results. This process is both inconvenient and time-consuming.We build TKP although they have tradeoffs on memory usage. The reason is that TKO utilizes minimal node utilities for further decreasing overestimated utilities of item sets. Even though it spends time and memory 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 it spends time and memory 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. LITERATURE SURVEY

No.	Author Name	Basic Concept	Remarks
1	Vincent S. Tseng, Cheng- Wei Wu, Philippe FournierViger, and Philip S. Yu, 2015	Closed high utility itemset, lossless and concise representation	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
2	Chowdhury Farhan Ahmed, Syed Khairuzzaman Tanbeer, Byeong-Soo Jeong, and Young-Koo Lee, 2009	Incremental mining, Interactive mining	Authors used pattern growth approach, which avoids the problem of level wise candidate generation
3	Vincent S. Tseng, BaiEn Shie, Cheng-Wei Wu, and Philip S. Yu, 2013	Utility mining, External utility and Internal utility	Improvement in the run time especially when database contains lots of long transactions
4	Chun-Jung Chu a, Vincent S. Tseng b, Tyne Liang, 2009	Negative values for utilities if itemsets are considered	The critical requirements of temporal and spatial efficiency for mining high utility itemsets with negative item values are met. High scalability in dealing with large databases is achieved.
5	Hua-Fu Li, Hsin-Yun Huang, Suh-Yin Lee, 201	Mining High Utility Itemsets based on BIT vector	To improve the efficiency of mining high utility itemsets two effective representations of an extended 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)

Vol. 4, Issue 10, October 2016

6	Sen Su, Shengzhi Xu, Xiang Cheng, Zhengyi Li, and Fangchun Ya,2015	Differentially private FIM algorithm	A novel smart splitting method is proposed to transform the database. For a given database, the pre-processing phase needs to be performed only once.
7	Vincent S. Tseng, Cheng-Wei Wu, Viger, Philip S. Yu, 2015	Framework for top-k high utility itemset mining	Empirical evaluations on both real and synthetic datasets show that the performance of the proposed algorithms is close to that of the optimal case of state-of-the-art utility mining algorithms. where k is the desired number of high utility itemsets to be mined

III. EXISTING SYSTEM APPROACH

FREQUENT item set mining is a fundamental research topic in data mining (FIM) 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 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 important topic in data mining and has received extensive attention in recent years. In utility mining, each item is associated with a utility (e.g. unit profit) and an occurrence count in each transaction (e.g. quantity). 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) if its utility is no less than a user-specified minimum utility threshold min_util. HUI mining is essential to many applications such as streaming analysis, market analysis, mobile computing and biomedicine.

Disadvantages Of Existing System:-

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

IV. PROPOSED SYSTEM APPROACH

The concept of transaction weighted utilization (TWU) model 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, 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. A classical 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 with one database scan.

Advantages Of Proposed System:

1. Two efficient algorithms named TKU (mining Top-K Utility items ets) and TKO (mining Top-K utility item sets in one phase) are proposed for mining the complete set of top-k HUIs in databases without the need to specify the min_util threshold.
2. 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.

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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 10, October 2016

V. SYSTEM ARCHITECTURE

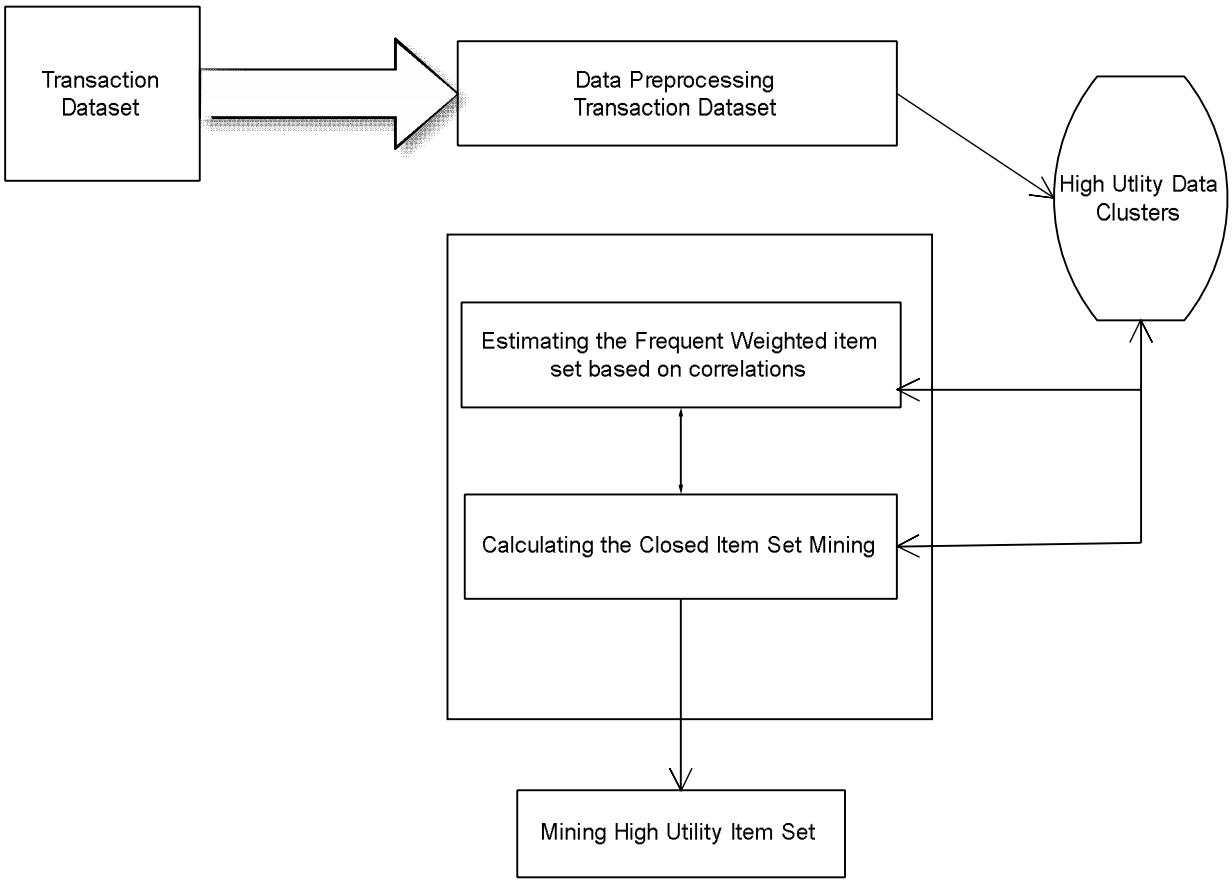

Fig No 01 System Architecture

VI. 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 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. TKU is the first two-phase algorithm for miningTop-k high utility item sets, 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 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.

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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 10, October 2016

[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.