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

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 5, Issue 10, October 2017

A Survey on an Efficient Hybrid Algorithm for High Utility Item Set in Parallel Mining Using Tku and Tko.

Nishant Singh[1], Harshit Kumar[1], Tanima Saha[1], Deepanshu Bhinda[1], P.S. Raskar[2], P.N Railkar[2]

B. E Students, Department of Computer Engineering. SKNCOE, Pune University, Pune, India[1]

Professor, Department of Computer Engineering. SKNCOE, Pune University, Pune, India[2]

ABSTRACT: High utility item sets (HUIs) mining is a rising subject in information mining.HUIs determined least utility edge min utility. Setting an appropriate min utility threshold is a difficult problem for users. In the event that min utility 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 utility is set too high, it is likely that no HUIs will be found. In this paper the above issues by proposing a new framework named efficient hybrid algorithm for high utility item set in parallel mining using TKU and TKO, where k is the desired number of HUIs to be mined. Two sorts of proficient calculations named TKU (mining Top-K Utility item sets) and TKO (mining Top-K utility item sets in one stage) are proposed for mining such item sets without the need to set min utility. We give an auxiliary examination of the two calculations with talks on their preferences and restrictions. TKO and TKU have excellent performance and scalability.

KEYWORDS:Utility mining, high utility item set mining, top-k pattern mining, top-k high utility item set mining

I. INTRODUCTION

Data mining is the process of analyzing data from different angles and summarizing it into useful data. Data mining is a tool for analyzing data. It allows users to analyze data from different levels or angles, arrange it, and the relationships among the data are found. Data mining is the process of finding patterns among sufficient of fields in thelarge relational database.A classical Top K model-based algorithm consists of two phases. In the first phase, called phase I, the complete setof 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. 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 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 resources they consume. On the contrary, if the threshold is set too high, no HUI will be found.

Background:

Frequently generate a huge set of HUIs and their mining performance is degraded consequently. Further, in case of long transactions in thedataset or low thresholds are set, then this condition may become worst. The huge number of HUIs forms a challenging problem to the mining performance since the more HUIs the algorithm generates, the higher

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

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 5, Issue 10, October 2017

processing time it consumes. Thus to overcome this challenges the efficient algorithms presented. Top k will not work on the parallel mining.

Motivation:
Set the value of k which is more intuitive than setting the threshold because k represents the number of Itemsets that users want to find whereas choosing the threshold depends primarily on database characteristics, which are often unknown to users.
Using a parameter k instead of the min_util threshold is very desirable for many applications. Top-k frequent pattern mining space cannot be straightly applied to top-k high utility Itemset mining.

Goals/Objectives:
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.Incontrast, UP-Growth performs better only when min_util is small. This is because when a 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. For 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 a 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. For the reasons mentioned above, the proposed algorithms UP-Growth and UP-Growth +achieve better performance than IHUP algorithm.

II. RELATED WORK

1. Vincent S. Tseng, Cheng- Wei Wu, Philippe FournierViger, and Philip S. Yu, 2015 - Closed high utility itemset, lossless.

Refer points-

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 candidate [1].

2.ChowdhuryFarhan Ahmed, Syed KhairuzzamanTanbeer, Byeong-SooJeong, and Young-Koo Lee, 2009 - Incremental mining, Interactive mining

Refer Points-

Authors used pattern growth approach, which avoids the problem of level wise candidate generation[2]

3. Vincent S. Tseng, BaiEnShie, Cheng-Wei Wu, and Philip S. Yu, 2013 - Utility mining, External utility and Internal utility

Refer points-

Improvement in the runtime especially when database contains lots of long transactions[3]

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

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 5, Issue 10, October 2017

4. Chun-Jung Chu a, Vincent S. Tseng b, Tyne Liang, 2009- Negative values for utilities if itemsets are considered

Refer points-

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[4]

5.Hua-Fu Li, Hsin-Yun Huang, Suh-Yin Lee, 201- Mining High Utility Itemsets based on BIT vector

Refer Points-

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[5]

6.Sen Su, ShengzhiXu, Xiang Cheng, Zhengyi Li, and Fangchun Ya,2015 - Differentially private FIM algorithm

Refer Points-

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[6]

7. Vincent S. Tseng, Cheng-Wei Wu, Viger, Philip S. Yu, 2015- Framework for top-k high utility Itemset mining

Refer Points-

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[7]

Existing System Approach:

In existing, frequently search item mining is a research (FIM) topic in data mining. However, the traditional mining 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:

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. Clearance search space for HUI mining is difficult because a superset of a low utility item set can be ahigh utility.

Proposed System Architecture:

In theProposed system, we address the above issues by proposing a new framework efficient hybrid algorithm for high utility item set in parallel mining using TKU and TKO.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

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

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 5, Issue 10, October 2017

the need to set min utility. We provide a structural comparison of the two algorithms with discussions on their advantages and limitations. Evaluations of 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.

Advantages:

1. Two efficient algorithms named TKU (mining Top-K Utility items etc.) 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 number of nodes maintained in memory could be reduced and the mining algorithm could achieve better performance

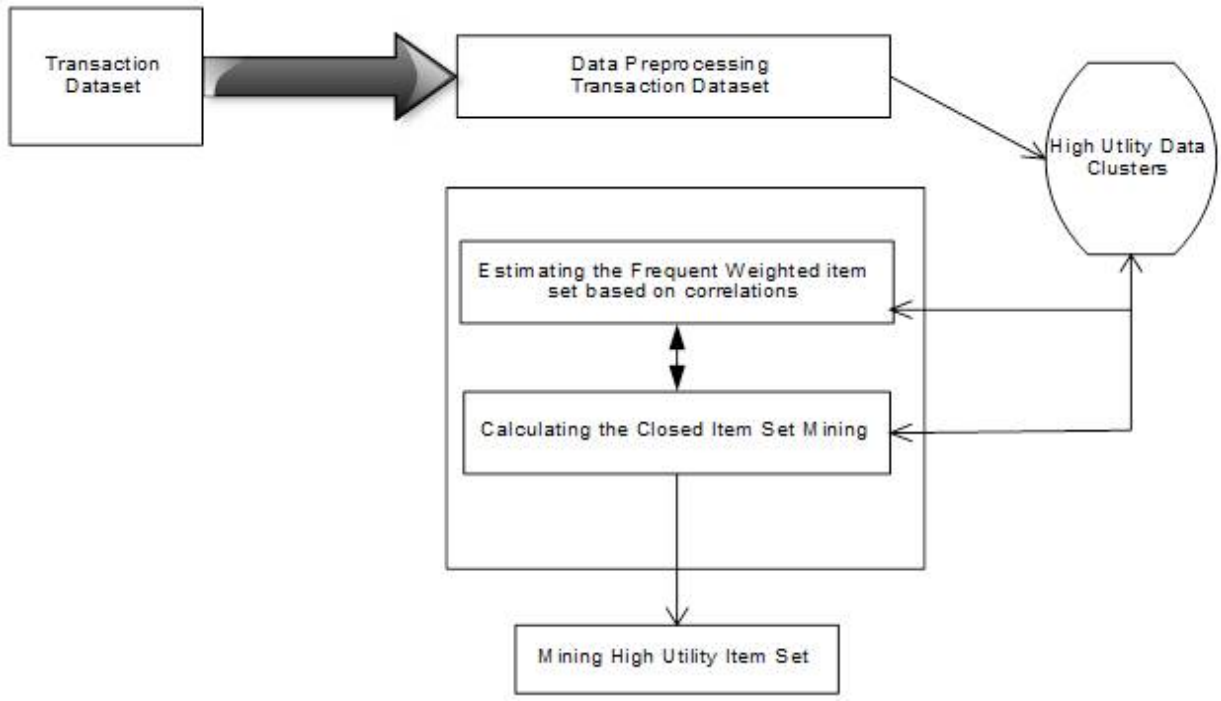

Fig 1. Proposed System Architecture

III. CONCLUSION

Here In thepaperthe problem of extracting high utility Item sets, Two efficient TKU algorithms (Top-K Utility Item sets) and TKO (mining Top-K utility set of items in one phase) are here proposed for extracting such sets of articles without using minimal useful thresholds concept. TKU is the first two-phase algorithm for extracting high-k high-utility objects. On the other hand, TKO is the first one-phase algorithm developed for HUI top-k extraction. evaluation of different types of Synthetic data sets that show the proposed algorithms have good scalability in large data sets and performance of the algorithms proposed is close to the optimal case of algorithms for extracting two-phase and two-phase good.

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

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 5, Issue 10, October 2017

REFERENCES

1.Vincent S. Tseng, Cheng- Wei Wu, Philippe FournierViger, and Philip S. Yu, 2015 – "Closed high utility item set, lossless".
2.ChowdhuryFarhan Ahmed, Syed KhairuzzamanTanbeer, Byeong-SooJeong, and Young-Koo Lee, 2009 - Incremental mining, Interactive mining.
3. Vincent S. Tseng, BaiEnShie, Cheng-Wei Wu, and Philip S. Yu, 2013 - Utility mining, External utility and Internal utility
4. Chun-Jung Chu a, Vincent S. Tseng b, Tyne Liang, 2009- Negative values for utilities if itemsets are considered
5.Hua-Fu Li, Hsin-Yun Huang, Suh-Yin Lee, 201- Mining High Utility Itemsets based on BIT vector
6.Sen Su, ShengzhiXu, Xiang Cheng, Zhengyi Li, and Fangchun Ya,2015 - Differentially private FIM algorithm
7. Fournier-Viger and V. S. Tseng, "Mining top-k sequentialrules," in Proc. Int. Conf. Adv. Data Mining Appl., 2011, pp. 180–194.