ISSN:0975-9646

Shrewd Technique for Mining High Utility Itemset via TKU and TKO Algorithm

R.Nandhini, Dr.N.Suguna
Department of Computer Science and Engineering
Akshaya College of Engineering and Technology
India

Abstract -High utility item sets (HUIs) mining is an emanate topic in data mining, which refers to discovering all item sets having a utility meeting a user-specified minimum utility threshold. We used three efficient algorithm AprioriCH (Apriori-based algorithm for mining High utility closed itemset), Apriori HC-D (AprioriHC algorithm with removing unpromising and isolated items) and CHUD (Closed High Utility itemset Discovery) algorithm. However, setting min_util appropriately is a hard problem for users. Finding an appropriate minimum utility threshold by trial and error is a tough process for users. Setting very low min_util, HUIs will be generated in large, which may cause the mining process to be very inefficient. In another way, by setting too high min_util, no HUI will be found. So the above issues is addressed by proposing a framework for mining top-k high utility item set , where k is the desired number of HUIs to be mined. Two types of proficient algorithms named 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 the need to set min_util.

Keywords -- Utility Mining, Data Mining, High Utility Item Set, Closed High Utility Item Set, Minimum Utility Threshold, Top-K High Utility Item Set, Lossless and Concise representation.

1. INTRODUCTION

High utility itemsets mining identifies itemsets where its utility satisfies a given threshold. It allows users to quantify the advantage of items using different values. Thus, it reflects the impact of distinct items. High utility itemsets mining is used in decision-making process of many applications, like retail marketing and Web service, since items are very different in many aspects in real applications. Experiments on real-world applications illustrate the significance of high utility itemsets in business decision-making; it also gives the difference between frequent itemsets and high utility itemsets. One of its main applications is market basket analysis. Market Basket Analysis is a important modelling technique based upon the theory that if you buy a certain set of items, you are more (or less) likely to buy another set of items. An itemset is called a high utility itemset (HUI) if its value is no lower than a user-specified minimum utility threshold; or else, it is called a low utility itemset. Utility mining is the most important task and has a wide range of applications such as website click stream analysis, cross marketing.

1.1 Data Mining

Data mining is considered with analysis of large volumes of data to automatically discover interesting regularities or relationships which are in turn leads to good understanding of the Underlying processes. The primary goal is to find hidden patterns, unexpected trends in the data. Data mining activities uses combination of technique from database technologies, statistics, and artificial intelligence and includes machine learning also.

The term is often misused to mean any form of large amount data or information processing. The actual data mining task is the automatic or semi-automatic analysis of big quantities of data to extract previously unknown interesting patterns. Over the last two decades data mining has appeared as a important research area .This is primary due to the inter-disciplinary nature of the subject and the varied range of application domains in which data mining based products and techniques are being engaged. These all includes bioinformatics, genetics, medicine, clinical research, education, retail and marketing research.

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

1.2 Sequential pattern mining

Sequential pattern mining has emerged as an important topic in data mining. It has proven to be very essential for handling order-based critical business problems, such as behavior analysis, gene analysis in bioinformatics and weblog mining. For example, sequence analysis is widely employed in DNA and protein to discover interesting structures and functions of molecular or DNA sequences. The collection of interesting sequences is generally based on the frequency/support framework: sequences of high frequency are treated as significant. Under this framework, the downward closure property plays a fundamental role for varieties of algorithms designed to search for frequent sequential patterns.

1.3 Utility itemset mining

Utility itemset mining, also generally called utility pattern mining, was first introduced. Every item in the itemsets is associated with an additional value, called internal utility which is the quantity (i.e. count) of the item. An external utility is attached to an item, showing its quality (e.g. price). Mining high utility itemsets is much more testing than discovering common itemsets, since the fundamental downward closure assets in frequent itemset mining does not hold in value itemsets.

1.4 Frequent Itemset Mining

An itemset can be definite as a non-empty set of items. An itemset with k dissimilar items is termed as a k-itemset. For e.g. Consider the combination of bread, butter, and milk

may signify a 3-itemset in a supermarket transaction . Frequent itemsets are the itemsets that appear frequently in the communication. The goal of frequent itemset mining is to identify all the itemsets in a transaction dataset. Frequent itemset mining acting as an necessary role in the theory and practice of many important data mining tasks , like mining association rules , long patterns ,emerging patterns, and dependency rules. It has been useful in the meadow of telecommunications, census analysis and text analysis. The criterion of being frequent is uttered in terms of support value of the itemsets. That value of an itemset is the percentage of transactions that contain the itemset.

2. RELATED WORK

In this section, we introduce the preliminaries associated with high utility itemset mining, closed itemset mining and Compact representations of high utility itemsets.

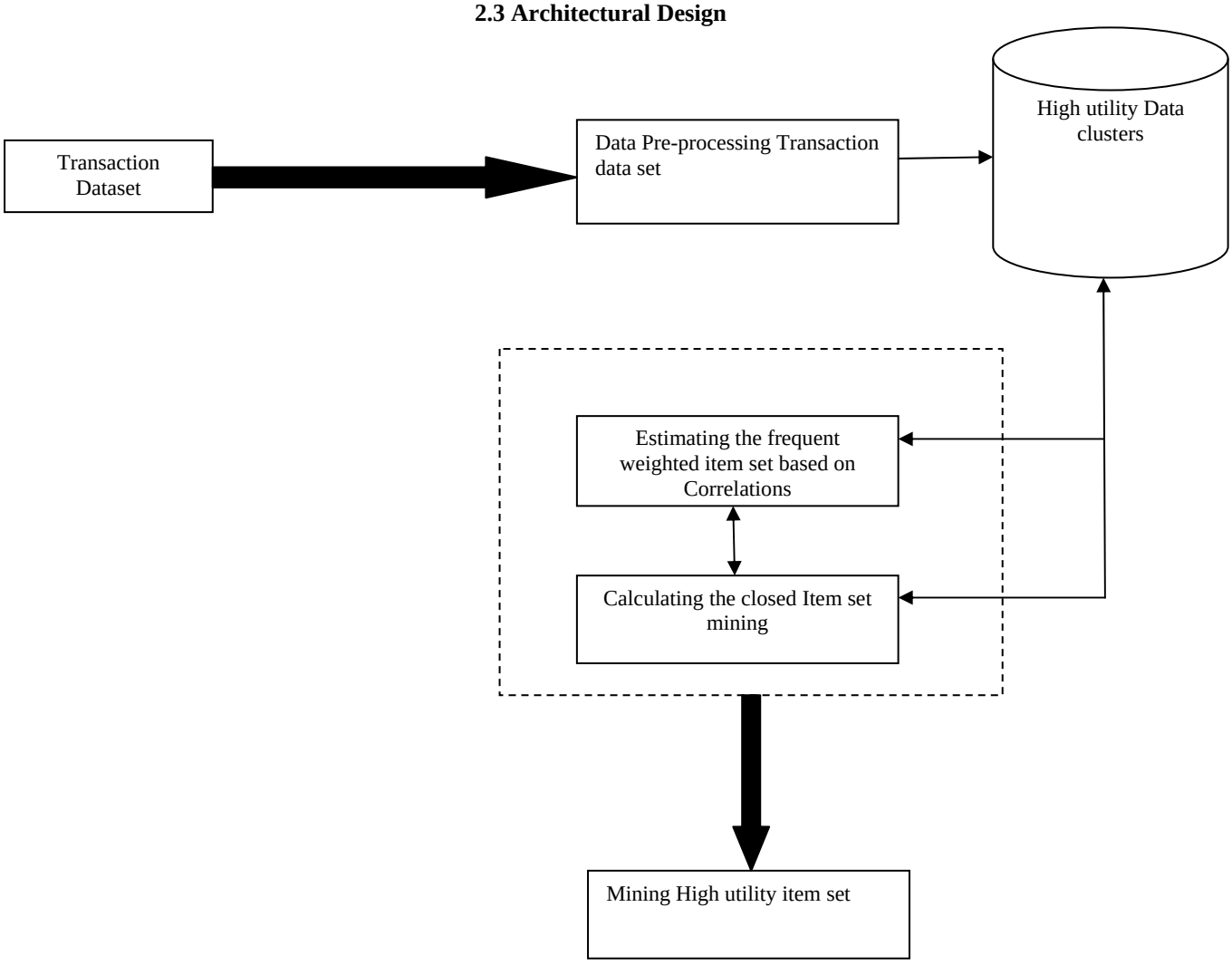
2.1 High Utility Itemset Mining

High utility quantitative Itemset mining refers to discover sets of items that cannot carry only high utilities (e.g., high profits) but also quantitative attributes like redundant data (duplicate data). Duplicate data will lead to large data consumption in resultant set. Proposed technique adopts a Compact representation to maintain the utility information of itemsets in databases with several efficient strategies integrated to prune the search space.

2.2 Closed high utility Itemset

Closed high utility Itemset with compact and lossless representation is a technique for mining high utility Itemset. This closed high utility itemsets extraction which is combination of concept closed Itemset into high utility itemset mining through proper Thresholding. The proposed representation is lossless owed to a new structure named as utility unit array that allow getting better to all High Utility item sets and their utilities professionally. 2) The proposed representation is also compact with three efficient algorithms named AprioriHC (Apriori based algorithm for mining High utility Closed itemset), AprioriHC-D (AprioriHC algorithm with Discarding unpromising and isolated item) and CHUD (Closed High Utility itemset Discovery)to find this demonstration. The AprioriHC and AprioriHC-D algorithms employs breadth first search to find CHUIs and inherits some nice properties from the well-known Apriori algorithm. The CHUD algorithms contain three novel strategies named REG, RML and DCM that greatly enhance its performance. A top-down method named as DAHU (Derive All High Utility itemsets) is implemented for efficiently recovering all HUIs from the set of CHUIs. The grouping of CHUD and DAHU provide a different way to obtain all HUIs and outperforms UP-Growth, one of the currently best methods for mining HUIs.

2.3 Architectural Design



3. EXISTING SYSTEM

Closed high utility Itemset with compact and lossless representation is a proposed technique for mining high utility Itemset. In High utility mining, closed high utility itemsets extraction which is combination of concept closed Itemset into high utility itemset mining through proper Thresholding. The proposed representation is lossless due to a fresh structure named utility unit array that allows recovering all High Utility item sets and their utilities efficiently.2) The proposed symbol is also compact with three efficient algorithms named AprioriHC, AprioriHC-D and CHUD to find this representation. The AprioriHC and AprioriHC-D algorithms employs breadth first search to find CHUIs and inherits some nice properties from the well-known Apriori algorithm. The CHUD algorithm includes three novel strategies named REG, RML and DCM that greatly enhance its performance. A top-down method named DAHU(Derive All High Utility itemsets) is implemented for efficiently recovering all HUIs from the set of CHUIs. The combination of CHUD and DAHU provides a new way to obtain all HUIs and outperforms UP-Growth, one of the currently best methods for mining HUIs.

4. PROPOSED SYSTEM

4.1 Top k utility mining Technique

The idea is to let the users specify k, i.e., the number of desired itemsets, as a alternate of identifying the minimum utility threshold.

- Set the value of k which is more intuitive than setting the threshold because k represent the number of itemsets that the 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 that rely on anti-monotonicity to snip the search space cannot be straightly applied to top-k high utility itemset mining.
- The second challenge is how to include the concept of top-k pattern mining along with the TWU model. Although the TWU model is widely used in utility mining, it is hard to adapt this model to top-k HUI mining because the exact utilities of itemsets are unknown in phase I.
- When a HTWUI is generated in phase I, we cannot guarantee that its utility is higher than other HTWUIs and that it is a top-k HUI before performing phase II.
- To guarantee that all the top-k HUIs can be captured in the set of HTWUIs, a immature approach is to run the algorithm with min_util = 0. However, this approach may face the problem of a very large search space.
- The third challenge is that the min_util threshold is not given in advance in top-k HUI mining. In traditional HUI mining, the search space can be efficiently pruned by the algorithms by using a given min_util threshold.
- However, in the scenario of top-k HUI mining, no min_util threshold is provided in advance.

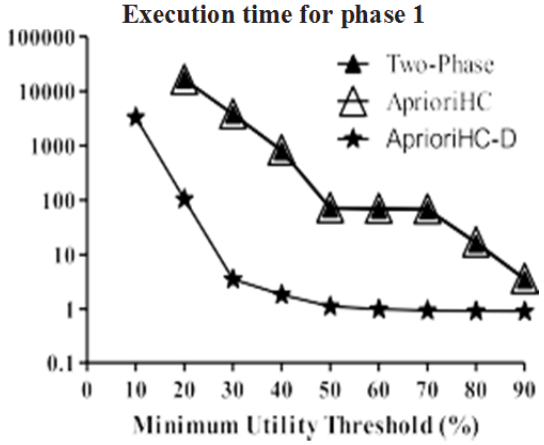
The minimum utility threshold is initially set to 0 and the designed algorithm has to gradually raise the threshold to prune the search space. Such a threshold is an internal parameter of the designed algorithm and is called the border minimum utility threshold min_util Border in this paper.

It is different from the external parameter min_util that is given by users in advance. If an algorithm cannot raise the min_util Border threshold effectively and efficiently, it would produce too many intermediate low utility itemsets during the mining process, which may degrade its presentation in terms of execution time and memory usage.

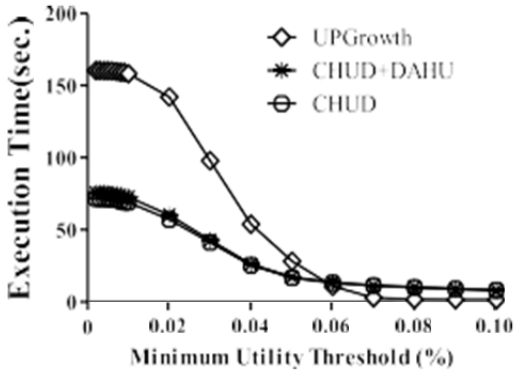
5. CONCLUSIONS

In this paper, we addressed the difficulty of redundancy in high utility itemset mining by proposing a lossless and compact representation named closed high utility itemsets, which has not been explored so far. To mine this representation, we used three efficient algorithms named AprioriHC, AprioriHC-D and CHUID. AprioriHC-D is an improved version of AprioriHC, which incorporates strategies DGU and IIDS for pruning candidates. AprioriHC and AprioriHCD perform a breadth-first search for mining closed high utility itemsets from horizontal database, while CHUD performs a depth-first search for mining closed high utility Itemsets from vertical database. The strategies incorporated in CHUD are efficient and novel. They have never been worn for vertical mining of high utility itemsets and closed high utility itemsets.

Execution time for phase 1



Total Execution Time



To efficiently recover all high utility Itemsets from closed high utility itemsets, we proposed an efficient method named DAHU (Derive All High Utility itemsets). To overcome all problems in existing we proposed apriori using TKU and TKO. Here the variable k is used to set the desired numbers of HUIs by the user itself. So that the problem of setting the minimum utility threshold for items is solved.

6.FUTURE WORK

High utility quantitative Itemset mining refers to discovering sets of items that carry not only high utilities (e.g., high profits) but also quantitative attributes like redundant data (duplicate data). Duplicate data will lead to large data consumption in resultant set. Proposed technique adopts a Compact representation to maintain the utility information of itemsets in databases with several effective strategies integrated to prune the search space.

REFERENCES

[1] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules," in Proc. 20th Int. Conf. Very Large Data Bases, 1994, pp. 487–499.

[2] C. F. Ahmed, S. K. Tanbeer, B.-S. Jeong, and Y.-K. 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.

[3] J.-F. Boulicaut, A. Bykowski, and C. Rigotti, "Free-sets: A condensed representation of Boolean data for the approximation of frequency queries," Data Mining Knowl. Discovery, vol. 7, no. 1, pp. 5–22, 2003.

[4] T. Calders and B. Goethals, "Mining all non-derivable frequent itemsets," in Proc. Int. Conf. Eur. Conf. Principles Data Mining Knowl. Discovery, 2002, pp. 74–85

[5] 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.

[6] R. Chan, Q. Yang, and Y. Shen, "Mining high utility itemsets," in Proc. IEEE Int. Conf. Data Min., 2003, pp. 19–26.

[7] A. Erwin, R. P. Gopalan, and N. R. Achuthan, "Efficient mining of high utility itemsets from large datasets," in Proc. Int. Conf. Pacific-Asia Conf. Knowl. Discovery Data Mining, 2008, pp. 554–561. 738 IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 27, NO. 3, MARCH 2015

[8] K. Gouda and M. J. Zaki, "Efficiently mining maximal frequent itemsets," in Proc. IEEE Int. Conf. Data Mining, 2001, pp. 163–170.