Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 5.258

IJCSMC, Vol. 5, Issue. 7, July 2016, pg.297 – 303

USING EFFICIENT MINING ALGORITHMS TO OBTAIN HIGH PROFITABLE ITEM SET

Mr. Bharath Bhushan S [1], Ms. Leena Shibu [2]

[1]M.Tech Student Department of Computer Science and Engineering, New Horizon College of Engineering, Bangalore, India

[2]Assistant Professor, Department of Computer Science and Engineering, New Horizon College of Engineering, Bangalore, India

[1] Bharath93316@gmail.com; [2] linashibu@gmail.com

Abstract- High Utility Item sets (HUIs) are profitable data mined from database. Presenting too many of High Utility Item sets to users degrades the efficiency of mining process. For achieving efficiency in mining process and provide compact mining results to user, proposed system uses a framework for mining closed high utility item sets (CHUIs),which serves as a brief and lossless representation of HUIs. System proposes three efficient algorithms named AprioriHC, AprioriHC-D and CHUD (Closed High Utility Item set Discovery) to find this representation. Proposed algorithms are very efficient and approaches achieve a massive reduction in the number of HUIs. In addition, all HUIs can be recovered by DAHU (Derive All High Utility Item sets), the combination of CHUD and DAHU performs better than the existing algorithms for mining HUIs.

KEYWORDS: High utility itemsets, utility mining, Closed high utility Itemsets

I. INTRODUCTION

The objective of this paper is to solve the problem of redundancy in high utility item set mining, by proposing a lossless and brief representation of HUIs, which are presented to users. Utility mining is the application of data mining techniques to discover patterns from the datasets. Utility itemsets typically consist of items with different values such as utilities, and the aim of utility mining is to identify the itemsets with highest utilities. Utility

is a measure of how ‘‘useful” an itemsets. The definition of the utility of an itemset X, u(X), states that it is equal to the sum of the utilities of X of all the transactions containing X. The goal of utility mining is to identify high utility itemsets, which drive a large portion of the total utility. Traditional association rules of mining models assume that the utility of each item is always 1 and that the quantity of sales is either 0 or 1; thus it is only a special case of utility mining in which the utility or the quantity of sales of each item can be any number. If u(X) is greater than a specified utility threshold, X is a high utility itemset; otherwise, it is a low utility itemset.

High utility item sets (HUIs) mining is an emerging topic in data mining, which refers to discovering all item sets having a utility meeting a user-specified minimum utility threshold.

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 up on 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

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 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.

II. RELATED WORK

In this section, we introduce the preliminaries associated with high utility itemset mining, closed high itemset mining.

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 due to a new structure named as utility unit array that allow getting better to all High Utility item sets and their utilities professionally. 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.

III. EXISTING SYSTEM

Many studies were proposed for mining HUIs, but they often present a large number of high utility itemsets to users. A very large number of high utility itemsets makes it difficult for the users to comprehend the results. It may also cause the algorithms to become inefficient in terms of time and memory requirement, or even run out of memory. It is widely recognized that the more high utility itemsets the algorithms generate, the more processing they consume. The performance of the mining task decreases greatly for low minimum utility thresholds or when dealing with dense database.

Disadvantages

The disadvantages of the existing system are listed below:-

- A very large number of high utility itemsets makes it difficult for the users to comprehend the results.
- It may also cause the algorithms to become inefficient in terms of time and memory requirement, or even run out of memory.
- The performance of the mining task decreases greatly for low minimum utility thresholds or when dealing with dense databases

IV. PROPOSED SYSTEM

The proposed mechanism is divided into the following steps:-

- The proposed representation is lossless due to a new structure named utility unit array that allows recovering all HUIs and their utilities efficiently.
- The proposed representation is also compact.
- We propose three efficient algorithms named AprioriHC (Apriori-based algorithm for mining High utility Closedþ itemset), AprioriHC-D (AprioriHC algorithm with Discarding unpromising and isolated items) and CHUD (Closedþ High Utility itemset Discovery) 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. Results show that CHUD is much faster than the state-of-the-art algorithms for mining all HUIs.
- We propose a top-down method named DAHU (Derive All High Utility itemsets) for efficiently recovering all HUIs from the set of CHUIs.

Advantages

The advantages are listed below:-

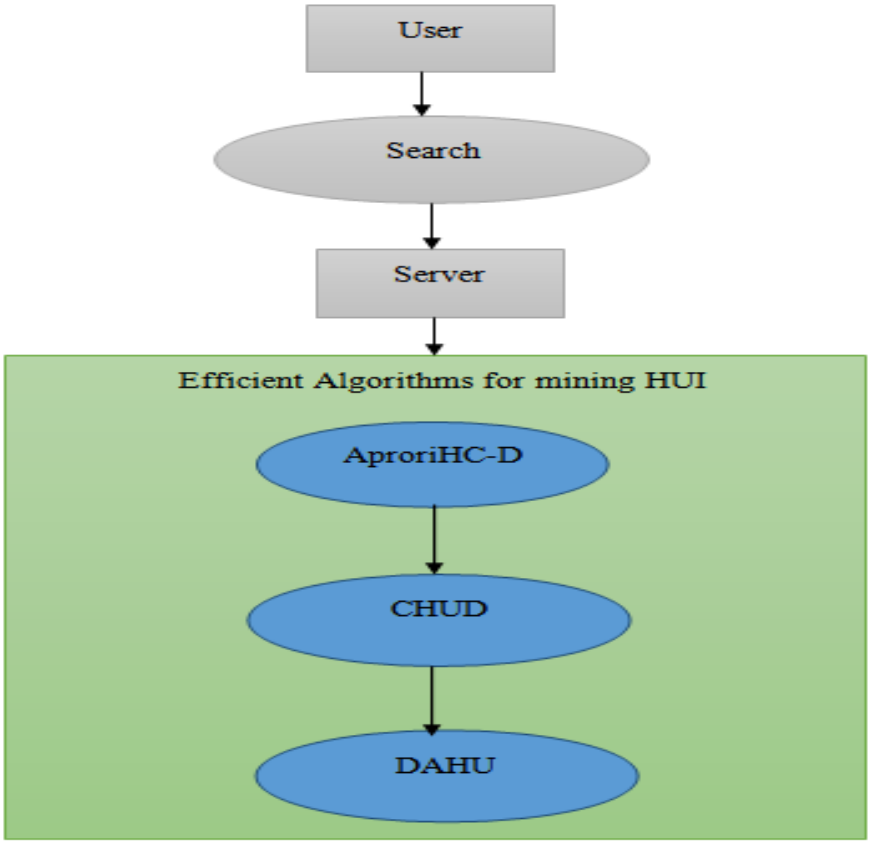
- Due to the new structure in the proposed mechanism all HUI itemset and their utilities are recovered efficiently.
- In the proposed mechanism, we have proposed 7 algorithm which performs based on the HUI which helps to break and work efficiently without any delay.
- As the work is divided among different algorithms so the utility of the search mechanism increases as compared with the previous ones.

SYSTEM DESIGN

The system design process builds up general framework building design. Programming outline includes speaking to the product framework works in a shape that may be changed into one or more projects. The prerequisite indicated by the end client must be put in a systematical manner. Outline is an inventive procedure; a great configuration is the way to viable framework. The framework "Outline" is characterized as "The procedure of applying different systems and standards with the end goal of characterizing a procedure or a framework in adequate point of interest to allow its physical acknowledgment". Different configuration components are taken after to add to the framework. The configuration detail portrays the components of the framework, the segments or components of the framework and their appearance to end-clients.

The architectural configuration procedure is concerned with building up a fundamental\ basic system for a framework. It includes recognizing the real parts of the framework and interchanges between these segments. The beginning configuration procedure of recognizing these subsystems and building up a structure for subsystem control and correspondence is called construction modelling outline and the yield of this outline procedure is a portrayal of the product structural planning.

The proposed architecture for this system is given below. It shows the way this system is designed and brief working of the system.



V. IMPLEMENTATION

MODULES

1. Push Closed Property into High Utility Itemset Mining

The first point that we should discuss is how to incorporate the closed constraint into high utility itemset mining. There are several possibilities. First, we can define the closure on the utility of itemsets. In this case, a high utility itemset is said to be closed if it has no proper superset having the same utility. However, this definition is unlikely to achieve a high reduction of the number of extracted itemsets since not many itemsets have exactly the same utility as their supersets in real datasets.

2. Efficient Algorithms for Mining Closed þ High Utility Itemsets

In this module we introduce three efficient algorithms AprioriHC (An Apriori-based algorithm for mining High utility Closed þ itemsets), AprioriHC-D(AprioriHC algorithm with Discarding unpromising and isolated items) and CHUD (Closed þ High Utility itemset Discovery) for mining CHUIs. They rely on the TWU-Model and include strategies to improve their performance. All algorithms consist of two phases named Phase I and Phase II. In Phase I, potential closed high utility itemsets (PCHUIs) are found, which are defined as a set of itemsets having an estimated utility (e.g. TWU) no less than abs_ min_ utility. In Phase II, by scanning the database once, CHUIs are identified from the set of PCHUIs found in Phase I and their utility unit arrays are computed.

3. Efficient Recovery of High Utility Itemsets

In this module, we present a top-down method named DAHU (Derive All High Utility itemsets) for efficiently recovering all the HUIs and their absolute utilities from the complete set of CHUIs. It takes as input an absolute minimum utility threshold abs_ min_ utility, a set of CHUIs HC and ML the maximum length of itemsets in HC. DAHU outputs the complete set of high utility itemsets $H = U^{k}_{i=1} H_k$ respecting abs_ min _utility, where H_k denotes the set of HUIs of length k.

CONCLUSION

In this paper providing a solution to finding a high utility itemset from set of closed high utility itemsets using efficient mining algorithms and solving the problem of redundancy, performance results of the algorithms also better compare to existing algorithms and giving effective results to the users.

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.

[9] T. Hamrouni, "Key roles of closed sets and minimal generators in concise representations of frequent patterns,"Intell. Data Anal.,vol. 16, no. 4, pp. 581–631, 2012.

[10] J. Han, J. Pei, and Y. Yin, "Mining frequent patterns without candidate generation," in Proc. ACM SIGMOD Int. Conf. Manage.Data, 2000, pp. 1–12.

[11] T. Hamrouni, S. Yahia, and E. M. Nguifo, "Sweeping the disjunctive search space towards mining new exact concise representations of frequent itemsets," Data Knowl. Eng., vol. 68, no. 10, pp. 1091–1111, 2009.

[12] H.-F. Li, H.-Y. Huang, Y.-C. Chen, Y.-J. Liu, and S.-Y. Lee, "Fast and memory efficient mining of high utility itemsets in data streams," in Proc. IEEE Int. Conf. Data Mining, 2008, pp. 881–886.

[13] C.-W. Lin, T.-P. Hong, and W.-H. Lu, "An effective tree structure for mining high utility itemsets," Expert Syst. Appl., vol. 38, no. 6,pp. 7419–7424, 2011.

[14] G.-C. Lan, T.-P. Hong, and V. S. Tseng, "An efficient projectionbased indexing approach for mining high utility itemsets," Knowl. Inf. Syst, vol. 38, no. 1, pp. 85–107, 2014.

[15] H. Li, J. Li, L. Wong, M. Feng, and Y. Tan, "Relative risk and odds ratio: A data mining perspective," in Proc. ACM SIGACT-SIGMOD-SIGART Symp. Principles Database Syst., 2005, pp. 368–377.

