IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 11, November 2016

Survey on Map-Reduce Based High Utility Item Sets Mining Using Concise and Lossless Representation

Ashwini Bhosale [1], **Veena Bhende** [2]

Research Scholar, Dept of Computer Science, AISSMSIOIT, Pune [1, 2]

Abstract: Now days high utility item sets mining (HUIs) from the large datasets is becoming the vital task of data mining in which discovery of item set with high utilities. But the existing previous methods are representing large number of HUIs to end user which resulted into inefficient performance of utility mining. To overcome this problems, in this project we are presenting the hybrid novel framework of HUIs with goal achieving the high efficiency for the mining task and provide a concise mining result to users using parallel data computing technology to process large dataset fast. The hybrid framework is proposed which is mining close dþ high utility item sets (CHUIs), which serves as a compact and lossless representation of HUIs. We used recent three efficient algorithms such as Apriori CH (Apriori-based algorithm for mining High Utility Closed þ item set), Apriori HC-D (Apriori HC algorithm with Discarding unpromising and isolated items) and CHUD (Closed þ High Utility Item set Discovery) to find this representation. Finally, the method DAHU (Derive All High Utility Item sets) is used to recover all HUIs from the set of CHUIs regardless of accessing main database. To improve the time performance of this approach our contribution is to used map-reduce framework to discover HUIs from last dataset faster as compared to existing recent method.

Keywords: Frequent item set, close dþ high utility item set, lossless and concise representation, utility mining, data mining.

I. INTRODUCTION

In selective marketing, decision analysis, and business management, the discovery of frequent relationship among a huge database has been known to be useful. By searching for sets of items that are frequently purchased together a popular area of its applications is the market basket analysis, which studies the buying behaviours of customers. Specifically, let I = {x1, x2... xm} be a set of items. A set X ⊆ I with m = | X | is called an m-item set or simply an item set. Formally, in the database that contain X, is larger than the minimum support threshold, indicating that the presence of item set X is signify can't in the database, an item set X refers to a frequent item set or a large item set if the support of X, i.e., the fraction of transactions. However, from two inherent obstacles, namely, (1) the subtle determination of the minimum support; (2) the unbounded memory consumption, it is reported that discovering frequent item sets suffers. Specifically, in previous, works without specific knowledge, a critical problem "What is the appropriate minimum support?" is usually left unsolved to users. Note that in an extremely large size of frequent item sets at the cost of execution efficiency setting the minimum support is quite subtle since a small minimum support may result.

Oppositely, for marketing decisions setting a large minimum support may only generate a few item sets, which cannot provide enough information. In order to obtain a desired result, to tune the minimum support over a wide range, users in general need. For the applicability of mining frequent item sets, this is very time-consuming and indeed is a serious problem. Furthermore, in practice is the large memory consumption, another issue which will be faced. Especially when the minimum support is small or the database size is large a large memory, which may not be affordable in most personal computers nowadays, is in general required during the mining process.

From executing the frequent item set mining It will result in the serious "out of memory" system crash, making users shy away. Note that users may tolerate to mine frequent item sets offline. In every night as long as users are able to make their marketing decisions in the morning for example, frequent item sets can be discovered. In contrast, in a commercial mining system the system crash due to the "out of memory" error is repulsive.

II. LITERATURE SURVEY

K. Chuang, J. Huang, and M. Chen, "Mining top-k frequent patterns within the presence of the memory constraint," VLDB J., vol. 17, pp. 1321–1344, 2008. [5], we tend to explore during this paper a feasible fascinating mining task to retrieve top-k (closed) item sets within the presence of the memory constraint.

Specifically, as against most previous works that focus on rising the mining potency or on reducing the memory size by best effort, we tend to initial decide to specify the

DOI 10.17148/IJARCCE.2016.51167

accessible higher memory size that may be used by mining frequent item sets. To accommodates the edge of the memory consumption, two efficient algorithms, known as MTK and MTK_Close, are devised for mining frequent item sets and closed item sets, severally, except specifying the refined minimum support. Instead, users just want to provides a additional human-understandable parameter, particularly the required range of frequent (closed) item sets k.

C. F. Ahmed, S. K. Tanbeer, B.-S. Jeong, and Y.-K. Lee, "Efficient tree structures for high utility pattern mining in progressive databases," IEEE Trans. Knowl. Data Eng., vol. 21, no. 12, pp. 1708– 1721, Dec. 2009. [2], recently, high utility pattern (HUP) mining is one in all the foremost important analysis issues in processing due to its ability to consider the non-binary frequency values of things in transactions and fully various profit values for every item.

On the other hand, progressive and interactive data processing provide the flexibleness to use previous data structures and mining results in order to reduce back redundant calculations once a database is updated, or once the minimum threshold is changed. Into this paper, we tend to propose three novel tree structures to expeditiously perform progressive and interactive HUP mining.

R. Chan, Q. Yang, and Y. Shen, "Mining high utility item sets," in Proc. IEEE Int. Conf. Data Min., 2003, pp. 19–26. [6], ancient association rule mining algorithms just generate an outsized range of extremely frequent rules, however these rules don't offer helpful answers for what the high utility rules are. we tend to develop a unique plan of Top-K objective-directed data processing, that focuses on mining the Top-K high utility closed patterns that directly support a given business objective.

To association mining, we tend to add the conception of utility to capture extremely fascinating statistical patterns and available a level-wise item-set mining algorithm. With each positive and negative utilities, the ant monotone pruning strategy in Apriori algorithm no more holds. In response, we develop a replacement pruning strategy supported utilities that permit pruning of low utility item set to be done by means that of a weaker however ant monotonic condition.

A. Erwin, R. P. Gopalan, and N. R. Achuthan, "Efficient mining of high utility item sets from large datasets," inProc. Int. Conf. Pacific Asia Conf. Knowl. Discovery Data Mining, 2008, pp. 554–561. [7], High utility item sets mining extends frequent pattern mining to discover item sets in a transaction database with utility values above a given threshold. However, mining high utility item sets presents a greater challenge than frequent item set mining, since high utility item sets lack the anti-monotone property of frequent item sets. Transaction Weighted Utility (TWU) proposed recently by researchers has anti-monotone property, but it is an overestimate of item set utility and therefore leads to a larger search space. We propose an algorithm that uses TWU with pattern growth based on a compact utility pattern tree data structure. Our algorithm implements a parallel projection scheme to use disk storage when the main memory is inadequate for dealing with large datasets.

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. [9], The last years witnessed an explosive progress in networking, storage, and processing technologies resulting in an unprecedented amount of digitalization of data. Hence, there has been a considerable need for tools or techniques to delve and efficiently discover valuable, non-obvious information from large databases.

In this situation, data mining is an important research field which offers efficient solutions for such an extraction. Much research in data mining from large databases have focused on the discovery of frequent patterns which are then used to identify relationships between sets of items in a database, through for example association rule derivation. In practice, however, the number of frequently occurring patterns is very large, hampering their effective exploitation by the end-users. In this situation, many works have been interested in defining manageably-sized sets of patterns, called concise representations, from which redundant patterns can be regenerated.

III. PROBLEM DEFINITION

In data mining domain, initially frequent item set mining is basic research topic. But later frequent item set mining approach suffering from problems. Frequent Item set mining failed to satisfy the need of users who desire to find item set with high utilities from dataset like high profits.

To address these problems, utility mining is introduced in data mining. The item set utility represents the importance of that item set. Such utility is measured in terms of profit, quantity, cost, weight etc. information based on preferences of user. HUI (high utility item set) is known if it's utility greater than user specified threshold of minimum utility. There are many research methods presented on HUI, but still this method is suffered from various research problems as HUI is not easy task because of down-ward closure property not included in utility mining, and this makes HUI inefficient in terms of time, memory, poor mining performances for large datasets.

To address these issues recently concise representations for utility mining is proposed called close dþ high utility item set (CHUIs). This representation approach practically showing the massive reduction on total number of high utility item set. The problem with this approach that time required for processing large datasets and extracting the HUIs is again major concern.

DOI 10.17148/IJARCCE.2016.51167

Architecture:

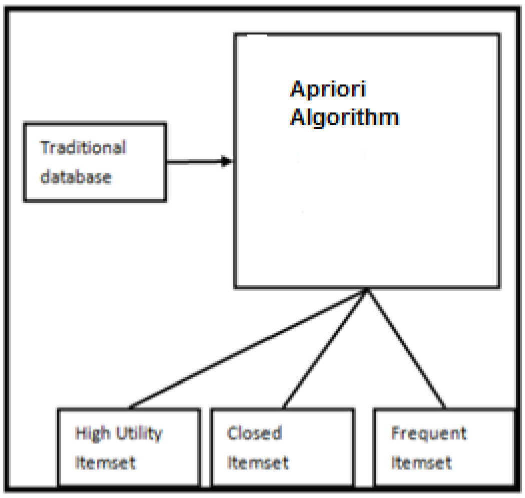

Fig.1. Architecture Diagram

IV. LIMITATIONS OF EXISTING WORK

- Existing HUI methods does not use down-ward closure property
- Existing HUI methods frequently present a large number of high utility item sets to users which makes difficult to end user to comprehend the results.
- Existing algorithms are inefficient in terms of time and memory requirement, or even run out of memory.

The recent efficient HUI representation approach is overcoming previous limitations but still time improvement is major research challenge.

V. WORK DONE

In this section we are discussing the practical environment, scenarios, performance metrics used etc.
Input either Mushroom or food mart or BMSWebView1 datasets.

Hardware and Software Configuration

➢ **Software Requirements**

• Front End	: Java
• Backend	: MySQL
• Tools Used	: JDK, My eclipse, NetBeans
• Operating System	: Windows 7 or onwards

➢ **Hardware Requirements**

• Processor	: Pentium Iv 2.6 GHz or onwards
• Ram	:2 GB
• Monitor	:15” Color
• Hard Disk	:20 Gb
• Keyboard	: Standard 102 Keys
• Mouse	:3 Buttons

VI. AIMS AND OBJECTIVES

The main aim of this project is to present the hybrid framework for HUIs from large dataset for improving the speed and efficiency of utility mining.

- To present literature review of existing methods of utility mining.
- To present limitations of existing techniques.
- To present proposed algorithms and framework.
- To present practical analysis and performance evaluation.

VII. CONCLUSION

In this paper, we proposed three efficient algorithms named Apriori HC (Apriori-based approach for mining High Utility Closed item set), Apriori HC-D (Apriori HC algorithm with Discarding unpromising and isolated items) and CHUID (Closed þ High Utility item set Discovery). for mining frequent item set. In FIM, to reduce the computational cost of the mining task and present fewer but more important patterns to users, many studies focused on developing concise representations, such as free sets, non-derivable sets, odds ratio patterns, disjunctive closed item sets, maximal item sets and closed item sets. These representations successfully reduce the number of item sets found, but they are developed for FIM instead of HUI mining. In future work we are presenting the hybrid novel framework of HUIs with goal achieving the high efficiency for the mining task and provide a concise mining result to users using parallel data computing technology to process large dataset fast.

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 item sets,” 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 item sets,” 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 item sets from large datasets,” inProc. 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 item sets,” inProc. 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.

DOI 10.17148/IJARCCE.2016.51167

[10] J. Han, J. Pei, and Y. Yin, “Mining frequent patterns without candidate generation,” inProc. 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 item sets,” 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 item sets in data streams,” inProc. IEEE Int. Conf. Data Mining, 2008, pp. 881–886.

BIOGRAPHIES

Ashwini S. Bhosale is M.E candidate at university of Pune at P.K Technical Campus. Working as an assistant professor at AISSMS-IOIT, Pune. She received her BE in 2012 from Pune University Of Information Technology. Her research interest are in the areas of Data mining and cloud computing, with current focus identify encryption with outsource revocation in cloud computing.

Veena S. Bhende is completed M.Tech candidate from JNTU Hydrabad. Working as an assistant professor at AISSMS-IOIT, Pune. She received her BE in 2012 from Shivaji University Of Information Technology. Her research interest are in the areas of Data mining and cloud computing, with current focus identify encryption with outsource revocation in cloud computing.

DOI 10.17148/IJARCCE.2016.51167