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

IJCSMC, Vol. 3, Issue. 11, November 2014, pg.489 – 493

REVIEW ARTICLE

A Review on Efficient Mining of High Utility Itemsets from Transactional Database

Pranoti Meshram[1], Prof. Vikrant Chole[2]

[1]Department of C.S.E, GHRAET College, Nagpur University, Maharashtra India

[2]Department of C.S.E, GHRAET College, Nagpur University, Maharashtra India

[1] pnmeshram301@gmail.com; [2] vikrantchole@gmail.com

Abstract— Mining high utility itemsets from a transactional database refers to the discovery of itemsets with high utility like profits. Although a number of relevant approaches have been proposed in recent years, but they incur the problem of producing a large number of candidate itemsets for high utility itemsets. Such a large number of candidate itemsets degrades the mining performance in terms of execution time and space requirement. The situation may become worse when the database contains lots of long transactions or long high utility itemsets. An emerging topic in the field of data mining is utility mining which not only considers the frequency of the itemsets but also considers the utility associated with the itemsets. The term utility refers to the importance or the usefulness of the appearance of the itemset in transactions quantified in terms like profit, sales or any other user preferences. The main objective of High Utility Itemset Mining is to identify itemsets that have utility values above a given utility threshold. Thus Utility mining plays an important role in many real-time applications and is an important research topic in data mining system to find the itemsets with high profit. In this paper we present a literature review of the present state of research and the various algorithms for high utility itemset mining.

Keywords:— Data mining,Candidate itemsets , high utility itemset, utility mining

I. INTRODUCTION

A. Data Mining

Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Knowledge Discovery in Databases (KDD) is the non-trivial process of identifying valid, previously unknown and potentially useful patterns in data. These patterns are used to make predictions or classifications about new data, explain existing data, summarize the contents of a large database to support decision making and provide graphical data visualization to aid humans in discovering deeper patterns. Discovering useful patterns hidden in a database plays an essential role in several data mining tasks, such as frequent pattern mining, weighted frequent pattern mining, and high utility pattern mining. Among them, frequent pattern mining is a fundamental research topic that has been applied to different kinds of databases,

such as transactional databases, streaming databases, and time series databases, and various application domains, such as bioinformatics, Web click-stream analysis, and mobile environments.

The basic goal of frequent itemset mining is to identify all frequent itemsets In past to find this frequent itemsets the generations of association rules and Apriori algorithm was used, once the frequent itemsets are identified and producing the itemsets with candidate and without candidates. But it is not producing the customer requirement like profit, sales in particular item. The unit profits and purchased quantities of items are not considered in the framework of mining frequent itemset. Hence, it cannot satisfy the requirement of the user who is interested in discovering the itemsets with high sales profits .Thus Mining high utility itemsets from databases refers to finding the itemsets with high profits.

B. Utility Mining

The traditional ARM approaches consider the utility of the items by its presence in the transaction set. The frequency of itemset is not sufficient to reflect the actual utility of an itemset. For example, the sales manager may not be interested in frequent itemsets that do not generate significant profit. Recently, one of the most challenging data mining tasks is the mining of high utility itemsets efficiently.

In view of this, utility mining emerges as an important topic in data mining for discovering the itemsets with high utility like profits.

The basic meaning of utility is the importance or profitability of items to the users. The utility of items in a transactional database consists of two aspects:

1. External utility: The importance of distinct items, which is called external utility.

2. Internal Utility: The importance of the items in the transaction, which is called internal utility.

Identification of the itemsets with high utilities is called as Utility Mining. The utility can be measured in terms of cost, quantity, profit and user preference .for this Utility mining model was proposed to define the utility of itemset. In this model by considering u(X) as a utility of an itemset X, which is the sum of the all utilities of itemset X in all the transactions containing X. then an itemset X is called a high utility items if its utility greater or equal to user- defined minimum utility threshresold.

The below diagram depicts the complete chain process of calculating and displaying the high utility itemsets. In this comparing with threshold value gives the High utility item sets as the results.

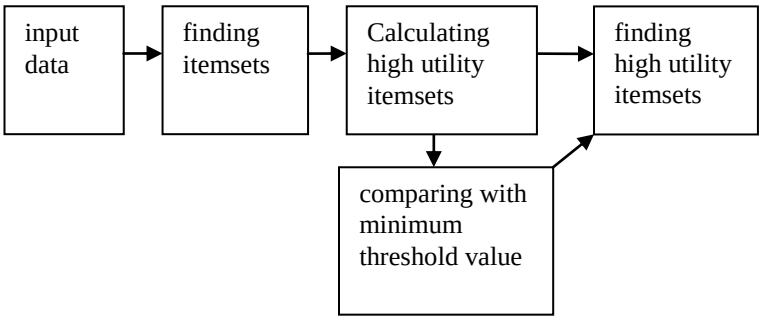

Fig 1: Data Flow Diagram

The main objective of high-utility itemset mining is to find all those itemsets having utility greater or equal to user-defined minimum utility threshold. In this paper we are presenting the literature survey study over the concept of high utility itemset mining using the concepts of data mining. In section II we are presenting the existing system of mining frequent itemset from transactional database. In section III we are presenting the Related Work done by various researches which describe various algorithm to find out high utility itemsets from transactional database.

II. EXISTING WORK

This paper provide application spectrum is wide in many real-time applications and is an important research issue in data mining area. Utility mining emerges as an important research topic in data mining field. Here high utility item sets mining refers to importance or profitability of an item to users. Number of algorithms like Apriori , FP growth has been proposed in this area, but they cause the problem of generating a large number of candidate itemsets. That will lead to high requirement of space and time and so that performance will be less and it is not at all good when the database contains transactions having long size or high utility itemsets which also having long size.

Existing studies [2] applied overestimated methods to facilitate the performance of utility mining. In these methods, potential high utility itemsets (PHUIs) are found first, and then an additional database scan is performed for identifying their utilities. However, drawback of existing methods is that it generate a huge set of PHUIs and their mining performance is degraded consequently. This situation may become worse when databases contain many long transactions or low thresholds are set. Thus the huge number of PHUIs forms a challenging problem to the mining performance since the more PHUIs the algorithm generates, the higher processing time it consumes.

III. RELATED WORK

A brief overview of various algorithms, Mining Frequent pattern defined in different research papers have been given in this section which is as follows:

A. Fast Algorithms for Mining Association Rules

R. Agrawal et al in [3] proposed Apriori algorithm, it is used to obtain frequent itemsets from the database. In Apriori algorithm there are two processes involve to find out all larger itemsets from the database . In the first Process step simply counts item occurrences to further determine the large one itemsets. for this it First generates the candidate sequences and then it chooses the large sequences from the candidate ones. Next, the database scan is performed to count the support of candidates itemsets. Then second process step was performed which involves generating association rules from frequent itemsets. After identifying the large itemsets, only those itemsets are allowed which have the support greater than the minimum support allowed. but disadvantage of using Apriori Algorithm is that it generates lot of candidate item sets and scans database every time and when a new transaction is added to the database then it should rescan the entire database again.

B. Mining Frequent Pattern without Candidate generation.

Mining frequent patterns in transaction and many other kinds of databases has been popularly important research in data mining . The previous studies adopt an Apriori-like candidate set generation-and-test approach. However, candidate set generation is still costly, especially when there exist long patterns. for this J. Han et al in [4] proposed a novel method of frequent pattern tree (FP-tree) structure, an extended prefix tree structure for storing crucial information about frequent patterns into compressed structure and develop an efficient FP-tree based mining method is Frequent pattern tree structure. Pattern fragment growth mines the complete set of frequent patterns using the FP-growth. It constructs a highly compact FP-tree, which is usually substantially smaller than the original database, by which costly database scans are saved in the subsequent mining processes. It applies a pattern growth method which avoids costly candidate generation. but FP-Growth Consumes more memory and performs badly with long pattern data sets. Thus it is not able to find high utility itemsets.

C. Mining Association Rules with Weighted Items

W. Wang et al in [5] proposed weighted association rule. This method extends the traditional association rule problem by allowing a weight to be associated with each item in a transaction, to reflect intensity of the item within the transaction. This provides us in turn with an opportunity to associate a weight parameter with each item in the resulting association rule. We call it weighted association rule (WAR). In WAR, we use a twofold approach. First it generates frequent itemsets. In second for each frequent itemset the WAR finds that meet the support, confidence. However, the weighted association rules does not hold downward closure property, mining performance cannot be improved.

D. Two Phase Algorithm

To address the above problem Liu et al. proposed [6] an algorithm named Two-Phase algorithm to efficiently prune down the number of candidates and can precisely obtain the complete set of high utility itemsets. In the first phase, a model that applies the transaction-weighted downward closure property (TWDC) on the search space to expedite the identification of candidates. In the second phase, one extra database scan is performed to identify the high utility itemsets. It performs very efficiently in terms of speed and memory cost. Although two-phase algorithm reduces search space by using TWDC property but it still generates too many candidates to obtain HTWUIs and requires multiple database scans.

E. Isolated Items Discarding Strategy for Discovering High Utility Itemsets

Traditional methods of association rule mining consider the appearance of an item in a transaction, whether or not it is purchased, as a binary variable. However, customers may purchase more than one of the same item, and the unit cost may vary among items. Utility mining, a generalized form of the share mining model, attempts to overcome this problem. Since the Apriori pruning strategy cannot identify high utility itemsets, developing an efficient algorithm is crucial for utility mining. To overcome this problem, Li et al. [7] proposed an isolated items discarding strategy (IIDS) to reduce the number of candidates. By pruning isolated items during level-wise search, the number of candidate itemsets for HTWUIs in phase one can be reduced. However, this algorithm still scans database for several times and uses a candidate generation-and-test scheme to find high utility itemsets and thus cannot improved performance.

F. Efficient Tree Structures for High Utility Pattern Mining in Incremental Databases

Ahmed et al. [8] proposed a tree-based algorithm, named IHUP. A tree based structure called IHUP-Tree is used to maintain the information about itemsets and their utilities. Although IHUP achieves a better performance than IIDS and Two-Phase, it still produces too many HTWUIs in phase one. Since the overestimated utility calculated by TWU is too large. Such a large number of HTWUIs will degrade the mining performance in phase one substantially in terms of execution time and memory consumption. Moreover, the number of HTWUIs in phase one also affects the performance of phase second due to more execution time required for identifying high utility itemsets .

IV. COMPARATIVE ANALYSIS

Sr. No	Author name	Algorithm used	Features	Problem
I	R. Agrawal and R. Srikant	Apriori Algorithm	Frequent itemsets and candidate generation	Rescan database every time and lot of candidates generated .
II	J. Han, J. Pei, and Y. Yin	FP-growth	finds frequent itemsets without generating any candidate itemset.	Consumes more memory and performs badly with long pattern data sets.
III	W. Wang, J. Yang, and P. Yu	weighted association rule	first propose concept of weighted items and weighted association rules.	does not hold downward closure property, mining performance cannot be improved.
IV	W. Wang, J. Yang, and P. Yu,	Two-Phase algorithm	performs very efficiently in terms of speed and memory cost.	Generate too many candidates to obtain HTWUI , require multiple database scan.
V	Y.-C. Li, J.-S.Yeh, and C.-C. Chang	isolated items discarding strategy (IIDS)	reduce candidates and to improve performance.	This algorithm still scan database for several times.
VI	C.F.Ahmed, S.K.Tanbeer, B.-S. Jeong, and Y.-K. Lee	a tree-based algorithm, named IHUP.	Maintain the information about itemsets and their utilities in the form of tree structure.	It generates huge set of PHUIs. Their mining performance is degraded consequently.

V. CONCLUSION

In this paper we have presented a brief overview of various algorithms for finding high utility itemset mining. In Data Mining, Association Rule Mining is one of the most important tasks. A large number of efficient algorithms are available for association rule mining, which considers mining of frequent itemsets. But an emerging topic in Data Mining is Utility Mining, which incorporates utility considerations during itemset mining. Utility Mining covers all aspects of economic utility in data mining and helps in detection of itemset having high utility. High Utility itemset mining is very beneficial in several real-life applications.

REFERENCES

[1] Sadak Murali & Kolla Morarjee, "*A Novel Mining Algorithm for High Utility Itemsets from Transactional Databases*," Global Journal of Computer Science and Technology Software & Data Engineering Volume 13 Year 2013

[2] Vincent S. Tseng, Bai-En Shie, Cheng-Wei Wu, and Philip S. Yu, Fellow," *Efficient Algorithms for Mining High Utility Itemsets from Transactional Databases*" IEEE Trans. Knowledge and Data Eng., VOL. 25, NO. 8, AUGUST 2013.

[3] R. Agrawal and R. Srikant, "*Fast Algorithms for Mining Association Rules*," Proc. 20th Int'l Conf. Very Large Data Bases(VLDB), pp. 487-499, 1994.

[4] J. Han, J. Pei, and Y. Yin, "*Mining Frequent Patterns without Candidate Generation*," Proc. ACM-SIGMOD Int'l Conf. Management of Data, pp. 1-12, 2000.

[5] W. Wang, J. Yang, and P. Yu, "*Efficient Mining of Weighted Association Rules (WAR*)," Proc. ACM SIGKDD Conf. Knowledge Discovery and Data Mining (KDD '00), pp. 270-274, 2000.

[6] Y. Liu, W. Liao, and A. Choudhary, "*A Fast High Utility Itemsets Mining Algorithm*," Proc. Utility-Based Data Mining Workshop,2005.

[7] Y.-C. Li, J.-S.Yeh, and C.-C. Chang, "*Isolated Items Discarding Strategy for Discovering High Utility Itemsets*," Data and Knowledge Eng., vol. 64, no. 1, pp. 198-217, Jan. 2008.

[8] 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. Knowledge and Data Eng., vol. 21, no. 12,pp. 1708-1721, Dec. 2009.

[9] Erwin, R.P. Gopalan, and N.R. Achuthan, "*Efficient Mining of High Utility Itemsets from Large Data Sets*," Proc. 12th Pacific-Asia Conf. Advances in Knowledge Discovery and Data Mining (PAKDD), pp. 554-561, 2008.

[10] H.F.Li, H.Y. Huang , Y.Cheng Chen, y. Liu, S.Lee, "Fast *and memory efficient mining of high utility itemsets in data streams*", Eight International Conference of Data Mining 2008.

[11] J.Pillai , O.P.Vyas ,"*Overview of itemset utility mining and its applications*" ,International Journal of Computer Applications (0975-8887), Volume 5-No.11 ,August 2010.

[12] H.Yao, H.J.Hamilton , "*Mining itemset utilities from transaction databases*", in Data and Knowledge Engineering 59(2006) pp.603-626.

