IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 16, Issue 2, Ver. I (Mar-Apr. 2014), PP 91-93
www.iosrjournals.org

Survey On: Mining High Utility Itemsets from Transactional Database

[1]Vinutha C, [2]Yogish H.K
IV Sem, Mtech, CSE department, East West Institute of Technology, Bangalore
Associate Professor, CSE department, East West Institute of Technology, Bangalore

***Abstract**: Utility mining emerges as an important topic in data mining field. Mining high utility itemsets from databases refers to finding the itemsets with high profits. The meaning of itemset utility is interestingness, importance, or profitability of an item to users. Mining high utility itemsets from a transactional database refers to the discovery of itemsets with high utility like profits. Although several relevant algorithm has been proposed in recent years, 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. This research work focuses on efficient method for high utility itemset mining, to reduce number of overestimated itemsets and to reduce the search space in order to improve the performance of high utility itemset mining.*
***Key Terms:** Candidate itemset, frequent itemset, high utility itemset, utility mining, data mining.*

I. Introduction

Now a days computers are used widely in different areas. Fast reliable and unlimited secondary storage provides a perfect environment for the users to collect and store large amount of the data. Computers are also used to extract the useful information from the mass of data. This is called as the knowledge discovery or Data mining.

Data mining is the process of revealing previously unknown and potentially useful information from large databases. The primary goal is to discover the hidden patterns in the data and to extract previously unknown interesting patterns. Discovering useful patterns hidden in a database plays an essential role in several data mining tasks, like frequent pattern mining and high utility pattern mining.

Among them the fundamental research topic is frequent pattern mining which can be applied to different kinds of databases such as transactional database, streaming database and time series database and many applications domains like bioinformatics, web click-stream analysis and mobile environment.

A. Frequent Itemset Mining

Frequent itemsets are the items that appear frequently in the transactions. The main goal of frequent itemset mining is to identify all the itemsets in the transaction data set, which are frequently purchased. Item sets are defined as a non empty set of items. If itemset is with k-different items is termed as a k-itemset. For ex{bread, butter, milk} may denoted as a 3-itemset in a supermarket transaction[1].

Let I= {i…..} be a set of items and D be a transaction database{…….} where each transaction T€D is a subset of I. The support or frequency of a pattern X{} is the number of transaction contained the pattern in transactional database.

The Apriori [1], algorithm is the initial solution for the frequent pattern mining problem.To overcomes the problems of Aprori, which generates more candidate sets and require more scans of database FP-Growth has been proposed [2]. Uses FP-Tree data structure without any candidate generation and using only two database scans. In the framework of frequent itemsets mining the importance of an item are not considered.

B. Weighted Frequent Itemset Mining

The weight of a pattern p is the ratio of the sum of all its weight to the length of p. The relative importance of an item is not considered in case of frequent itemset mining. The weighted association rule mining is introduced to address this problem. Here weights are given to items to represent the importance of an item to the users. Where this weight indicates the importance of itemset [3]. Since this framework does not support downward closure property mining performance cannot be improved. The problem of invalidation of downward closure property can be solved by Tae et al who proposed the improved model of weighted support measurement and exploit a weighted downward closure property. Weighted support is the fraction of the weight of the transaction that contains both A&B relative to weight of all transaction, where A&B are non empty subset [4].

C. High Utitlity Itemset Mining

Utility mining emerges as an important topic in data mining field. Mining high utility itemsets from databases refers to finding the items with high profits [6]. The meaning of an itemset utility is interestingness, importance or profitability of an item to users.

Utility of an itemset is defined as the product of its external utility and its internal utility, where the importance of distinct item is called as external utility and the importance of items in transaction is called an internal utility. If the utility of an item is not less than a user-specified minimum utility threshold then it is called high utility itemset, otherwise is called low utility itemset. By utility mining, many important business area decisions like maximizing revenue or minimizing marketing or inventory cost can be considered and knowledge about itemsets/customers contributing to the majority of the profit can be discovered.

II. Terms And Definition

Let I={i_1, i_2, ……..i_m} be a set of items and D be a transaction database {T_1, T_2,……T_n}where each transaction $T_i \epsilon$ D is a subset of I.

DEFINITION 1: The internal utility or local transaction utility value $l(i_p,T_q)$, represents the quantity of item i_p in transaction T_q.

DEFINITION 2: The external utility $p(i_p)$ is the unit profit value of item i_p.

DEFILITION 3: Utility $u(i_p,T_q)$ is the quantitative measure of utility for item i_p in transaction T_q defined by $U(i_p,T_q)= l(i_p,T_q)$ x $p(i_p)$.

DEFINITION 4: The utility of an item X in transaction T_q, by $U(X,T_q)$ is defined by $U(X,T_q) = \sum_{ip\in X} U(ip, Tq)$ where X={ i_1, i_2, ……..i_k}is a k-itemset $X\subseteq T_q$, and $1 \leq k \leq m$.

DEFINITION 5: The utility of an item X is defined by $U(X) = \sum_{Tq\in D} \sum_{ip\in X} U(ip, Tq)$.

DEFINITION 6: The transaction utility (TU) of transaction T_q denoted as $TU(T_q)$ describes the total profit of that transaction and is defined by $TU(T_q) = \sum_{ip\in Tq} U(ip, Tq)$.

DEFINITION 7: The minimum utility threshold δ is given by the percentage of total transaction utility values of the database minutil = δ X $\sum_{Tq\in D} TU(Tq)$

DEFINITION 8: An itemset X is a high utility itemset if $u(X) \geq$ minutil.

DEFINITION 9: Transaction-weightd utilization of an itemset X, denoted by TWU(X) is the sum of the transaction utilities of all transactions containing X, $TWU(X) = \sum_{X\subseteq Tq\in D} TU(Tq)$. The downward closure property can be maintained using transaction-weighted utilization.

DEFINITION 10: X is a high transaction-weighted utilization[HTWU] itemset if $TWU(X) \geq$ minutil.

Property 1(Transaction weighted downward closure): For any itemset X if X is not a HTWUI any superset of X is a low utility itemset.

III. Related Work

Discovering a hidden pattern in a database plays an important 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 fundamental research topic, has been applied to different kinds of databases like transactional databases [1], streaming databases and time series databases. Among the issues of frequent pattern mining the most famous are association rule mining [1] and sequential pattern mining.

One of the well-known algorithm for mining association rules is Apriori [1] by considering the problem of discovering association rules from a large database they presented two new algorithm called AIS and SETM and combined to form a new algorithm called Apriori hybrid. Since Apriori generates more number of candidate keys and require more number of scans. A novel frequent pattern tree structure & FP-Growth mining algorithm is used [2], achieves a better performance than apriori, since the frequent pattern treats all the products uniformly, importance of item is not considered, and the weighted association rule mining is proposed. Where weight of an itemset reflects the importance of an items [3].

Since the framework of weighted association rules does not have downward closure property Tao et al [4] proposed the concept of weighted downward closure property. Even though weighted association rule mining considers the importance of an items the quantities of an items in the transactions are not yet considered so the issue of high utility itemset mining is proposed and many studies [5], have addressed this problem Liu et al proposed an algorithm named two phase [6] which is mainly composed of two phases to efficiently prune the number of candidate sets and to obtain high utility itemset.

In phase I, defines the transaction weighted utilization mining model where high transaction weighted utility itemsets are identified. In phase I it employs an Apriori based level-wise method to enumerate HTWUI's candidate itemsets with length k are generated from length k-1 HTWUI's and their TWU's are computed by scanning the database once in each pass. In phase II, HTWUI's that are high utility itemsets are identified with an additional database scan. Even though this two phase algorithm reduces the search space it generates too many candidates in phase I. In order to reduce the candidate sets in phase I they introduced a new strategy [7]

called isolated item discarding strategy. One of the disadvantage of this is it scans the database for several times. To efficiently generate a high utility itemsets in phase I and to avoid scanning database too many time a tree based algorithm named as IHUP Interactive high utility pattern [5] has been proposed. IHUP is a tree based structure which maintains information about structure which maintains information about itemsets and its utilities can be maintained. IHUP algorithm has 3 steps 1) Construction of IHUP-tree. 2) Generation of HTWUIs. 3) Identification of high utility itemsets.

Although IHUP is better than IIDS it produces same number of HTWUI as produces by two phase method. To increase the efficiency they proposed compressed utility pattern tree for mining high utility itemsets. A large number of HTWUIs will degrade the mining performance in phase I substantially in terms of execution time and memory consumption and also affects the performance of phase II. The more HTWUI's generates in phase I, the more execution time for identifying high utility itemsets it requires in phase II.

As stated above the number of generated HTWUI's is a critical issue for the performance of algorithms, so this study aims at reducing the itemset's overestimated utilities and by using several strategies the number of generated candidates can be highly reduced in phase I and highly utility itemsets can be identified more efficiently in phase II.

IV. Conclusion

The theoretical model and definitions of high utility pattern mining cannot maintain the downward closure property of Apriori. The Two-Phase algorithm was developed based on the definitions of to find high utility itemsets using the downward closure property. They have defined the transaction weighted utilization, and proved that the downward closure property maintained throughout. Utility mining is a comparatively new area of research and most of the literature work is focused towards reducing the search space while searching for the high utility itemsets.

The key contribution of this paper is to provide an efficient research method for high utility itemset mining from high transactional database. By using several strategies, to decrease overestimated utility and enhance the performance of utility mining especially when databases contain lots of long transactions or a low minimum utility threshold is used and focuses towards reducing the search space of high utility itemset mining.

References

[1] 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.

[2] 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.

[3] C.H. Cai, A.W.C. Fu, C.H. Cheng, and W.W. Kwong, "Mining Association Rules with Weighted Items," Proc. Int'l Database Eng. and Applications Symp. (IDEAS '98), pp. 68-77, 1998.

[4] F. Tao, F. Murtagh, and M. Farid, "Weighted Association Rule Mining Using Weighted Support and Significance Framework," Proc. ACM SIGKDD Conf. Knowledge Discovery and Data Mining (KDD '03), pp. 661-666, 2003.

[5] 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.

[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.