Review on Hiding the Sensitive High Utility Itemsets

S.Kannimuthu, PhD
Associate Professor,
Computer science and engineering,
Coimbatore institute of engineering and technology

S.Gunasekaran, PhD
Head of Department,
Computer science and engineering,
Coimbatore institute of engineering and technology

S.Roopa
PG scholar,
Computer science and engineering,
Coimbatore institute of engineering and technology

ABSTRACT

The Association Rule Mining is the traditional mining technique which identifies the frequent itemsets from the databases and this technique generates the rules by considering the each items. The traditional association rule mining fails to obtain the infrequent itemsets with higher profit. Since association rule mining technique treats all the items in the database equally by considering only the presence of items within the transaction. The above problem can be solved using the Utility Mining technique. The Utility Mining technique identifies the product combinations with high profit but low frequency itemsets in the transactional database. Hiding High Utility Itemsets (HUIs) is the main challenges faced in the utility mining. In this paper, a detailed discussion about the traditional Association Rule Mining and the various algorithms involved in Utility Mining is explained.

Keywords

Association rule mining, Data mining, High utility itemsets, utility mining.

1. INTRODUCTION

Data mining is the process of retrieving the useful information from the large database to aid the users for making the efficient decisions. The derived knowledge can be simply classified into association rules, sequential patterns, Classification, Clustering & Utility Mining. Among these, Association Rule Mining is the most commonly used technique to determine the relationship of the items that are purchased in the large transactional database. Data mining technique has many advantages like discovering the valuable and interesting itemsets from the large databases. The itemsets are the set of items in the transactional database. Data mining technique is performed in the large database since it contains more sensitive information that is to be concealed from the unauthorized users to do malpractice. Due to the development of electronic data in the government and in many organization they implicitly contains the confidential information which may leads to a privacy threats if they are accessed by the unauthorized users.

The most challenging tasks of the data mining is the process of mining high utility itemsets and hiding the sensitive high utility itemsets in an efficient manner. The itemsets are identified using the high utilities called as Utility Mining.

1.1 Association Rule Mining

Association Rule Mining (ARM) is the concept involved in the frequent itemset mining. ARM is the most widely used traditional technique in the data mining and in knowledge discovery and it has a enormous application in the business , inventory predictions, supply chain management, product marketing etc.., the ARM identifies the frequent itemsets from the transactional database and it generates the rules for the itemsets by considering the each item a single value.

The main objective of the ARM is to find the frequently occurring itemsets in the database. It finds all the itemsets frequency that is beyond the minimum support threshold and then the ARM generates the rules for the co-occurrence of the frequent itemsets.

The frequent itemsets are the itemsets that occurs frequently in the database with some profit. The frequent itemsets may or may not contribute to the high profit. The non-frequent itemsets may sometimes contribute to the higher profit.

The frequent itemset mining concentrates only on the itemsets which occurs frequently in the transaction database neither concentrating on the profit of the itemsets. For example, considering the itemsets such as ‘Printers’ and the ‘scanners’, the frequency or the occurrence of the printer in the transactional database be 8 and the occurrence of the scanners be 5 then the ARM determines only the itemsets which are high in the frequency, ie. Obviously the printers. The profit of the 8 printers may be of 30% and the profit of the 5 scanners be 50% but this profit is not considered in the frequent itemset mining only the higher frequency of the itemsets are considered.

1.2 Utility Mining

In the real world application, the transactions usually consist of quantities and the profit. The lesser quantity with higher profit may occur rarely in the database. The quantity with lower profit occurs frequently in the transaction database. In order to address the ARM problem of finding the quantity of item with higher profit in the transaction, a new research issue is considered called Utility Mining.

Utility Mining is a concept of identifying the high utilities. Utility Mining which not only considers the frequency utility of the itemsets but also it considers the utilities associated with the itemsets. The term utility refers to the usefulness of the appearance of the itemsets in the transactions. The objective of the Utility Mining is to identify the high utility itemsets by considering the quantifiers such as profit, quantity, cost or other user preferences. The high utility itemsets are the itemsets that have the utility value above the user-specified threshold value.

Utility Mining is the extension of the frequent itemsets mining, which mines or discovers the itemset that occurs frequently in the transaction database. The measures used for the utility of items are local transaction utility and external utility.

The local transaction utility of an item is directly obtained from the transaction datasets such as quantity of the items sold in a particular transaction. The external utility of an item is just like the user preferences, ie, it is given by the user like a profit value of that particular itemsets. These values are stored in a utility table or represented using the utility function.

The ratio of the utilities for an itemset in all the transactions that are higher or equal to that of the user specified threshold values are said to be the high utility itemsets.

2. RELATED WORKS

In this section, a brief review of the different algorithms, techniques, concepts and approaches that have been defined in various research journals and applications are discussed.

A Fast High Utility Itemsets Mining Algorithm using Two phase algorithm is proposed [2], to efficiently prune down the number of candidates and in order to precisely obtain the complete set of high utility itemsets. It applies the transaction weighted property in the first phase to identify the candidates and uses extra database scan in the second phase to identify the high utility itemsets. Algorithm is parallelized on shared memory multi-process architecture using Common Count Partitioned Database (CCPD) strategy [2].

Mining Itemset Utilities from transaction databases is proposed [3] for mining the itemsets. Generally the frequent itemset only reflects the statistical correlation between the items without specifying the semantic significance of the items. The author proposed a utility based itemset mining approach to overcome the above limitation, by allowing the user to quantify the preferences by considering the usefulness of the itemsets using utility values. The algorithms used here are the UMining algorithm and the UMining_H algorithm. The UMining algorithm is used for mining all high utility itemsets using some pruning strategy. UMining_H, a heuristic method for mining most high utility itemsets. The former algorithm may be preferable to the later because it guarantees the discovery of all high utility itemsets .

High-Utility Pattern Mining: A method for discover of high-utility item sets is proposed [4] for mining. An algorithm is proposed for the frequent item set mining that identifies the high utility itemset combinations. The goal is to find the segments of data, which is defined through the combinations of few items. The author tackles the problem of mining frequent combinations of items with the highest contribution to a particular business objective. The task is considered as a optimization problem and it is solved using the specialized partition trees called High-Yield Partition Trees. Binary partition tree is used to solve a problem of identifying significant defining patterns, where each tree node represents a particular pattern and includes all of its subscribing transactions.

An Efficient algorithm for mining temporal high utility itemsets from data streams is proposed by the author [5]. The temporal high utility itemsets are mined i.e., itemsets whose support is larger than a pre-specified threshold in current time window of the data stream. The algorithm Temporal High Utility Itemsets (THUI)-mine is proposed for mining efficiently and effectively by generating fewer candidate itemsets such that the execution time can be reduced. This algorithm is based on the principle of Two phase algorithm, extended it with the sliding window filtering (SWF) technique. Temporal mining problem can be decomposed into two procedures such as pre-processing procedure and incremental procedure. THUI mine is faster than the two phase by 2-10 times and the performance gain becomes more significant. The future work can be of extending the proposed concept to discover other interesting patterns in data streams like utility items with negative profit.

Mining Top-K High utility itemsets is proposed [7] as Mining high utility itemsets refers to the discovery of the itemsets with utilities higher than that of the user specified threshold value. Setting an appropriate minimum utility threshold is a difficult problem for the user, since if min_util value is too low then too many high utility itemsets will be generated. If min_util value is high then no utility itemsets will be generated." Top-K high utility itemset mining" is proposed to solve the above mentioned problem where K is the desired number of high utility itemsets that are to be mined. The challenges like anti-monotone property and the requirement of lossless results are solved. No pattern missing during the mining process and search space and the number of checked candidates are effectively reduced.

A Utility-based association rule mining, a marketing solution for cross selling is proposed [8], a utility-based association rule mining method that evaluates the association rules by measuring the specific business benefits accuring to firms. The three key elements are identified such as opportunity, effectiveness, and probability to define and operate using the user preference as a utility function. Frequent itemset mining focuses only on the occurrence of frequency of itemsets with little references to the semantic significance to the user. To address this problem, HUIM is validated through market basket analysis. Pruning strategies were involved to facilitate efficient discovery of high utility itemsets from large volumes of data.

Research article-Reducing side effects of hiding sensitive itemsets in privacy preserving data mining is proposed [9], as Hiding missing artificial utility (HMAU) algorithm is proposed to hide the sensitive itemsets through transaction deletion technique. Transaction deletion are evaluated by concerning three dimensions such as hiding failure dimension, missing itemset dimension and artificial itemset dimension. The transaction with the sensitive itemsets are first evaluated by the designed algorithm to find the minimal HMAU values among transactions, so the transactions with minimal HMAU value are directly deleted from the database.

On-shelf utility mining with negative item values is proposed [10] for mining items. On-shelf mining not only considers profits and quantities of items in transactions but also their on-shelf time periods in stores. The author proposed an efficient three-scan mining approach to efficiently find high on-shelf utility itemsets with negative profit values from the temporal databases. Effective itemset generation method is developed to avoid generating a large number of redundant candidates and to effectively reduce the number of data scans during mining process. Proposed TS-HOUN approach carry out only three data scans to finish the problem of on-shelf utility mining with negative item values. This algorithm is higher than the conventional TP-HOUN algorithm since the conventional algorithm adopts the level wise technique to handle the problem of on shelf utility mining with the negative item profits.

An Incremental mining algorithm for high utility itemsets is proposed [12] as in the past, Two phase algorithm was designed for fast discovering high utility itemsets from databases. If a data comes intermittently, the approach needs to process all the transaction in the batch way, this problem is handled by proposing an incremental mining algorithm for efficiently mining high utility itemsets. It is based on the concept of fast-update approach. Experimental results show that the performance of the proposed algorithm executes faster than the two-phase algorithm in the intermittent data environment.

3. LITERATURE REVIEW

A Modified Hiding High Utility Item First algorithm (HHUIF) with item selector (MHIS) is proposed [1] for hiding sensitive itemsets. The MHIS algorithm computes the sensitive itemsets by utilizing the user defined utility threshold values. In order to hide the sensitive itemsets, the frequency value of the items is changed. The main objective of the MHIS algorithm is to select the best items from the sensitive itemsets having same high utility value from the database. By modifying the frequency values of items in the sensitive itemsets based on the minimum utility threshold value, the high utility value itemset are hidden. The proposed algorithm reduces the computational complexity as well as improves the hiding performance of the itemsets.

In the structure of the proposed algorithm, transactional database is used from which high utility itemsets are found using the utility mining algorithms. It discovers the sensitive itemsets based on threshold function. On applying MHIS algorithm on sensitive itemsets, the user can acquire the sanitized database. See figure 1 for the basic structure of hiding the sensitive high utility itemsets.

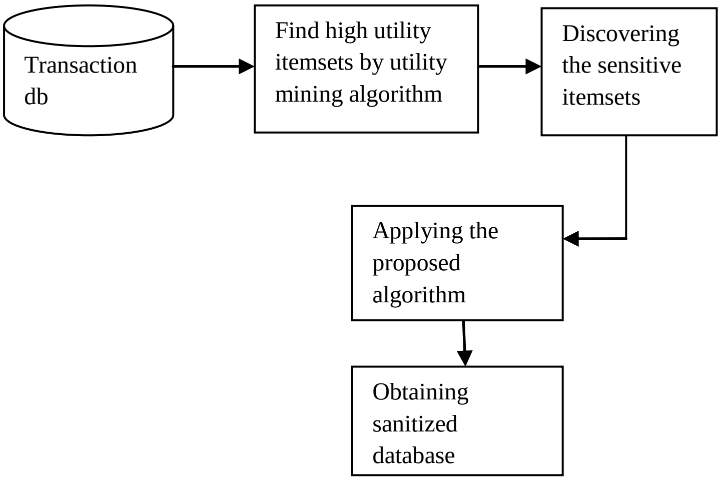
Transaction db
Find high utility itemsets by utility mining algorithm
Discovering the sensitive itemsets
Applying the proposed algorithm
Obtaining sanitized database

Figure 1: Basic structure of hiding the sensitive high utility mining

Novel algorithm for privacy preserving utility mining is proposed [6] for preserving the privacy. A tradeoff between the privacy protection and knowledge discovery in the sharing process is an important issue. The study focuses on privacy preserving utility mining (PPUM) and presents two novel algorithms HHUIF Hiding High Utility Itemsets First (HHUIF) and Maximum Sensitive Itemsets Conflict First (MSICF) to achieve the goal of hiding sensitive itemsets so that the adversaries cannot mine them from the modified database. The main goal of HHUIF is to decrease the utility value of each sensitive itemset by modifying the quantity values of items contained in the sensitive itemset. The goal of MSICF is to reduce the number of modified items from the original database, it selects an appropriate item which has the maximum conflict count among items in the sensitive itemsets. These algorithms reduce the impact on the source database of privacy preserving utility mining.

A GA-based approach to hide sensitive high utility itemsets is proposed [11] for hiding the sensitive itemsets. Genetic algorithm is used in many research issues since they could provide feasible solutions in a limited amount of time. A GA based privacy preserving utility mining method is proposed to find appropriate transactions to be inserted into the database for hiding sensitive high utility itemsets. Privacy preserving factors such as quantities and profits are considered. The operations in the GA such as selection, crossover and mutation are performed and there occurs the changes in the original database. But privacy preserving utility mining using GA does not change the original database. The side effects like hiding failures, missing costs and artificial costs are considered during the hiding process. The proposed algorithm provides less information loss and protects high risk information in the database.

4. PERFORMANCE ANALYSIS

The performance analysis of the algorithms is carried out using the parameters such as Hiding failure, Miss cost and the Dissimilarity. From the above analysis of the algorithms the table 1, shows that the GA based HSHUI algorithm has less percentage of hiding failure with the specified minimum utility threshold values and it also shows that the algorithm GA-based HSHUI has the less percentage of miss cost when compared to that of the other algorithms. The above table shows the dissimilarity percentage of the different algorithms.

4. CONCLUSION

Hiding the sensitive high utility itemsets from the unauthorized users plays a major role in the utility mining. The challenges faced in the utility mining like preserving the privacy by hiding the highly sensitive high utility itemsets are analyzed with different algorithms. The above reviews show that the process of hiding the sensitive high utility itemsets has some original information loss in the transactional database. The problem has to be addressed by identifying a new algorithm for preserving the privacy to the transactional database. In future, algorithm which provides privacy to the highly sensitive high itemsets are to be developed in order to avoid the unauthorized access.

Table 1: Analysis of Algorithm in Terms of Hiding Failure, Miss Cost and Dissimilarity

S.no	Algorithm	Dataset used	No. Of transactions	No. of distinct items	Minimum utility threshold (α)	Hiding failure (%)	Miss cost (%/)	Dissimilarity (%)
1	HHUIF [1]	Sample Dataset	100	10	2500 3000 3500	11 23 7	10 13 17	1 1 5
2	MHIS [1]	Sample Dataset	100	10	2500 3000 3500	7 0 0	0 18 19	3 2.5 4.5
3	GA-based HSHUI [11]	Simulation dataset Food mart dataset	1 to 5 21,556	0.01 to 10 1,559	0.005 0.01 0.015 0.005 0.01 0.015	2 4 6 1.4 1.6 1.8	2 7 8 1.6 1.9 2	Not available
4	HHUIF [6]	Data.ntrans1. nitems0.1	1000	100	2000< min util 3000 4000>3000 3000=3000	100 0 0	0 62.04 0.93	1.93 0.81 2.95
5	MSICF [6]	Data.ntrans1. nitems0.1	1000	100	4000> min util 3000 3000=3000 2000<3000	0 0 100	70.37 12.96 0	1.93 0.81 2.54

5. REFERENCES

[1] Rajalakshmi selvaraj and Venu Madhav Kuthadi, 2013 A Modified Hiding High Utility Item First Algorithm(HHUIF) with Item Selector(MHIS) for Hiding Sensitive Itemsets in International Journal of Innovative Computing and Control.

[2] Ying Liu, wei-Keng Liao, Alok Choudhary, 2005 A Fast Itemsets Mining Algorithm, in UBDM'05, Chicago ,USA .

[3] Hong Yao, Howard J.Hamilton, 2005 Mining itemset utilities from transaction database, in ELSEVIER .

[4] Jianying Hu, Aleksandra Mojsilovic, 2007 High-utility pattern mining: A method for discovery of high-utility item sets, in pattern recognition ELSEVIER .

[5] Chung-jung Chu, Vincent S.Tseng, Tyne Liang, 2008 An efficient algorithm for mining temporal high utility itemsets from data streams, in science direct, the journal of system and software, ELSEVIER.

[6] Jieh-Shan Yeh, Po-Chiang Hsu, 2010 HHUIF and MSICF: Novel algorithms for privacy preserving utility mining, in expert systems with application, ELSEVIER .

[7] Cheng Wei Wu, Bai-En Shie, Philip S. Yu, Vincent S. Tseng, 2012 Mining Top-K High Utility Itemsets, KDD'12, Beijing, china.

[8] Dongwon Lee, Sung-Hyuk Park, Songchun Moon, 2012 Utility-based association rule mining: A marketing solution for cross-selling", in expert systems with applications ELSEVIER.

[9] Chun-Wei Lin, Tzung-Pei Hong, Hung-Chuan Hsu, 2014 Research article- Reducing side effects of hiding sensitive itemsets in privacy preserving data mining, in Hindawi publishing corporation The scientific world journal.

[10] Guo-cheng Lan, Tzung-pei Hong, Jen-peng Huang, Vincent S.Tseng, 2013 On-shelf utility mining with negative item values, in expert systems with applications, ELSEVIER .

[11] Chun-Wei Lin, Tzung-Pei Hong, Jia-Wei Wong, Guo-cheng Lan, wen-yang Lin, 2014 A GA-Based Approach to Hide Sensitive High Utility Itemsets, in Hindawi publishing corporation, the scientific world journal.

[12] Chun-Wei-Lin, Guo-Cheng Lan, Tzung-pei Hong, 2012 An incremental mining algorithm for high utility itemsets", in Expert system with applications ELSEVIER

[13] V. Ciriani, S. De Capitani di Vimercati, S. Foresti, and P. Samarati, 2008 K-Anonymous Data Mining: A survey in Springer US, Advances in Database Systems .