International Journal of Scientific Engineering and Applied Science (IJSEAS) – Volume-2, Issue-6,June 2016
ISSN: 2395-3470
www.ijseas.com

Mining Closed High Utility Itemsets using Secured Approach

Ms.Archana Kisan Dere
Second Year Master of Engineering
Department of Computer Engineering.
Sharadchandra Collage OfEngineering.,
Dumberwadi,Otur,Pune-(India).
Email:-archu.dere@gmail.com

Prof. S .A. Kahate
Assistant professor
Department Of computer Engineering.
Sharadchandra Collage Of Engineering.,
Dumberwadi,Otur,Pune-(India)
Email:-sandip.kahate@gmail.com

***Abstract* — Some itemset are useful in data mining for the market base analysis purpose. Current model of market base analysis produces low important itemsets and frequent lose selling value. To avoid this utility mining is essential. Extraction high itemset from database is very difficult task. So formulated HUI is compact in size that's why the efficiency will degraded of mining process. Some ideas are available to achieve the efficiency are free sets closed itemsets etc. Closed high utility itemset (CHUI) proposed to achieve this goal. To outsourced the party for mining of data. Data privacy problem occur because one party never trustworthiness on another party. Outsourcing of this task may have privacy leak problems. We proposed a system to handle the load of computation, storage and processing to another party with preservation of privacy of outsourced high utility mining and also to produce the concise representation of HUIs using existing work.**

Keywords — Compact representation of HUIs, Frequent Itemset Mining, HUIs, Utility mining.

I. Introduction

To get useful and important information for the knowledge the huge mining of data required. Some organization required such data for their purpose. Market basket analysis is popular method in the market is FIM. Some issues are in the FIM model it generated itemset having low importunacy item and low selling tags. It may generate high importunacy itemset. FIM possess same importunacy, weights to all itemsets. Utility mining is established to overcome this problem to reduce cost weight and height.

Some applications such as website click stream analysis, mobile commerce environment, biomedical applications etc has high utility mining. If the utility of an itemset is greater than user-specified minimum utility threshold then it is said to be high utility itemset otherwise it is considered as a low utility itemset. Large amount of high utility itemsets causes difficulty to user for result analysis of HUI. Require more memory and more processing time causes less efficiency. For more efficiency to reduce cost overhead and mining task and to provide concise representation, different approaches like Freeset, Non derivable sets, Closed Itemset are introduced in FIM. But applying these techniques to HUI produced several challenges:

1) How to recover all HUI's from the concise representation.

2) Lossy representation of all HUI which is not meaningful to user.

3) Algorithms may not be efficient.

4) Significant reduction in the extracted patterns may not be achieved

The concept of closed itemset into high utility itemsets mining to address mentioned challenges and named iii) Closed+ High Utility itemset Discovery proposed by algorithm. Information and storage require extra memory and processing them to get a CHUI is very difficult step. Organizations have huge data and tend to outsource the task of HUI mining to another party for the analysis. In given system outsourced party will have information and does not have privacy module to protect data privacy. Proposed system is to reduce the load of computation, storage and processing to another property with preservation of privacy of outsourced high utility mining.

II. Literature Riview

Itemset mining is popular application to generate frequently purchased itemsets in market basket analysis. But it can generate high amount of frequent itemsets if the data is highly correlated and set minimum support threshold is very low. Instead of mining all frequent items the solution is to construct concise representation of frequent itemsets. T. Calderset.al. [1] aimed to identify the redundancy of frequent itemsets to reduce the result of mining operation. Paper presented the deduction rules allows the minimal representation of all frequent itemsets. Non Derivable Itemsets considered for concise representation. Experiments showed that mining concise representation first and then from this creating frequent itemsets give better results than existing algorithms. High Utility Pattern mining has several applications in broader aspect.

U. Yun et.al.[2] To prune the search spaces i.e. removing the itemsets which are not satisfying certain condition many algorithms used a support constraint. This scheme allows for basic pruning but the resulting patterns have weak affinity after mining datasets for obtaining frequent patterns. This paper proposed an efficient mining algorithm called WIP (Weighted Interesting Pattern mining). The algorithm founded on mining weighted frequent patterns. This paper determine the concept of a weighted hyperclique pattern that uses a new measure, called weight-confidence, to consider weight affinity and prevent the generation of patterns with substantially different weight levels. Experimental study showed that WIP is efficient in weighted frequent pattern mining and it generates less but more worthful patterns for users.

W. Cheung et.al.[3] introduces Compressed and Arranged Transaction Sequences Tree or CATS Tree and CATS Tree algorithms. Once CATS Tree is built, it can be used for multiple frequent pattern mining with different supports. Furthermore, CATS Tree and CATS Tree algorithms allow single pass frequent pattern mining and transaction stream mining. In addition, transactions can be added to or removed from the tree at any time. CATS Tree extends the idea of FP-Tree to improve storage compression and allow frequent pattern mining without generation of candidate itemsets.

S.F. Shie et.al.[4] Some algorithms, for e.g. Apriori, suffer from the costs to handle a huge number of candidate sets and scan the database repeatedly. An algorithm, named FP-growth, for mining the complete set of frequent itemsets from FP-tree is developed. This approach avoids the costly generation of a large number of candidate sets and repeated database scans, which was regarded as the most efficient strategy for mining frequent itemsets. Updates to the transaction database could invalidate existing rules or introduce new rules. The problem of updating the association rules can be reduced to finding the new set of frequent itemsets in the updated database. A simple solution to the update problem was to re-mine the frequent itemsets of the whole updated database. However, it was clearly inefficient because all the computations done in the previous mining are wasted. This problem is avoided in proposed algorithm, called AFPIM(Adjusting FP-tree for Incremental Mining), to efficiently find new frequent itemsets with minimum re-computation.

Y.C. Li et.al.[5] The pruning method used in Apriori cannot be applicable and not able to identify high utility itemsets. This paper proposed the Isolated Items Discarding Strategy (IIDS), which can be applied to any existing level-wise utility mining method to reduce candidates and to improve performance. The most efficient known models for share mining are ShFSM and DCG, which also work adequately for utility mining as well. By applying IIDS to ShFSM and DCG, the two methods FUM and DCG+ were implemented, respectively. For both synthetic and real datasets, Results after the experiments showed that the performance of FUM and DCG+ is more efficient than that of ShFSM and DCG, respectively. Therefore, IIDS is an effective strategy for utility mining.

Erwin et.al. [6] used pattern growth approach (Han, Wang, & Yin, 2000) and proposed an efficient algorithm for utility mining. Revent research by authors (Erwin, Gopalan, & Achuthan, 2007) on utility mining had produced an algorithm that ran efficiently on dense data, but performed unsatisfactorily on sparse data. They felt the need to improve the performance on sparse data, motivated by this they have formulated a new compact data representation named Compressed Utility Pattern tree (CUP-tree) which extends the CFP-tree of (Y.G Sucahyo & R.P Gopalan, 2004) for utility mining, and a new algorithm named CTU-PRO for mining the complete set of high utility itemsets.

Jian Pei et.al.[7] In frequent pattern mining, due to numerous problems like, huge set of candidate sequence generation, number of repeated scans of the database and problem of Apriori while mining long sequential patterns, this paper proposed an efficient sequential pattern mining method named as Frequent pattern projected Sequential pattern mining i.e. FreeSpan. It integrates the mining of frequent sequences with frequent mining and in order to confine the search space used projected sequence database and it reduces the efforts of candidate subsequent generation.

R. C.W. Wonget.al.[8] considers the scenario of data stream. Data is continuously arriving through the data stream in terms of sliding window. To mine the K most interesting itemsets of varied sizes in a data stream the authors adopted new model using the sliding window concept. Sliding window is partitioned into buckets. For the transactions in each bucket, statistics of the frequency counts of the itemsets is maintained. Experimental studies showed that algorithm guarantees no false negatives for any data distributions and also show that the number of false positives returned is typically small according to Zipfian Distribution. Experiments on synthetic data showed that the memory used is tens of times smaller than that of a naive approach, and the false positives are negligible.

Vincent S et.al. [9] used closed itemset mining technique and proposed a novel solution called Closed High Utility Itemset Mining. Methodology in this paper is extends further to achieve privacy in the outsourced mining task.

III. System Design

System Modules

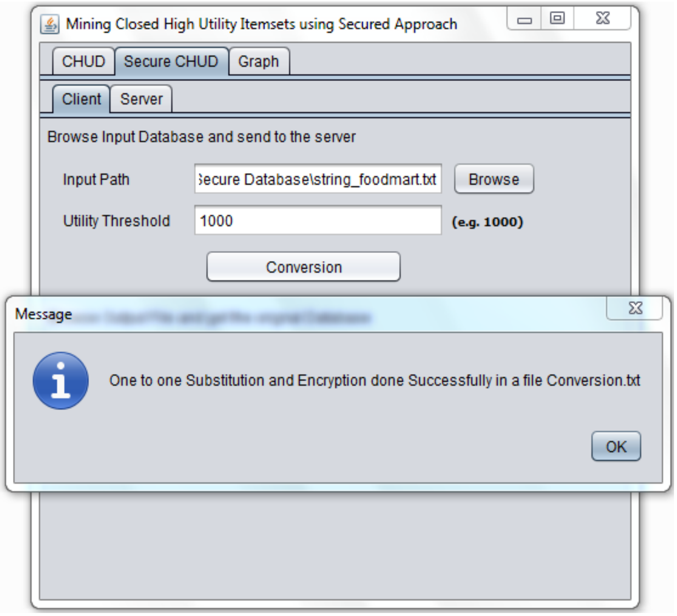
System aims at mining closed high utility Itemsets in privacy preserving manner. Mining concise representation of Closed High Utility Itemsets, methodology in [10] is used. Privacy in outsourced task is achieved by this system. Secured techniques are applied before the mining task and then outsourced the further process to another party. Our system contains following modules:

Module 1: Preprocessing:-

First, every item in the transaction table is substituted with its respective numerical Id. Each item in item set is assigned with numerical id randomly.

Output of this process is converted dataset D', Map of the items and their numerical ids. This map is kept securely at client side and converted dataset D' is forwarded to next step. For example Item I1 is converted to random numeric id 78 and so on for other items.

Then we applied Homomorphic encryption on weights of the data. Weights of the items are encrypted using homomorphic cryptosystem. The system is using homomorphic encryption because it allows mathematical operation on encrypted data and when results of this mathematical encryption are decrypted then decrypted results reflects the mathematical operation's effect.
Our system used one of the homomorphic cryptosystem called Pallier cryptosystem, which exhibits following properties:
i. Homomophic addition
Dsk (Epk (a+b)) = Dsk(Epk(a) * Epk(b) mod N^2)
ii. Homomorphic Mulitplication
Dsk(Epk (a*b)) = Dsk(Epk $(a)^b$ Mod N^2)
Where Epk is encryption function with Key public key Pk derived using N and g where N is product of two prime numbers of similar two lengths and g is generator in Z_{N2}. Also, let Dsk be the decryption function with secret key sk.
iii. Semantic Security: It is impossible to figure out information about plain text using cipher text.


Mining Closed High Utility Itemsets using Secured Approach

Browse Input Database and send to the server

Input Path: ;ecure Database\string_foodmart.txt

Utility Threshold: 1000 (e.g. 1000)

Message

One to one Substitution and Encryption done Successfully in a file Conversion.txt

Figure 1. GUI of converting database into secured form.
Results: Encrypted weights of the items, Converted Database.

This converted database and encrypted utility threshold given to third party for high utility mining.

To achieve CHUI mining task, the algorithm used is Closed High Utility Item sets Discovery

Module 2: Mining of CHUIs in secured way:-

A - Converts the database in vertical database and simultaneously calculates the utility for each transaction and also its transaction weighted utility of items. When transaction is fetched, its Tid and transaction utility are stored into global table named Global Utility Table (GUT).

B – In this algorithm process to scans database and collect promising items having estimated utility greater than abs_min_utility into ordered list which is sorted in increasing order of support. Then utilities of unpromising items are removed from Global Utility table (GUT).

C – In this step CHUD generates candidates in recursive manner starting from candidates containing a single promising item and recursively joining items to them to form larger candidates.

D - Here performs Subsume check on X (item) which verifies if there exists an item which is included in a closed item set that has already been found and supersets of X do not need to be implemented.

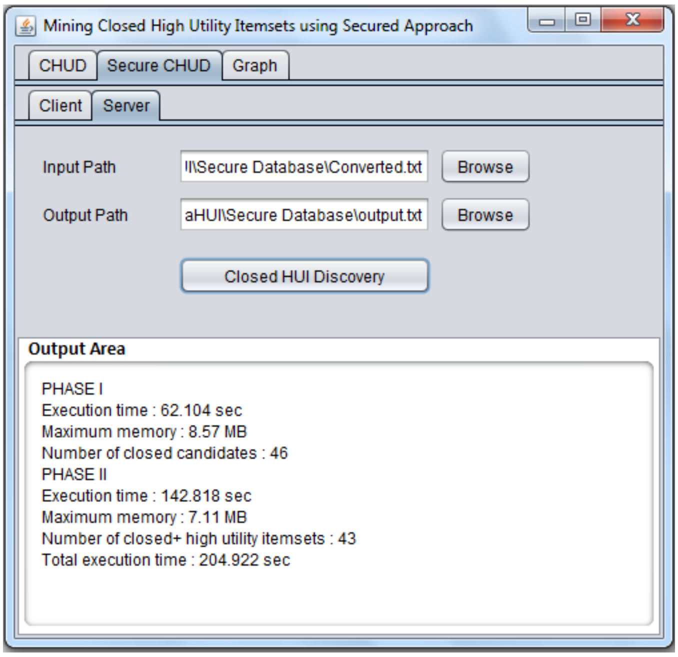
E – Then next computes the closure of an Xc = C(X) of X. Then the estimated utility is calculated.

F – After that DCM strategy (Discarding candidates with maximum utility less than minimum utility threshold) is applied i.e. it computes the maximum utility of Xc. It discards the candidate whose estimated or maximum utility is less abs_min_utility; otherwise Xc is outputted with its estimated utility.

G – Here a node N (Xc) is created and the procedure Explore is called for finding candidates that are supersets of Xc i.e. potential CHUIs. Then RML strategy is applied to remove minimum utility items from local transaction utility table.

H – This step consists of taking each candidate X and calculates its utility. Each candidate of low utility is discarded and candidate with high utility than absolute minimum utility is outputted.

Results: First the converts the database vertically and gives Global Utility Table of all promising items and then extracts candidate closed high utility itemsets and finally extracts the final closed+ high utility itemsets.



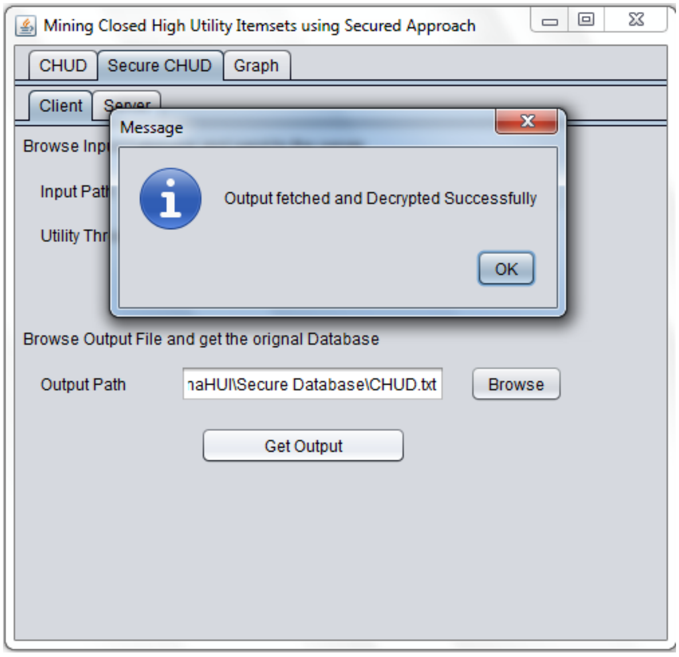
Input Path: ll\Secure Database\Converted.txt
Output Path: aHUI\Secure Database\output.txt

```
PHASE I
Execution time : 62.104 sec
Maximum memory : 8.57 MB
Number of closed candidates : 46
PHASE II
Execution time : 142.818 sec
Maximum memory : 7.11 MB
Number of closed+ high utility itemsets : 43
Total execution time : 204.922 sec
```

Figure 2. GUI of extraction of CHUIs.

Module 3 Conversion of encrypted output: -

This phase converts the encrypted output back to plain text readable to the client side which includes number of closed high utility itemsets, containing transaction and item utilities Complete set of CHUIs which is in substituted form, is obtained using locally kept map of items and its numerical Ids.



Output fetched and Decrypted Successfully

Browse Output File and get the orignal Database

Output Path: ıaHUI\Secure Database\CHUD.txt

Figure 3. GUI of converting output from server.

IV. Comparison of Existing and Proposed System

Existing system does the task of mining Closed High Utility Itemsets. There is no difference between number of extracted candidate CHUI and final CHUIs by existing and proposed system respectively. The main difference is, the CHUI mining task is outsourced to another party and is in secured way i.e. with privacy preservation. Though the items and utility information is given to another party the privacy is preserved. This is achieved by applying items substitution and homomorphic encryption technique.

Following figure show the time required for the existing system and proposed system for CHUI generation.

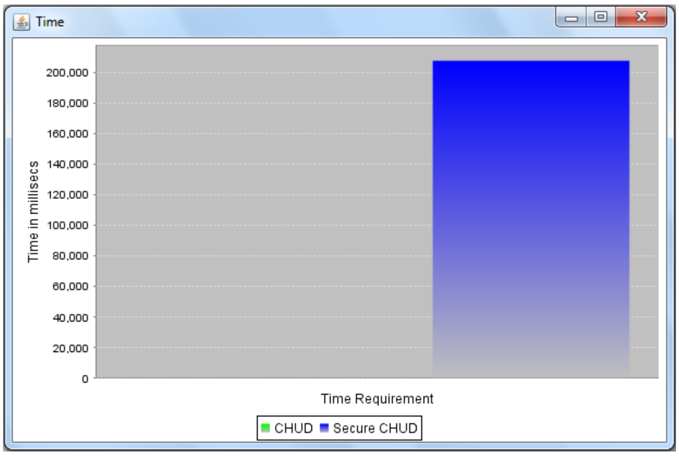


Fig. 1. GUI of comparison of time requiement of CHUD and Secure CHUD

This figure show the time required for the existing system and proposed system for CHUI generation.

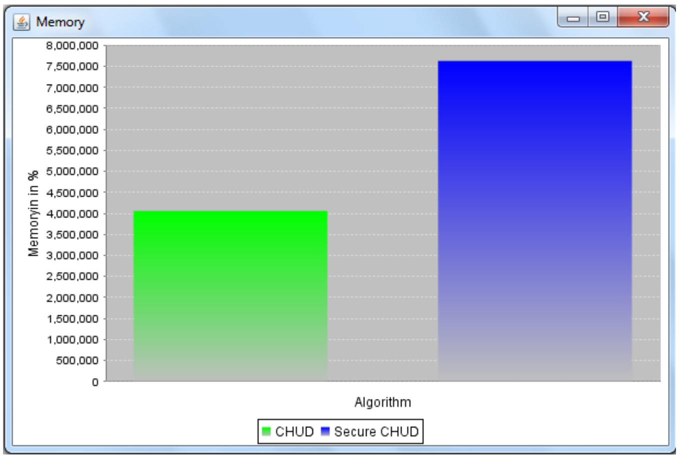


Fig. 2. GUI of comparison of time requiement of CHUD and Secure CHUD

This figure show the memory required for the existing system and proposed system for CHUI generation.

V. Conclusion

In this paper, we have discussed various techniques of generating HUIs and techniques which addressed the issues in high utility mining and also improve the efficiency and accuracy of process generating the set of HUIs. We discussed some compact representations available for HUIs. Amongst them Closed High Utility Mining proved better and efficient. But conducting all these processes, organizations have to carry the storage and computation overhead. We proposed a system to alleviate the process of generation of compact HUIs to another party and also fulfilled the aroused need of security and privacy of the data as outsourced party may not be trusted one.

VI. FUTURE SCOPE

In current CHUI discovery, the user gives the minimum utility threshold and with reference to that the CHUIs generated. It can be possible that the user may provide minimum utility threshold lesser and thus can be result in more number of and unnecessary high utility itemsets. Also the user may provide very high minimum utility threshold and fewer number of CHUIs may generated and important Itemsets may get ignored. In future work we can apply the technique which will select the optimum minimum utility threshold to avoid the mentioned problem.

REFERENCES

[1] 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.
[2] U. Yun, “Efficient Mining of Weighted Interesting Patterns with a Strong Weight and/or Support Affinity,” Information Sciences, vol. 177, pp. 3477-3499, 2007.
[3] W. Cheung and O.R. Zaı¨ane, “Incremental Mining of Frequent Patterns without Candidate Generation or Support Constraint,” Proc. Seventh Int’l Database Eng. and Applications Symp. (IDEAS ’03), pp. 111-116, 2003.
[4] J.L. Koh and S.F. Shie, “An Efficient Approach for Maintaining Association Rules Based on Adjusting FP-Tree Structures,” Proc. Ninth Int’l Conf. Database Systems for Advanced Applications (DASFAA ’04), pp. 417-424, 2004.
[5] 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, pp. 198-217, 2008.
[6] Erwin, A.Gopalan, R.P.Achuthan, N. R.: A Bottom-Up Projection Based Algorithm for Mining High Utility Itemsets. Proceedings of the 2nd International Workshop on Integrating Arti_cial Intelligence and Data Mining, Volume 84, Gold Coast, Australia, 2007.
[7] Jian Pei,Jiawei Han, “FreeSpan: Frequent pattern projected Sequential pattern mining. School of computing science, Simon Fraser University. 2001.
[8] R. C.W. Wong and A.W. Fu. Mining top-k itemsets over a sliding window based on zipfian distribution. In Proc. of SIAM SDM, 2005.
[9] Vincent S. Tseng, Cheng-Wei Wu, Philippe Fournier-Viger, and Philip S. Yu,”Efficient Algorithms for Mining the Concise and Lossless Representation of High Utility Itemset,” Knowl. And data Engg., vol. 27, no.3,2015.