International Journal of Research in Advent Technology, Vol.7, No.5, May 2019
E-ISSN: 2321-9637
Available online at www.ijrat.org

Comparison between High utility mining of rare item sets Algorithms over Transactional Database

Sunidhi Shrivastava[1], Neha Garg[2], Pankaj Gugnani[3]
Department of Computer Science and Applications [1,2,3],ITM university Gwalior[1,2,3]
Email:sunidhishrivastava5@gmail.com[1],nehagarg179@gmail.com,[2]pankajitmcse@gmail.com[3]

Abstract-Privacy Preservation of highly profitable itemsets is very important term in utility based data mining. In the process of mining profitable items from sales database whether they are frequent or non-frequent utility pattern. It is very important for making market strategies. In this paper, comparison between two algorithms is done based on their properties for analyzing high utility rare itemsets. High utility rare itemset algorithm executes on transactional database in which, two thresholds will be used for extracting High utility rare items. First thresholdused for discovering HUI and, another for extracting high utility rare itemsets. Second algorithm, privacy preservation of rare itesets using genetic algorithm used the process of privacy preservation using Genetic algorithm. HURI using GA approach is more beneficial when it comes to better security purpose. As a future work more optimization algorithm can be implemented for obtaining better result.
Keywords -Data mining; utility mining; rare itemsets; privacy preservation; Genetic algorithm.

1. INTRODUCTION

Data mining is the process of identifying interesting patterns and knowledge from huge amount of data. The process of discovering knowledge from data involves Data Cleaning, Data Integration, Data Transformation, Data Mining, Pattern Evaluation and Data Presentation. Mined data can be used in many applications like Production Control, Science Exploration [1][2].

Privacy preservation is an emerging field in the data mining ,how to keep confidential information private in case of making better market strategy, basically sanitized data base for protecting values and keep them safe from outside world. It can be considered as a strategy for marketing [3][4][5].

2. LITERATURE SURVEY

A brief numerous algorithm overview and methods defined in various research papers has been provided in this part.

R. Agrawaland A. Swamiin [6], proposed Apriori algorithm, through applying two thresholds called, minimum confidence and min. support.

H. Yao et al proposed in [7], the utility problem based approach in mining to determine the item sets that are significant according to their utility values. Two different new pruning schemes were presented to decrease the worth of discovery high profitable item set. With these pruning approch, to calculate its actual utility value a n -item set with a upper bound utilty, less than min_utility can be pruned instantly without go during the database. On the basis of these pruning strategies, UMining and UMining_H algorithms were introduced to provide effective solutions to itemset

Mining drawback mainly based upon utility. But, there are also some limitations first, since first depth search approaches such as FP growth have several advantages over level wise approaches. Second, the difficulty of how to classify high utility rules from the high profitable item set could be investigated.

J. Hu et al. in [8], proposed frequent item set mining algorithmic rule through that classifies high profitable item groupings. In compare to the traditional association rule and frequent item mining methods, the objective of the algorithm is to discover data segments, defined by few items (rules) groupings, which fulfill various situations present an effective estimate to solve it by particular partition trees, known as high yield partition trees and investigated the a diversity of splitting schemes performance.

Liu et al. proposed in [9], two different stage algorithms for discovering item sets those have higher utility. On the first stage, "transaction-weighted downward closures, property" applied on a model to accelerate the candidate's identification on the search space. In another stage, one additional database scan is the high profitable item sets identify performed. t and w-confidence are above few provide threshold.

Hua.Fu. Li et al. proposed in [10], specially two algorithms known as MHUI-TID and MHUI-BIT, for mining HUI in database. These two distinctive successful thing learning representation and an amplified lexographical tree-based rundown information structure is created to expand the mining high utility thing sets proficiency.

V.S. Tseng et al.proposed in [11], A new technique Temporal HUI (THUI), it's a time based high utility item sets mining from data streams. Temporal high utility item sets discovery is a significant mining interesting pattern, procedure for example from data streams association rules. The process of identify temporal high profitable itemset can be achieved easily to considering less memory and fewer candidate generation which cause less processing and

CPU time. So this algorithm satisfies the needs of time and space for mining data streams.

G.C.Lan et al. proposed a novel pattern type, known Rare Utility Item sets in [12], which consider not only individual profits and quantities but also usual current periods and items branches in a multi database atmosphere. A fresh advance of mining known as Two-Phase algorithm for mining Rare Utility Item sets in several Databases (TP-RUI-MD) was proposed to successfully find out rare utility item sets.

David j. haglin et al introduced Minimal Infrequent Itemsets (MINIT) finding method which was the first algorithm produced specially for identifying Minimal Infrequent Item Set (MIIs) [13]. An association surrounded by the number of MIIs and the amount of calculation, identified by the computational time required on the four datasets. The issue related to this method is NP complete.

Pillai, Jyothz et al. [14], proposed HURI which can generate high profitable rare itemsets based on support, utility threshold and user's interest. The future work includes the integration of chronological and fuzzy concept in HURI and using it for finding those rare items, which provide highest profit to a transaction.

AnujaPalhade and RashmiDeshpande in [15], proposed high utility mining approach instead of utilizing the customary methodology focused around frequency. Author proposed a novel skeleton; in meticulous Generation of maximal high Utility Item sets from Data streams (GUIDE), which knowledgeably mine maximal high HUI from vast datasets. The proposed reduced information structure UP-Tree is coordinated for putting away vital data in information streams.

Jyothi Pillai et al. proposed in [16], new advance for mining high utility infrequent itemset using Fuzzy concept, FHURI is a comprehensive version of HURI algorithm. FHURI algorithm has practical meaning to business strategies such as minimizing purchasing expenses of HURI, score suppliers by rating the quality of their supplies and military recognize the the majority successful promotions; recognize profitable itemsets.

3. PROPOSED WORK

In a transactional database two thresholds will be used for extracting HURI. First one is use for discovering HUI and, another for extracting high utility rare itemsets. We propose an analysis on the High Utility Rare itemsets using high utility pattern rare itemset (UPRI) algorithm.[17][18][19]

3.1 Min-utility threshold: In utility mining minimum utility is a user defined value, item sets having value greater than the min –utility threshold consider as a high utility item sets.

3.2 Min-support threshold: A frequent itemset is the itemset containing frequency support higher than a minimum user specified threshold and the rare item Set which has the support lower than the minimum user specified threshold.

The propose method comprises of three steps-

1. Compute transaction utility of each transaction, transaction weighted utility and item's utility.
2. Apply minimum utility threshold on transaction weighted utility, items having lesser value will be discarded.
3. Apply minimum support threshold, items having lesser value will be extracted as rare itemset. So finally high utility rare item set are received from the transaction.

EXAMPLE- Given a table of transaction having some items and their value of occurrence in every transaction.

Table 1. Transaction Database

Transaction	A	B	C	D	E	F	G
T1	1	1	2	1	0	0	1
T2	2	0	5	0	2	1	0
T3	3	1	1	5	1	0	0
T4	0	5	2	2	1	0	0
T5	1	1	2	0	1	1	1

Let us consider a transactional database in Table 1, comprising of seven items and five transaction, each transaction showing the occurrence value of an item in that particular transaction. Each item is linked with a profit value which is given in Table 2.

Table 2. Unit Profit associated with items

Item	A	B	C	D	E	F	G
Profit	5	7	3	1	3	4	2

3.3 Item's utility: An items utility will be calculate by multiplying profit and items value.

U(A)=TI(A)+T2(A)+T3(A)+T5(A)*profit(A)
=1+2+3+1*5=35

T4 having 0 value of A that's why it will not added. Then, U (A) = 35 similarly, we can found all item's utility.

3.4 Transaction's utility:Now we can also calculate transactions utility by multiplying each item with their utility from 1 transaction at one time –

TU(T1)=A(T1)*P(A)+B(T1)*P(B)+C(T1)*P(C)+
D(T1)*P(D)+E(T1)*P(E)+F(T1)*P(F)+G(T1)*P(G)

=(1*5)+(7*1)+(2*3)+(1*1)+(0*3)+(4*0)+(2*1)
=21

TU(T1)=21, TU(T2)=35, TU(T3)=33, TU(T4)=46, TU(T5)=27.

***3.5 Transactions weighted utility*:** Transaction weighted utility of an itemset is the sum of the transaction utilities of all transaction containing that item.

TWU(A)=TU(T1)+TU(T2)+TU(T3)+TU(T5)=116

TWU(B)=127, TWU(C)=162, TWU(D)=100, TWU(E)=141, TWU(F)=62, TWU(G)=48

If minimum utility threshold is 50 than item G will be discarded because it's less than of minimum utility threshold, than minimum support threshold which is 0.7 will be applied on the remaining database. Support will be calculated as- transaction's weighted utility divided by maximum transactions weighted utility.

3.6 Support: A=0.7, B=0.7, C=1, D=0.6, E= 0.8, F= 0.3

Finally, items which are below from the min-support threshold will be considered as high utility rare itemset.

So we have high utility rare items- D, F. and there related itemsets pattern shown in Table-3. Elapsed time is 0.298553 seconds.

Itemset	Support	Profit
D	0.6	100
F	0.3	62
AB	0.6	77
AD	0.4	52
AE	0. 6	93
AF	0.4	60
BD	0.6	98
BE	0.6	104
CD	0.6	98
CF	0.4	60
DE	0.1	79
EF	0.4	60
ABC	0.6	77
ABD	0.4	52
ACD	0.4	58
ACE	0.4	52
ACF	0.6	93
AEF	0.4	60
BCD	0.6	98
BCE	0.6	104
BDE	0.4	79
ABCD	0.4	52
ABCE	0.4	58
BCDE	0.4	60

Table 3. Support and Profit for all itemsets

Now, when we are add a new item H in Table-1 Transaction databasewith its profit value is 2.

Transaction	H
T1	0
T2	2
T3	2
T4	1
T5	3

Table 4. Add new item

We have high utility rare items- **D, F** and **G** and their related high utility rare itemsets. Elapsed time is 0.291898 seconds.

Table5. Add more transactions

T6	2	4	1	2	2	0	0	0
T7	0	0	4	2	1	4	2	0

When we add two more transaction item given in Table – IX, than HURI will be- **A**, **F, G, H**. and their related high utility rare itemsets. Elapsed time is 0.300303 seconds.

We have analyzed:

1. If an item having large number of 0 values in all transactions, it will defiantly present in the high utility infrequent item sets.
2. The utility and the database are updated along with the update of the database.

3 new scenarios are generated for the working of the updated database.

4. It is recommended that this system is also works on frequent changes in the transactional database.

4. HURI USING GENETIC ALGORITHM

Input:

1. A transactional dataset
2. A profit dataset
3. Min_support threshold
4. Min_utility threshold
5. Dynamical addition of transactions

Output: A sanitized database with no sensitive item sets.

Method: In the proposed methodology on the basis of two user specified threshold, we have find out High Utility Rare Itemstes (HURI).

4.1. Min_utility threshold- In utility mining minimum utility is a user defined value, item sets having a value greater than the min_utility threshold consider as a high utility item set.

4.2. Min_support threshold- A frequent itemset is the itemset containing frequency support higher than a minimum user specified threshold and the rare item set which has the support lower than the minimum user specified threshold [20].

The propose method comprises of following steps-

1. First step to calculate transaction utility of each transaction, transaction weighted utility and item's utility of each item in transactions.
2. Apply minimum utility threshold on transaction weighted utility, items having lesser value will be discarded.
3. Apply minimum support threshold, items having lesser value will be extracted as rare itemset. so finally high utility rare item sets received from the transaction.
4. Apply genetic algorithm for hiding high utility rare items.

5. Take maximum length item set from the high utility rare itemsets.
6. The point crossover will be used for flipping the values from the database.
7. Finally, we have sanitized data base.
8. At the end total elapsed time is calculated.

6. RESULT ANALYSIS

The graph is plotted by taking occurrence of items at the x-axis and the number of transactions on the y-axis. From the graph it can easily identify that all the high utility rare itemsets are hidden. So by using the proposed algorithm based on two threshold values, all the HURI will be optimized.

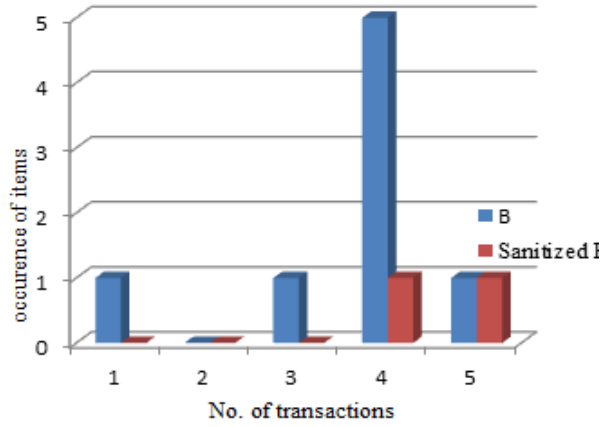

Fig.1.Sanitized database

ACKNOWLEDGMENTS

Due to the broadness of the term, defining privacy is quite challenging. Even in the limited scope of information privacy, several definitions have been presented. In fact, there is always a fair amount of subjectiveness due to individuals' own privacy concept, beliefs and risk assertions. When it comes to making market strategies, how to hide sensitive information from other competitors is important.
According these two utility algorithms, high utility rare patterns can be extracted and privacy preservation of these patterns can be done.
In our analysis we have reached on this decision that genetic approach is more efficient than the previous one on the basis of security.
As a future work more optimization algorithm can be implemented for obtaining better result.

REFERENCES

[1] P. N. Tan, M. Steinbach, V. Kumar. "Introduction to Data mining". 2009.
[2] A. Raorane and R.V. Kulkarni. "Data Mining Techniques:A Source For Consumer Behavior Analysis". International Journal of Database Management Systems, Vol.3,No.3,pp: 45-56, Aug. 2011.
[3] R. Agrawal and R. Srikant. "Fast Algorithms for Mining Association Rules". InProceedings of the 20th International Conference Very Large Databases, pp. 487-499,1994.
[4] V. Jaideep, and C. Clifton. "Privacy preserving association rule mining in vertically partitioned data". ACM, pp: 639-644, 2002.
[5] H. Jiawei, J. Pei, and Y. Yin. "Mining frequent patterns without candidate generation". In ACM Sigmoid Record, Vol. 29, No. 2, pp. 1-12, 2000.
[6] R. Agrawal, T. Imielinski and A. Swami. "Mining association rules between sets of itemsin large databases". ACM SIGMOD International Conference on Management of data, pp: 207-216, 1993.
[7] H. Yao and H. J. Hamilton, "Mining itemset utilities from transaction databases". Data and Knowledge Engineering, vol. 59, pp. 603-626 2006.
[8] J. Hu. And A. Mojsilovic, "High-utility pattern mining: A method for discovery of high-utility item sets", Pattern Recognition 40, 2007.
[9] Liu, Y., Liao, W., and A. Choudhary. "A Fast High Utility Itemsets Mining Algorithm".InProceedings of the Utility- Based Data MiningWorkshop, August 2005.
[10] H.F. Li, H.Y. Huang, Y.Cheng Chen, and Y. Liu and S. Lee. "Fast and Memory EfficientMining of High Utility Itemsets in DataStreams". Eight IEEE International Conference on Data Mining, 2007.
[11] V. S. Tseng, C.J. Chu, T. Liang. "Efficient Mining of Temporal High Utility Itemsets from Data streams". Proceedings of SecondInternational Workshop on Utility-Based Data Mining, August 20, 2006.
[12] G.C.Lan, T.P. Hong and V.S. Tseng. "A NovelAlgorithm for Mining Rare-Utility Itemsets ina Multi-database environment"
[13] D.Haglin and A.Manning. "On Minimal Infrequent Itemset Mining". In 2007 International Conference on Data Mining, pp: 141- 147, 2007.
[14] P. Jyothi, and O. P. Vyas. "High Utility Rare Item Set Mining (HURI): An Approach for Extracting High Utility Rare Item Sets". I-Manager's Journal on Future Engineering and Technology 7, No. 1, 2011.
[15] P.Anuja, and R. Deshpande. "An Effective Up-Growth Algorithm for Discovering High Utility Itemset Mining".International Journal of Science and Research (IJSR), 2013.
[16] P. Jyothi, O. P. Vyas and M. K. Muyeba. "A Fuzzy Algorithm for Mining High Utility Rare Itemsets-FHURI".International Journal on Recent Trends in Engineering & Technology, 2014.
[17] W. Dong, H. Jiang, L. Chen and G. Liu. "Incremental updating algorithm for infrequent itemsets on weighted condition".International Conference on Computer Design And Applications, 2010.

[18] A. Gupta, A. Mittal, and A. Bhattacharya. "Minimally Infrequent Itemset Mining Using Pattern-Growth Paradigm and Residual Trees" .Conf. Management of Data (COMAD), pp: 57-68, 2011.

[19] S. Shrivastava ,P.K. Johari. "Analysis on High Utility Infrequent ItemSets Mining Over Transactional Database". IEEE International Conference On Recent Trends In Electronics Information Communication Technology, 2016.

[20] S. Shrivastava ,P.K. Johari."Privacy Preservation of Infrequent Itemsets Mining Using GA Approach". Recent Developments in Intelligent Computing, Communication and Devices, pp 97-104, 2017.