An Algorithm for Privacy Preservation of High Utility Rare Itemsets

Kanika Middha[1], Prof. Jeetesh Kumar Jain[2]
[1, 2] Department of Computer Science and Engineering
[1, 2] SRCEM, Banmore, Gwalior, M.P., India

Abstract- *Data mining is the field of extraction or mining of important information. The mining of information is somehow leading to the loss of information also due to semi-honest adversaries and fraudulent parties. Thus, to preserve the data, concept of privacy preservation is included in data mining. The frequent itemsets i.e. the itemsets having much use have been preserved till date but, the rare itemsets also have their importacnce and thus, need to be saved. This, paper proposes a new method of finding rare itemsets and their preservation in the best possible way. Thus, contributing to the privacy preservation field. The results show that all the rare itemsets are preserved well without the loss of data.*

Keywords- Data Mining, Association Rule Mining, Utility Mining, Rare Itemsets, privacy preservation.

I. INTRODUCTION

Amid the late ten years, Data mining, additionally named as learning disclosure in databases has set up its position as an unmistakable and key investigation field. The target of data mining is to concentrate concealed data from plenitude of crude data[3]. Data mining has been used as a piece of various data areas. Data mining can be saw as an algorithmic method that takes data as info and yields designs, for instance, request rules, itemsets, association guidelines, or synopses, as yield [11].

Data Mining assignments are comprehensively portrayed into two classes, Descriptive Mining and Predictive Mining. The Descriptive Mining techniques, for instance, Clustering, Discovery of Association Rule, Discovery of Sequential Pattern, is used to find human-interpretable examples that delineate the data. The Predictive Mining strategies like Classification, Regression, Deviation Detection, use a couple of variables to foresee dark or future estimations of different variables.

Mining Association ruels is one of the investigation issues in data mining [1]. Given an arrangement of exchange where each exchange is a plan of items, an association guideline is a structure's affirmation XY, where X and Y are sets of items.

The issue of mining association principles was at first exhibited in [1] and later extended in [2], for the case of databases involving straight out resources alone.

Utility mining model was proposed in [19] to describe the utility of itemset. The utility is a measure of how supportive or gainful an itemset X is. The utility of an itemset X, i.e., u(X), is the utilities' sum of itemset X in each one of the exchange containing X. An itemset X is known as high utility itemset if and just ifu(X) >= min_utility, where min_utility is a client characterized slightest utility edge [11].The basic focus of high-utility itemset mining is the find each one of those itemsets having utility more noticeable or comparable to client characterized minimum utility edge.

1.1 Privacy Preservation

Security safeguarding [23] has started as an imperative worry with reference to the data's accomplishment mining. Security saving data mining (PPDM) manages ensuring the protection of individual data or delicate learning without giving up the data's utility. Individuals have turned out to be very much aware of the protection interruptions on their own data and are extremely hesitant to share their touchy data. This may prompt the incidental consequences of the data mining. Inside of the requirements of protection, a few routines have been proposed yet this branch of exploration is in its earliest stages.

Security Preserving Data Mining (PPDM) [24] turns into a well known exploration zone in data mining in the previous couple of years. In 1996, Cliftonetal. [25]analyzed that data mining can realize danger against databases and tended to conceivable answers for accomplish security assurance of data mining. In 2002, Rizvietal. examined the security protecting mining of association rules[26,27].

The cleaning algorithms for the security safeguarding mining on association guidelines can be separated into two classes :
(1)Data-Sharing method and
(2) Pattern-Sharing method

(1) Data-Sharing method: Sanitization procedure performs on information to the eliminate or hide the restrictive association rules group that include sensitive information.

(2) Pattern-Sharing method: Sanitizing algorithm performs on rules mined from a database instead of the data itself. Regarding pattern-sharing methods, the only known method that falls into this category was introduced in [28].

In protection safeguarding data mining, utility mining assumes an imperative part. In protection safeguarding utility mining, some delicate itemsets are covered from the database as indicated by certain security strategies. Concealing touchy itemsets from the foes is turning into a essential issue these days. Additionally, just not very many techniques are accessible in the writing to conceal the touchy itemsets in the database. One of the current security protecting utility mining routines uses two algorithms considering the utility variables as amounts and benefits in genuine applications by Yeh and Hsu. In the first place, they proposed the assurance algorithm for data sterilization to abstain from uncovering the delicate data [29]. In HHUIF algorithm, the most compelling item in the exchange which contains the touchy high utility itemsets is picked. The picked's amount item is in this manner diminished for concealing the touchy high utility itemsets.

In PPUM, the data disinfection for concealing the touchy data could be partitioned into two sorts, which are the item purification and the exchange cleansing. The exchange sterilization system is embraced in this paper to secure the touchy data. In protection safeguarding utility mining (PPUM) [29], the reason for existing is to shroud the touchy high utility itemsets with the insignificant reactions.

II. LITERATURE SURVEY

In the past area we have exhibited the essential thought of Association Rule mining, Data Minin, Rare Itemset Mining and Utility Mining. A brief survey of distinctive algorithms, thoughts and strategies portrayed in assorted examination papers have been given in this area.

The digging of association guidelines for finding the relationship between data itemsets in boundless databases is an overall thought about framework in data mining field with specialists methods like Apriori [1], [2]. ARM method can be broke down into two stages. The important step incorporates finding all standard itemsets in databases. The second step incorporates making association rules from successive itemsets.

In [6], Yao et al decided the issue of utility mining,a theoretical model called MEU, which finds all itemsets in an exchange database with utility qualities higher than the minimum utility breaking point. The numerical model of utility mining was characterized in perspective of utility bound property and the bolster bound property. This built up the system for future utility mining algorithms.

Liu et al proposed algorithm Two-Phase [8] for discovery high utility itemsets. In the first stage, a model that applies the "exchange weighted descending conclusion property" on the hunt space to help the acknowledgment confirmation of hopefuls. In the second stage, one extra database range is performed to perceive the high utility itemsets.

In paper [14], L. Szathmary et al formulated an exceptional procedure for part in order to processing all uncommon itemsets there are itemset mining errand into two stages. The important step is the acknowledgment of the negligible uncommon itemsets. In the second step, the negligible uncommon itemsets are taken care of remembering the complete objective to restore all uncommon itemsets. To recoup all uncommon itemsets from minimaal uncommon itemset(mRIs), a model algorithm called "A Rare Itemset Miner Algorithm(Arima)" was proposed. Arima makes the arrangement of all uncommon itemsets, parts into two sets: the arrangement of uncommon itemsets having a zero backing and the arrangement of uncommon itemsets with non-zero backing. In case an itemset is uncommon then any increase of that itemset will come to fruition an uncommon itemset.

R. Agrawal et al in [10] proposed Apriori algorithm, which is utilized to get incessant itemsets from the database.The itemsets which show up every now and again in the exchanges are called regular itemsets. MINIT (MINimal Infrequent Itemsets), which is the first algorithm outlined particularly to mine insignificant occasional itemsets (MIIs)[11]. A negligible occasional item set is a rare item set that don't have a subset of items which frames a rare item set. MINIT is both insignificant and non-negligible (unweighted) rare itemset mining from unweighteddata.It depends on SUDA2 algorithm. Additionally demonstrated that the negligible occasional itemset issue is NP-complete issue.

Apriori-uncommon is an adaption of the Apriori algorithm used to mine incessant itemsets.To recoup all uncommon itemsets inside insignificant uncommon itemset (mRIs), a model algorithm named ―A Rare Itemset Miner Algorithm (Arima) was contrived in [17]. Arima produces the arrangement of all uncommon itemsets, isolates into two sets: the arrangement of uncommon itemsets having a zero backing

and the arrangement of uncommon itemsets with non-zero backing. In the event that an itemset is uncommon then any expansion of that itemset will results an uncommon itemset [17]

Problem Statement

Provide transactional database and user-specified minimum utility threshold and also minimum support threshold, the problem of mining high utility rare itemsets is to find the complete set of the itemsets whose utilities are higher than or equal to minimum utility threshold and whose support count is lesser than minimum support threshold and then to preserve them by bringing the utility of the itemsets below the threshold. Thus, applying the privacy preservation of the rare itemsets.

DEFINITIONS:

Table 1. An example database

TID	Transaction	TU
T1	(A,1) (C,1) (D,1)	8
T2	(A,2) (C,6) (E,2) (G,5)	27
T3	(A,1) (B,2) (C,1) (D,6) (E,1) (F,5)	30
T4	(B,4) (C,3) (D,3) (E,1)	20
T5	(B,2) (C,2) (E,1) (G,2)	11

Table 2. Profit table

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

Definition 1.**(Utility Mining)** Utility Mining finds all the itemsets in transaction database with utility values higher than the user defined minimum utility threshold.

Given a finite items set I={i1,i2,.....im}, all item has a unit profit pr(ip). A transaction database D ={T1, T2...,Tn}contains a transactions set, and each transaction has a unique identifier d, known TID. all item ip in transaction Td is associated with a quantity q(ip,Td), which the purchased quantity of ip in the Td.

Definition 2.**(Item Utility)** Utility of an item in a transaction database is the product of profit and quantity.$u(ip, Td)$ and dscribed as $p(ip) \times q(ip, Td)$. For example, in Table 1, $u(\{A\}, T1) = 5 \times 1 = 5$.

Definition 3.**(High utility itemset)** An itemset is known as a high utility itemset if its utility is no less than a user specified minimum utility threshold which is signified as min_util. Otherwise, it is known as a low-utility itemset.

Definition 4.**(Transaction Utility)** The transaction utilityvalue of a transaction is the sum of utility values of each itemsin a transaction. Transaction utility reflects utility in atransaction database It is denoted as TU(Td)and defined as u(Td , Td). For example, $TU(T1) = u(\{ACD\}, T_1) = 8$.

Definition 5.**(Transaction Weighted Utility)** Transaction weighted utility of an itemset X is the transaction sum utilities of each transactions including X.

Definition 6.**(High transaction weighted utility itemset)** An itemset X is known as a HTWUI if TWU(X) is no less than minimum utility threshold. Transaction-weighted utilization of an itemset X is the sum of the transaction utilities of each transactions including X, which is denoted as $TWU(X)$ and defined as

$$\sum_{X \subseteq T \wedge T \in D} TU(T_d)$$

For example, $TWU(\{AD\}) = TU(T1) + TU(T3) = 8 + 30 = 38$. If $TWU(X)$ is no less than minimum utility threshold, X is known as a *high transaction-weighted utilization itemset*.

Definition 7.**(Rare Itemset Mining)** Rare itemsets are the itemsets that occur infrequently in the transactional dataset. An itemset X is called as rare itemset if it is below the minimum specified threshold.

Definition 8.**(High Utility Rare Itemset)** A high itemset which is a high utility itemset, but occurs infrequently in the dataset.

III. PROPOSED WORK

A. Proposed Method

The framework of the proposed methods consists of two phases:

- **PHASE -1**

In the first level high utility itemsets are created having utility value greater than the transaction utility threshold.

- **PHASE -2**

In the second phase, rare itemsets are generated from high utility itemsets known as the high utility rare itemsets. Rare itemsets are the itemsets appearing below the threshold support value.

- **PHASE -3**

The rare itemsets that have been generated as a result of phase-2 are preserved in this phase thus, bringing them below the threshold.

PHASE -1:

Phase1 of the proposed method comprises of three steps to generate high utility itemsets.

1. Compute Transaction utility of all transaction, transaction weighted utility and support of all item
2. Construct a Growth Tree.
3. Generate high utility itemsets.

The Proposed Data Structure: Growth-Tree

To minimize scanning database repeatedly and to enhance the mining performance, we use a tree data structure, namely Growth Tree. The Growth-Tree is used to represent and maintain the itemsets knowledge and their utilities in transactions.

Removing local unpromising items

Suppose if there is a path in an item's conditional pattern base which contains a set of unpromising items. The set of unpromising items, do not favour the high utility itemsets generation .Thus, the unpromising items and their utilities can be eliminated from the path. Then the path is rearranged in fixed order.

PHASE-2

Mining High Utility Rare itemsets :

Minimum support threshold is specified by the user according to users' preference .After the completion ofphase1,high utility itemsets are generated, These items are mined to get high utility rare itemsets. The items which are below the minimum support threshold are known as rare itemsets. Support of an item in the transaction database is the proportion of occurrence of the item.

PHASE-3

Preserving High Utility Rare Itemsets

The high utility rare itemsets are now preserved in this phase by bringing the utility of the itemset below threshold. The user defined threshold is chosen. The values are brought below this threshold value. Whole phases are explained with a pseudo-code that explains the whole proposed algorithm in an easier way. The pseudo-code is as follows:-

Description: Privacy Preservation of High Utility Rare Itemsets of user's interest.

Input:-Db: original database
C : Candidate itemset of size k
L : Rare itemset of size k
$\{S_1, S_2, \ldots S_i\}$: Sensitive itemsets

Output:- the sanitized database Db'.

Algorithm 1:

1. Reading the data from the database Db.
2. Set the min_utility and min_support.
3. Finding the transaction weighted utility of the itemsets based on quantity * profit.
4. Calculating the transaction utility based on the transaction weighted utility.
5. Arrange the itemsets in decreasing order and save as plist.
6. Removing the elements which are below the mininmum support.
7. Generating all the frequent patterns.
8. Constructing Growth_Tree();
9. High utility items are derived from the tree.
10. If(item supp<min supp)
    Choose items and set them as high utility rare itemsets.
11. Now applying privacy preservation on the high utility rare itemsets by calling Preservation();
12. End

Algorithm 2: Growth_tree()
For trans 1 :n
For each item I in trans do

1. Arrange elements in trans in increasing order.
2. Create node based on the order.
   node$_i$_count = count(curr_node)
   node$_i$_nu = I
   if(child_exist == true)
   if(child_previously_exists == true)
   node_curr ++
   curr_utility = curr_utility + prev_utility;

End if

```
Algorithm 3: Preservation()
While(rare_itemset is not empty)
        Diff = utility – threshold
                If  item< diff
                        Item_value ==0
                        Diff = diff – utility.
                        Update database.
                Else
                        Item_value = item_value +
                ceil(diff/ item_value)
                        If item_value< 0
                                Item_value ==0
                                Update database.
                        End if.
                End if.
End while.
```

IV. RESULT ANALYSIS

Step 1 : The user is asked to input minimum utility and the minimum support.

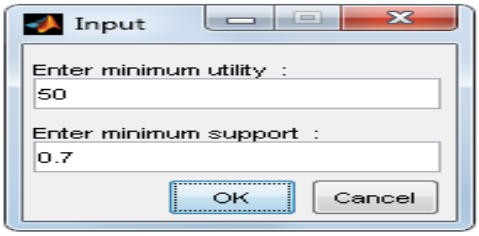

Fig. 1 : User input

Step 2 : Based on the proposed algorithm, the growth tree is constructed for the example taken. The Table 1.1 and Table1.2 are considered as an example for the growth tree construction. The tree generated is shown like this:-

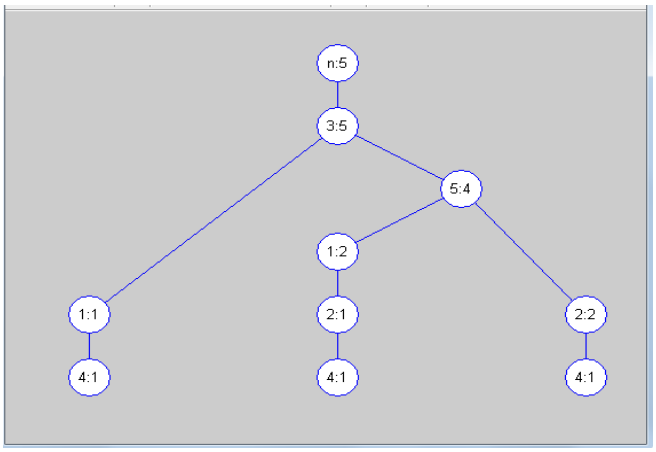

Fig 2: Growth Tree for the Example taken

Step 3: After Phase 1 & phase 2, the high utility rare itemsets generated are shown on the display screen. These are:-

Rate Items :

1
2
4

Step 4: the final display of the results is-

1. It displays the initial rare itemsets with utility value more than minimum threshold.
2. The number of rounds in which all the itemsets get preserved.
3. Total time taken for the whole process to take place.

HURI pairsHURI		Utility
1-Itemset	{1}	94
1-Itemset	{2}	90
1-Itemset	{3}	137
1-Itemset	{4}	87
1-Itemset	{5}	126
2-Itemset	{1,3}	72
2-Itemset	{1,5}	61
2-Itemset	{2,3}	74
2-Itemset	{2,4}	64
2-Itemset	{2,5}	74
2-Itemset	{3,4}	75
2-Itemset	{3,5}	100
2-Itemset	{4,5}	64
3-Itemset	{1,3,4}	72
3-Itemset	{1,3,5}	72
3-Itemset	{2,3,4}	74
3-Itemset	{2,3,5}	74
3-Itemset	{2,4,5}	64
3-Itemset	{3,4,5}	75
4-Itemset	{1,3,4,5}	72
4-Itemset	{2,3,4,5}	74

no_of_times =
8

Rate Items : No Rare items left
Elapsed time is 1.644965 seconds.

Effectiveness measurements:-

The effectiveness measurements can be shown with the help of the table below. The table have fields that effect the no. of itemsets and their preservation.

Table 3: Effectiveness measurements results

Threshold	Rare itemsets generated	Itemsets preserved	No. of times	Time taken
50	3	3	8	2.666458
55	3	3	4	1.259982
60	3	3	4	1.258999
65	3	3	4	1.214257
70	3	3	3	1.188320
75	3	3	1	0.989486
80	3	3	1	0.982073
85	3	3	1	0.969127

1. Comparison graph between the threshold to preserve rare itemsets and time taken is plotted. As the threshold increases, the time taken to preserve the itemsets decreases. This can be easily seen in the fig 3.

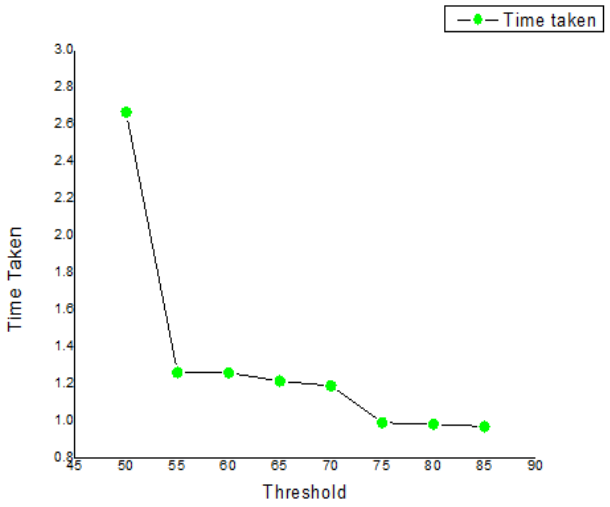

Fig 3: Comparison between threshold and time taken to preserve the rare itemsets generated.

2. Comparison to show the no. of rounds or loops that have been required to preserve the rare itemsets along with the no. of itemsets generated on various threhsolds. A 3-D graph is plotted to show the results.

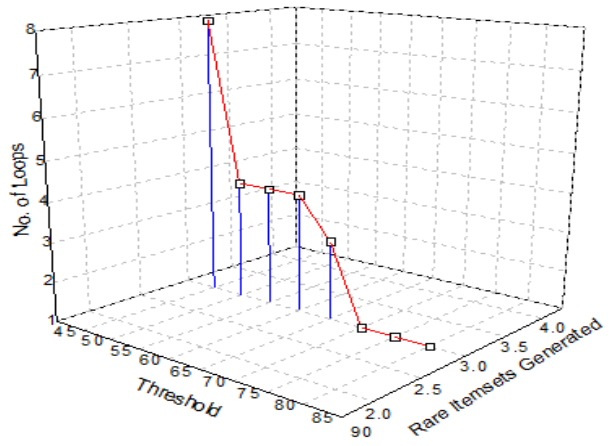

Fig. 4: Comparison to show the no. of loops and the threshold for various rare itemsets generated.

V. CONCLUSION

In modern day business, business strategies are made based on the user requirements and their patterns towards the data available. To introduce the right strategy, a new growth-tree based algorithm has been proposed to mine the high utility rare itemsets from the databases.

These are those itemsets that occur infrequently in the patterns but are important for various decision making issues. After applying the algorithm, the next focus is to work on the privacy preservation of the HURI. These itemsets are then preserved using a new proposed algorithm that preserves the rare itemsets. Thus, any third-party is not able to find them and this leads to the preservation od data in the best way. The algorithm can be easily used in the real time applications like hospitality, banking, supermarts, IT firms, etc.

REFERENCES

[1] R. Agrawal, T. Imielinski and A. Swami, 1993,"Mining association rules between sets of itemsin large databases", in Proceedings of the ACMSIGMOD International Conference on Management of data, pp 207-216.

[2] R. Agrawal and R. Srikant, 1994, "FastAlgorithms for Mining Association Rules", inProceedings of the 20th International ConferenceVery Large Databases, pp. 487-499.

[3] Attila Gyenesei, "Mining Weighted AssociationRules for Fuzzy Quantitative Items", Lecture notes in Computer Science, Springer, Vol.1910/2000, pages 187-219, TUCS TechnicalReport No.346, ISBN 952-12-659-4,ISSN 1239-1891, May 2000.

[4] R. Chan, Q. Yang, Y. D. Shen, "Mining Highutility Itemsets", In Proc. of the 3rd IEEE Intel.Conf. on Data Mining (ICDM), 2003.

[5] H. Yun, D. Ha, B. Hwang, and K. Ryu. "Mining association rules on significant rare data usingrelative support". Journal of Systems andSoftware, 67(3):181–191, 2003.

[6] H.Yao, H. J. Hamilton, and C. J. Butz, "A Foundational pproach to Mining Itemset Utilities from Databases", Proceedings of the Third SIAM International Conference on Data Mining, Orlando, Florida, pp. 482-486, 2004.

[7] G. Weiss. "Mining with rarity: a unifying framework", SIGKDD Explor. Newsl., 6(1):7–19, 2004.

[8] Liu, Y., Liao, W., and A. Choudhary, A., "A Fast High Utility Itemsets Mining Algorithm", In Proceedings of the Utility- Based Data Mining Workshop, August 2005.

[9] Lu, S., Hu, H. and Li, F. 2005. "Mining weighted association rules. Intelligent Data Analysis",5(3):211–225.

[10] V. S. Tseng, C.J. Chu, T. Liang, "Efficient Mining of Temporal High Utility Itemsets from Data streams", Proceedings of Second International Workshop on Utility-Based Data Mining, August 20, 2006.

[11] H. Yao, H. Hamilton and L. Geng, "A Unified Framework for Utilty-Based Measures for Mining Itemsets", In Proc. of the ACM Intel.Conf. on Utility-Based Data Mining Workshop(UBDM), pp. 28-37, 2006.

[12] A. Erwin, R.P.Gopalan and N. R. Achuthan, 2007,"A Bottom-up Projection based Algorithm formining high utility itemsets", in Proceedings of2nd Workshop on integrating AI and Data Mining(AIDM 2007)", Australia, Conferences in Research and Practice in Data Technolofy(CRPIT),Vol. 84.

[13] J. Hu, A. Mojsilovic, "High-utility pattern mining: A method for discovery of high-utility item sets", Pattern Recognition 40 (2007) 3317 –3324.

[14] L. Szathmary, A. Napoli, P. Valtchev, "Towards Rare Itemset Mining" Proceedings of the 19thIEEE International Conference on Tools with Artificial Intelligence, 2007, Volume 1, Pages: 305-312, ISBN ~ ISSN:1082-3409 , 0-7695-3015-X

[15] Kriegel, H-P et al. 2007. "Future Trends in Data Mining, Data Mining and Knowledge Discovery", 15:87–97.

[16] M. Adda, L. Wu, Y. Feng, "Rare ItemsetMining", Sixth International conference onMachine Learning and Applications, 2007, pp 73-80.

[17] H.F. Li, H.Y. Huang, Y.Cheng Chen, and Y. Liuand S. Lee, "Fast and Memory Efficient Mining of High Utility Itemsets in Data Streams", 2008 Eighth IEEE International Conference on Data Mining.

[18] M. Sulaiman Khan, M. Muyeba, Frans Coenen,2008. "Fuzzy Weighted Association Rule Mining with Weighted Support and Confidence Framework", to appear in ALSIP (PAKDD),pp.52-64.

[19] S. Shankar, T.P.Purusothoman, S.Jayanthi and N.Babu, "A Fast Algorithm for Mining High Utility Itemsets", Proceedings of IEEE International Advance Computing Conference(IACC 2009), Patiala, India, pages : 1459 - 1464

[20] Hu, J., Mojsilovic, A. "High-utility Pattern Mining: A Method for Discovery of High utility Item" Sets, Pattern Recognition, Vol. 40,3317-3324.

[21] G.C.Lan, T.P.Hong and V.S. Tseng, "A NovelAlgorithm for Mining Rare-Utility Itemsets ina Multi-Database Environment".

[22] J. Pillai, O.P. Vyas, S. SoniM. Muyeba "AConceptual Approach to Temporal Weighted Itemset Utility Mining", 2010 International Journal of Computer Applications (0975 - 8887) Volume 1 – No. 28.

[23] M. B. Malik, M. A. Ghazi and R. Ali, "Privacy Preserving Data Mining Techniques: Current Scenario and Future Prospects", in proceedings of Third International Conference on Computer and Communication Technology, IEEE 2012.

[24] C. Clifton and D. Marks, "Security and privacy implications of data mining," in Proceedings of the 1996 ACMSIGMOD Workshop on Data Mining and Knowledge Discovery,pp.15-19,1996.

[25] Y. Saygin,V. S. Verykios, and C.Clifton, "Using unknowns to prevent discovery of association rules," SIGMOD Record,vol.30,no.4,pp.45-54,2001.

[26] V.Verykios,E. Bertino, I. G. Fovino, L. P. Provenza,Y.Saygin, and Y. Theodoridis, "State-of- the-art in privacy preserving data mining", SIGMOD Record,vol.33,no.1,pp.50-57,2004.

[27] V. Verykios, A. Elmagarmid, E. Bertino, Y. Saygin, and E. Dasseni, "Association rules hiding," IEEE Transactions on Knowledge and Data Engineering,vol.16,no.4,pp.434-447,2004.

[28] H.Yao, H.J.Hamilton, and C.J.Butz, "A foundational approach to mining itemset utilities from Databases," in Proceedings of the 4th SIAM International Conference on DataMining,pp.482-486,2004.

[29] J. S. Yeh and P. C. Hsu, "HHUIF and MSICF: novel algorithms for privacy preserving utility mining,"Expert Systems with Applications, vol. 37, no. 7, pp. 4779–4786, 2010.