ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 3, Issue 1, July 2013

Hiding Sensitive Association Rules by Distorting RHS Items

Parvin Shirrouhi, Mohammad Naderi Dehkordi, Faramarz Safi
Computer Engineering Department, Islamic Azad University, Najafabad Branch

Abstract ***—Association rules is a data mining technique which extracts useful patterns in the form of laws. One of the major problems in applying this technique on a Dataset is the disclosure of sensitive information which would endanger their security and confidentiality. Privacy-preserving data mining is to preserve the privacy of the personal data identified by the data mining techniques. Hiding association rules is one of the methods in privacy-preserving. In this article, one hiding association rules algorithm has been discussed. In the proposed algorithm for hiding sensitive rules, Data Distortion Techniques— based on reduction of confidence rules— has been used. Reduction of confidence in sensitive rules, through the reduction of right side Items Set support, and working on a series of transactions, which totally support the sensitive rules, by choosing a transaction with the least amount of items. Our proposed algorithm with two reference algorithms on both compacted and non-compacted Dataset has been implemented, we observed that the execution time of the proposed algorithms are compared with both reference on both Dataset has been considerably reduced. Also, the number of lost rules, the non-compacted Dataset, the proposed algorithm is more efficient than the two reference algorithms. The effectiveness of the proposed algorithm has been analyzed through the implementation and comparing of the obtained results with the reference algorithms. The results indicate that the proposed algorithm is effective.***

Index Terms**—Data Mining, Hiding Association Rules, Privacy Preserving.**

I. INTRODUCTION

Data mining aims to find useful patterns from large amounts of data in the transactional Dataset [15]. Association rules mining is a data mining technique with which we extract useful patterns in the form of laws from the transactional Dataset [2]. If, in the opinion of the Dataset's owner, the extracted association rules are important and sensitive they will be termed as Sensitive Association Rules (RH), otherwise they will be called Non-Sensitive Association Rules (~RH). Therefore after the data-mining process on the primary Dataset, the extracted rules are divided into two categories: rules ~ RH and RH. With the disclosure of this RH information, their security and privacy will be jeopardized. For example, in the medical Datasets, sharing information about the patients is useful, but at the same time it is necessary to maintain their "identity". Here the Privacy-Preserving of the people must be protected. Thus, using data mining technology on Datasets of companies, organizations, etc., has brought about concerns about data security against unauthorized access, and therefore strengthened our resolve to reach an important goal, which is, Dataset security and privacy-preserving. Privacy-preserving is a term associated with the extraction of this information, so that we are able to hide some important information that we do not wish to disclose. Therefore the concept of data mining Privacy-preserving [1, 3] is the process of preserving private information of data mining algorithms. The purpose of data mining Privacy-preserving is to develop algorithms to transform the original data, so that private information and knowledge will not be discovered after the process of data mining techniques. Hiding association rules[2] is one of the methods in privacy-preserving. The hiding association rule algorithms intend to prevent the extraction of sensitive associations and laws after data mining process, by making minimal changes to the original Dataset. In this research data distortion techniques [16] have been used for Hiding Sensitive Association Rules, based upon reduction of the confidence level in sensitive law. Reduction of confidence in sensitive rules, through the reduction of right side Items Set support, and working on a series of transactions, which totally support the sensitive rules, by choosing a transaction with the least amount of items. In this algorithm for hiding any sensitive rules, the amount of transactions on which the changing operation is performed, is calculated and the same amount of delete operation is performed on the desired Items. After changing the transactions, the confidence level of Sensitive Rule is calculated, and we find that the confidence level of Sensitive Rule has been reduced and the Sensitive Rule is hiding. Compared to the reference algorithms, execution time for this algorithm is reduced considerably. Rest of the article is organized in seven sections. Section 2 presents the formal definition of the issue of hiding association rules. Section 3 presents a review of related performed work. Section 4 explains the Privacy-Preserving of Association Rules, symbols, and definitions and expressions of Sensitive Rule. Section 5deals with the proposed algorithm, symbols and explain the various stages. Section 6 compares and evaluates the proposed algorithm with algorithmsb.1 and ADSRRC. And section 7 presents future works, discussions and conclusion.

II. FORMAL DEFINITION OF THE ISSUE OF HIDING ASSOCIATION RULES

Issue input: Original Dataset D, which we wish to publish. Sensitive rules RH and Minimum Support Threshold (MST) and Minimum Confidence Threshold (MCT) have been specified.

A set of sensitive association rules RH ⊂ R that should be hidden. Issue output: Modified Dataset DM should be made in such a manner that the set of sensitive association rules RH cannot be extracted from the DM, the set of sensitive association rules RH ⊂ R should be extractable as much as possible, efforts should be made to have no Lost Rules, and Ghost Rules should not be produced. The process of Hiding

Association Rules is also known as Sanitization. Fig.1 shows the process of Sanitization of Association Rules.

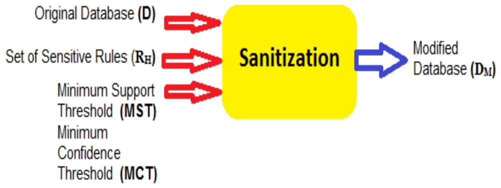
Original Database (D)
Set of Sensitive Rules (RH)
Minimum Support Threshold (MST)
Minimum Confidence Threshold (MCT)
Sanitization
Modified Database (DM)

Fig 1. Process of Sanitization of Association Rules

III. RELATED WORK

Many methods are proposed for Privacy-Preserving of Dataset knowledge which has been divided into the following classes: Heuristic Approaches, Border-base Approaches, and Exact Approaches. Many of the new methods belong to the Heuristic Approach. This approach is in itself divided into two groups, Data Modification and Data Reconstruction. In Data Modification group, to remove Sensitive Association Rules, the original Dataset is directly altered. In this method the two following technique are used: Data distortion and data blocking techniques. Data Distortion technique tries to hide Association Rules by increasing or decreasing the amount of support or increasing or decreasing the amount of Confidence. For increasing or decreasing the amount of support or increasing or decreasing the amount of Confidence in the selected transactions, if there are no items present we create them and if there are items present, then we have to eliminate them. In methods based on Distortion, by not revealing any useful information, good security can be achieved. Atallah et al [4, 7] for the first time provided an algorithm for hiding association rule by reducing the amount of support. Reduction of the amount of support in the set of Items is done using a data base Lattice graph. Dasseni et al [5] aimed at reducing the effect so non_sensitive Items. Decreasing the amount of Confidence with increasing the amount of supporting Antecedent (LHS) rules, through transactions which Partially Support adhere by the rules until the time Confidence reaches below its minimum Confidence and hides the rules. Olivera&Zaiane [6] for the first time proposed the method of Multiple Rule Hiding. To hide there is a need to scan the DataSet twice, regardless of the number of Sensitive Items. The first scan is to make index files to speed up the process of finding sensitive transactions and allow for efficient retrieval of data. The second scan is to apply algorithm Dataset selectively. Verykios et al [8] aimed at hiding an Item that has the maximum support among short length transactions. Hiding set of items in a round robin method. Items are eliminated through the round robin method until the amount of support reaches below the MCT. Pontitakis[9]Priority-based Distortion Algorithm (PDA): for a sensitive rule, the presence of sensitive Items in the desired transactions will eliminate them, and will decrease the amount of Confidence in a Sensitive Rule. Weight-based Sorting Distortion Algorithm (WSDA), focuses on the optimization of the Hiding process, and tries to minimize the side effects and complications. The process of elimination begins by prioritizing the transactions based on their weight. In this research we have tried to Hiding Sensitive Association Rules by using Distorting techniques, so that if there is a data mining process on the Dataset, the Sensitive Association Rules will not be discovered and thus their security and privacy are not compromised. Dehkordi et al [10] proposed a novel method for privacy preserving association rule mining based on genetic algorithms. It also makes sure that no normal rules are falsely hidden (lost rules) and no extra fake rules (ghost rules) are mistakenly mined after the rule hiding process using genetic algorithm. The algorithm sanitizes both rule and item set with minimal side effects by introducing new hiding strategies.

IV. PRIVACY PRESERVING OF ASSOCIATION RULES

A. *Symbol of foundation and definition*

Association Rules is one of the important techniques of Data Mining, whereby we extract the useful patterns in the form of rules from the transaction Dataset. Each Association Rule is in the form of $A \Rightarrow$, is pre-eminent and is called the left hand side of the rules, and Is the outcome and is called the right hand side of the rules. If the set of Items present in the Dataset is ={ , ,…, } and is a Dataset of transactions, in which each transaction T is a set of Items in which ⊆ . , Set of items in , if ⊆ , therefore $Support\,(A) = \frac{|}{|}$ in which | is the number of transaction which includes , and || is the total number of transactions in the Dataset. is a set of Items called Frequent, if Support( ) is greater than the M. To find frequent set of Items we can use popular algorithms such as Apriori Algorithm, Eclat Algorithm, etc. From now on, in this article S will stand for $Suppo$ and Co will stand for $Confiden$. Each $A \Rightarrow$ Association rule will be extracted as follows: ⊂ , ⊂ and $A \cap$ = Support Rule: Represents the frequency of Rule in the Dataset. $A \Rightarrow$ Support Rule is equal to $Support\,(A, B) = \frac{|A \cup}{|D}$, and Confidence Rule: Represents the amount of Rule power in the Dataset. $A \Rightarrow$ Confidence Rule is equal to $Conf\,(A \Rightarrow B) = \frac{Sup(A \Rightarrow}{Sup(A}$ defined as Minimum Support Threshold $M$and Minimum Confidence Threshold M. The prerequisite for extraction of a Rule is firstly Sup( )> M and secondly Conf( )> M. If the arrangement is Sup( )< M and or Conf( )< M Rule extraction will not take place.

B. *Terms of sensitive rules*

For a DataSet , a user specifies the minimum support threshold and the minimum confidence threshold and then extracts Association rules from the Dataset by using algorithms such as Apriori Algorithm. We call this set of Association rules set of . The user chooses a Isubset, which is a Sensitive Rule, and then tries to hide it under

⊆ Hiding sensitive rules means protection of Privacy-Preserving of the data and security of the Dataset. In this article we cannot formally define sensitivity. But we consider sensitivity in association with reducing the amount of support in a rule , or reducing the amount of confidence in a rule to below the level of minimum support or minimum confidence. A changed Dataset , after the process of Hiding, does not contain sets which are Sensitive Rules.

V. PROPOSED ALGORITHM

In the proposed algorithm, distorting technique has been used to hide Sensitive Association Rules based on decrease in the amount of Sensitive Confidence Rule. Reducing the amount of reliability of Sensitive Rules through reducing the amount of support on RHS Association Rule Items is sensitive, and work is done on sets of transactions which completely support the Sensitive Rules. From among Sets of transactions, firstly a transaction is chosen which has the lowest number of Items. In order to Hide Sensitive Rules in this algorithm, the number of transaction, on which the changing process is performed, is calculated and equal to the number of selected transactions, the operation for elimination is performed to remove an item. After changing the transactions the reliability of the Sensitive Rule is calculated, whereby a reduction in the reliability of Sensitive Rule and Hiding of Sensitive Rule is witnessed, and our assessment criteria, in which priority is to Hide Sensitive Rules without Hiding Failure, the second priority is reduction in runtime, and the third priority is that the missing Rules should be minimal, has been fulfilled. For this reason real Datasets such as Chess and Mushroom have been used in this research, and for mining operations “orange” tools have been used. For comparing and evaluation of the results b.1 and ADSRRC algorithms have been used.

C. Algorithm Terminology

The primary Dataset , and a set of sensitive rules that should be hidden, *M*: and *M*(are specified as input. The exit for the algorithm is the changed Dataset , which after the hiding process, does not contain the set which is a Sensitive Rule. All the Sensitive Rules within the set is marked as , and all transactions within the Dataset are marked as . One transaction may have both ends of a Rule , i.e., it may have the prior as well as the result at the same time. In this case we say that the transaction fully supports rule . A transaction may only include prior subsets of Rule (excluding the prior itself), or it may include result subsets of Rule (excluding the result itself). In this case we say that the transaction Partially Support Rule . sets includes transactions that fully support the Rule . For each Rule , is the number or transactions on which the changing operation should be performed in order to hide Rule . If each rule is like (⇒), then is the left hand side item of Rule r, and is the right hand side item of Rule .

```
INPUT: a set of rules to hide, the source Dataset ,
Themin_supthreshold( M: ),
The min_conf threshold( M( ).
OUTPUT: the DataSet transformed so that
 the rules in cannot be mined
Begin
1. Compute confidence of rules
2. Sort rule in ascending order by the confidence them
3. While(all the rules € are not hidden) {
4. Foreach rule in do
 {
5. = { in D / fully supports r}
6. For each transaction of count the number of items in it
7. Sort the transactions in in ascending order of the number items supported
8. Compute .
9. FOR := 1 TO .
   {
10. Choose the transaction in With the lowest number of items ( the first transaction in )
11. Choose the item in
12. delete from
13. Remove from
   }
14. Recomputed the confidence
15. If (conf( ) < min_conf) then Remove from
 }
 }
```

Fig 2. PROPOSED ALGORITHM

D. Stages proposed algorithm

In the proposed algorithm, distorting technique has been used to hide Sensitive Association Rules based on decrease in the amount of Sensitive Confidence Rule. Reducing the amount of reliability of Sensitive Rules through reducing the amount of support on RHS Rule Items and work is done on sets of transactions which completely support the Sensitive Rules. In the distorting technique, we can have a reduction in the set of right hand side Items , or a reduction in the set of left hand side Items . In transactions that fully support Rule , with the reduction of left side items , with regards to the supporting $r \Rightarrow r$ formula, is equal to $Support(r_l, r_r) = \frac{|r_l \cup}{|D}$ there is a decrease in Rule Support, but with regards to confidence rule formula $Conf \Rightarrow r_r) = \frac{Sup(r_l \Rightarrow}{Sup(r_l}$ it has no real effect in reducing the amount of confidence rule. But in these transactions a reduction in the right hand set of items of , reduction in support rule, and a reduction in confidence rule is witnessed. For this reason in the above mentioned algorithm we are faced with a reduction of confidence in sensitive rule through a reduction in right side rule , support. Work is performed on transactions which fully support the sensitive rules.

Steps 1 - Calculate the confidence of the sensitive rules R

Steps 2-Arrange in ascending order the set of sensitive rules Ibased on their values confidence.

Steps 3- The following steps should be repeated until all sensitive rules are hidden R

Steps 4- The following steps should be taken for each rule in the set of sensitive rules R

Steps 5- all the transactions in the Dataset D which fully support the rule , should be placed in .

Steps 6- Count the number of items in each transaction t within the :

Steps 7- arrange in ascending order the transactions within the Tbased on the number of items.

Steps 8- Calculate ..

$$\alpha = \frac{Conf}{MCl} \qquad (1)$$

$$x = \frac{|T_{Full\ Support}(}{\alpha} \qquad (2)$$

$$\beta = [\qquad (3)$$

$$|T_{Modify}| = |T_{Full\ support}(r)| - \qquad (4)$$

$$N_1 = |T_{Modify}| + \qquad (5)$$

Steps 9- Perform the following steps 1 to I

Steps 10- Select the first transaction from within the which has the least number of items.

Steps 11- Select Item from the right hand side items .

Steps 12- delete item from transaction .

Steps 13- delete of transaction from and return to Step 9.

Steps 14- Calculate the Confidence of rule .

Steps 15- If the confidence of rule is lower than the min_con, then remove rule from Iand return to step3. If all the sensitive rules in the I set are hidden, the program ends.

VII. COMPARISON AND EVALUATION OF THE PROPOSED ALGORITHM

Table I- Information Related to Dataset

Dataset Name	Number Of Records	Number Of Items	Number Of Items Per Record
Mushroom	8124	119	23
Chess	3195	75	37

Table II- Information Related to Dataset

Data set Name	MCT	*MST*	All rules generated Data set
Mushroom	%60	%40	3184
Chess	%92	%88	13834

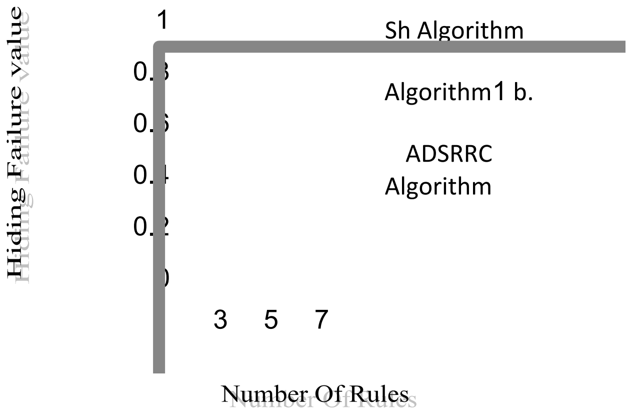

Fig 3.the amount of Hiding Failure for hiding rule 3, 5, and 7 in the Dataset. CHESS

In Fig.3 the amount of Hiding Failure for hiding rule 3, 5, and 7 with three algorithms in the Chess Dataset is equal to zero. The same can be witnessed in the Mushroom Dataset.

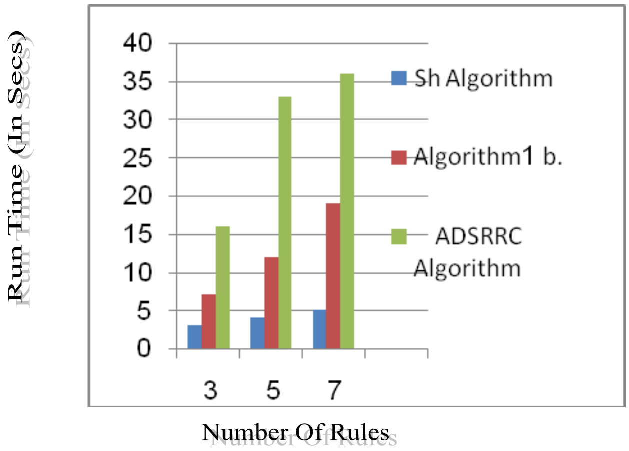

Fig 4. Runtime for hiding rule 3, 5, and 7 in the Dataset CHESS

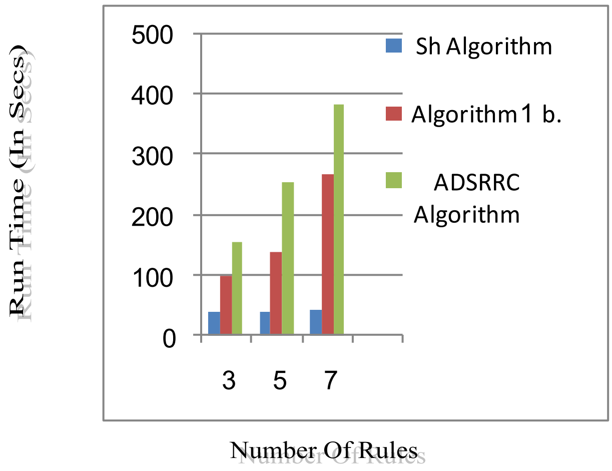

Fig 5. Runtime for hiding rule 3, 5, and 7 in the Data Set.

MUSHROOM

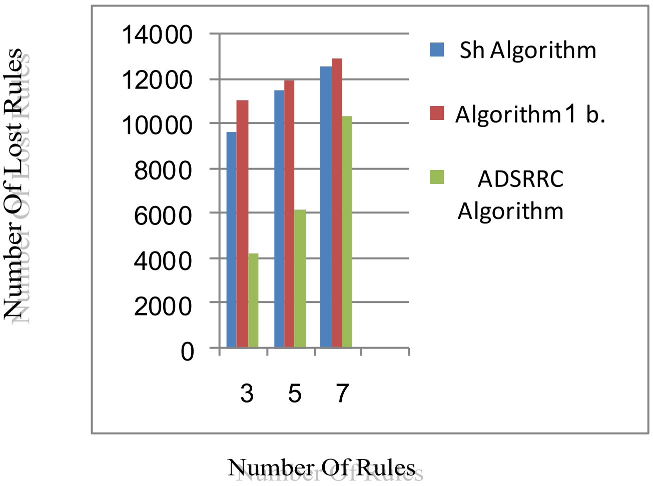

Fig 6. The amount of Lost Rule for hiding rule 3, 5, and 7 in the Dataset. CHESS

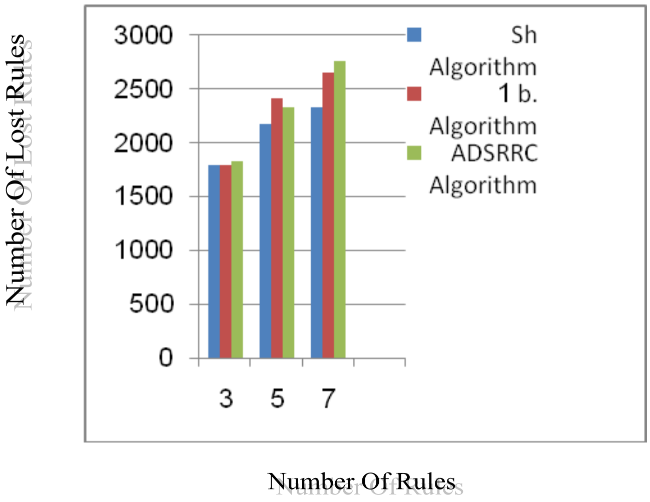

Fig7. The amount of Lost Rule for hiding rule 3, 5, and 7 in the Dataset. MUSHROOM

VIII. PRESENTS FUTURE WORKS, DISCUSSIONS AND CONCLUSION

In Table I, given the column number of the items included in each record, Chess=37 and Mushroom= 23, we can see that in Table II the number of rules produced in Chess Dataset is MST = 88% and MCT = %92, which is much higher than Mushroom Dataset. It is said that the Chess Dataset is denser than the Mushroom Dataset. In Fig.3 the amount of Hiding Failure for hiding rule 3, 5, and 7 with three algorithms in the Chess Dataset is equal to zero. The same can be witnessed in the Mushroom Dataset. In Fig.4 and Fig.5, in the proposed algorithm, firstly through some calculations the number of records to be changed is specified and the same number of selected transactions of the desired Item is deleted. Then the amount of support and confidence is calculated, and it is observed that the amount of confidence is reduced and the rules are hidden. But for hiding any rules in algorithm b.1, the desired item from the transaction is deleted and immediately the amount of confidence and support is measured. This phase will continue until there is a decrease in support or confidence levels till the rules are hidden. Therefore, the runtime for the proposed algorithm is much less than algorithm b.1. Clustering has been used for algorithm ADSRRC. There may be a few rules in one Cluster which have a common Item. This algorithm functions similar to b.1 but with a difference; in this phase instead of calculating the amount of Confidence and support of a rule, this process should be performed for many rules within a Cluster, thus increasing its runtime compared to the other two algorithms. This subject is evident in Fig.4 and Fig.5 in both the Chess as well as Mushroom Datasets for hiding rule 3, 5, and 7. In this case the proposed algorithm, compared with the other two, enjoys a higher performance level. Fig.5 shows the Mushroom Dataset, with regards to the number of its records in Table I is higher as compared to the Chess Dataset. As you can see, the proposed algorithm in this Dataset is of a high performance, followed by algorithm b.1 and then ADSRRC, which in this case due to the use of Clustering, has performed well. In Fig.6 we have the amount of Lost Rule for hiding rule 3, 5, and 7 in the Chess Dataset. The proposed algorithm is better than b.1 algorithm, but in this case ADSRRC algorithm due to the use of clustering enjoys a higher performance level. In Fig.7 we have the amount of Lost Rule for hiding rule 3, 5, and 7 in the Mushroom Dataset. In this case, with regards to the number of its records in the Mushroom Dataset, the proposed algorithm, compared with the other two algorithms, enjoys a higher performance level. Based on the present results, it can be concluded that the proposed algorithm in the two Datasets mentioned above – in terms of number of records, number of rules for hiding, in terms of reduced run time is superior to the other two algorithms. In terms of lost rules, the denser the Dataset the better it is for ADSRRC algorithm, since it uses a Clustering method. If the Dataset is non-compact the proposed algorithm is more effective, even with an increase in the number of hiding sensitive association rules, and the higher number of records in a Dataset. With regards to these results, we can have a combination of proposed algorithm and ADSRRC algorithm for our future work, i.e. to use Cluster in the structure of the proposed algorithm. In this case, we will benefit from the advantages of both algorithms, by shortening the runtime in the dense Datasets, and the number of lost rules would also be reduced.

REFERENCES

[1] Benjamin C. M. Fung, Ke Wang, Ada Wai-Chee Fu, and Philip S. Yu, “Introduction to Privacy-Preserving Data Publishing Concepts and Techniques,” 2011

[2] ArisGkoulalas-Divanis, and Vassilios S. Verykios, “Association Rule Hiding for Data Mining,” 2010.

[3] Charu C. Aggarwal, and Philip S. Yu, “Privacy-Preserving Data Mining Models and Algorithms,” 2008.

[4] M. Atallah, E. Bertino, A. Elmagarmid, M. Ibrahim, and V. S. Verykios, “ Disclosure limitation of sensitive rules,” In Proceedings of the 1999 IEEE Knowledge and Data Engineering Exchange Workshop (KDEX), pp. 45-52, 1999.

[5] E. Dasseni, V. S. Verykios, A. K. Elmagarmid, and E. Bertino, "Hiding association rules by using confidence and support," In Proceedings of the 4th International Workshop on Information Hiding, pp. 369-383, 2001.

[6] S. R. M. Oliveira, and O. R. Zaïane, "Privacy preserving frequent itemset mining," In Proceedings of the 2002 IEEE International Conference on Privacy, Security and Data Mining (CRPITS), pp. 43–54, 2002.

[7] D. A. Simovici, and C. Djeraba, "Mathematical Tools for Data Mining: Set Theory, Partial Orders, Combinatorics," Springer Publishing Company, Incorporated, 2008.

[8] V. S. Verykios, A. K. Emagarmid, E. Bertino, Y. Saygin, and E. Dasseni, "Association rule hiding," IEEE Transactions on. Knowledge and Data Engineering, 16(4): pp. 434–447, 2004.

[9] E. D. Pontikakis, A. A. Tsitsonis, and V. S. Verykios, " An experimental study of distortion based techniques for association rule hiding," In Proceedings of the 18th Conference on Data Set Security (DBSEC), pp. 325–339, 2004.

[10] Mohammad NaderiDehkordi, KambizBadie, and Ahmad KhademZadeh, "A Novel Method for Privacy Preserving in Association Rule Mining Based on Genetic Algorithms," Journal of software, vol. 4, no. 6, August 2009.

[11] Ali Amiri . Dare to share, "Protecting sensitive knowledge with data sanitization," 2006.

[12] Vassilios S. Verykios, and Emmanuel D. Pontikakis, "Efficient algorithms for distortion and blocking techniques in association rule hiding," 2007.

[13] YuhongGuo,"Reconstruction-Based Association Rule Hiding,"2007.

[14] Komal Shah, Amit Thakkar, and Amit Ganatra, "Association Rule Hiding by Heuristic Approach to Reduce Side Effects & Hide Multiple R.H.S. Items,"2012.

[15] D. Hand,H. Mannila and P. Smyth," Principles of Data Mining," MIT Press, Cambridge, 2001.

[16] Oliveira. S.R.M.andZaiane. O.R," Privacy preserving frequent item set mining," In Proceedings of the IEEE ICDM Workshop on Privacy, Security and Data Mining, pp. 43–54, 2002.