A Theoretical Approach for Hiding Sensitive Association Rules

Geeta S. Navale
Sinhgad Institute of
Technology and Science, Pune
Research Scholar
SKNCOE, Pune, India

Suresh N. Mali, PhD
Sinhgad Institute of Technology
and Science, Pune
Principal,
SITS, Pune, India

ABSTRACT
Data mining process is used to extract knowledge from the database. Large numbers of data mining tools are available to get the useful information. These tools can be utilized to break the privacy and security of useful sensitive information present in the database. This sensitive information may be personal information, patterns, facts etc. This sensitive information if mined will result in loss of business logics of database owners. Hence there is a need to hide sensitive knowledge. The hiding process must ensure that the knowledgeshould be mined without disclosing sensitive association rules to the users with minimum impact on nonsensitive association rules. Also, intentional as well as unintentional attackers who are trying to retrieve sensitive association rules should not be successful once they are hidden. In this paper, the authorspropose a methodology to hide sensitive association rules.

Keywords
Privacy Preserving Data Mining, Association Rules Hiding, Knowledge Hiding

1. MOTIVATION
In today's dynamic internet world, a lot of importance has to be given to the security of information while sharing it with different organizations such as e-business enterprises, government agencies, online company sellers etc. Present-day advanced data mining techniques discovers unknown information about various associations of itemsets in large amount of data stored in repositories which will be useful to the organizations to increase their businesses. However, in many cases the organizations having the source of information may not intended to share such associations with any other organization or they want to have rights of information. Therefore, there is a need to hide the sensitive associations in the database former to share it among in the organizations.

The rest of the paper is arranged as follows. In Section 2 the authors present some related information and the terms that are used in the rest of the paper. Section 3 describes the proposed system to hide sensitiveassociation rules. Finally, the authors clinch our discussion in Section 4.

2. RELATED INFORMATION AND TERMS
Agarwal et al. [1] were the first to introduce and mine association rules. Let us consider simple example as shown in Table 1.

Table 1. Transaction Database

T-id	Items
101	x, y
102	x, z, u, v
103	y, z, u, w
104	x, y, z, u
105	x, y, z, w

Itemset: Collection of one or more items from any transaction is called as Itemset. For example {x, y, z, u} is Itemset.

There are various measures based on which association rules can be mined [2][3][4][5][6]. Here the authors are considering two measures support and count.

Support Count: It is frequency of occurrence of an Itemset I in the given transactions.

For example:

1. Support Count of ({y, x}) = 3 and
2. Support Count of ({y, x, z}) = 2

Confidence Count: It is the strength of relation between set of items.

For example:

The authors have considered two thresholds 'S_min' minimum support count threshold and 'C_min' minimum confidence count threshold.

Frequent Itemset: An Itemset whose support count is greater than or equal to 'S_min' threshold is called as Frequent Itemset.

Mining the association rules is atwo-step process: finding the Frequent Itemset and then finding the association rule. Therefore the process of discovery of association rules is as shown in Figure 1.

After obtaining Frequent Itemset (F), one can easily discover any interesting Association Rules such as X $\rightarrow$ Y [1].

Finding Frequent Itemset generally is pretty easy but very costly as size of the data base increases. The simple method would be to count all Itemsets that appear in any transaction in a given database. Given a set of items of size 'm', there are

2^m subsets that are possible. The possible number of itemsets is $2^m - 1$.

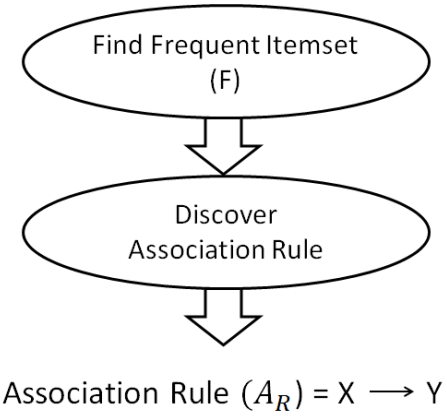

Fig 1: Use of Frequent Itemset to find Association Rule

With the help of database transactions, items, frequent itemsets and threshold values of Confidence (C_min) and Support (S_min) one can develop as algorithm to extract all possible association rules as shown in Figure 2.

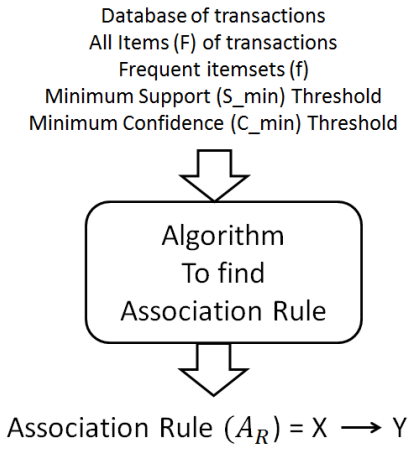

Fig 2: Use of Frequent Itemset, Support, and Confidence to find Association Rule

As the minimum threshold support 'S_min' decreases, the number of frequent itemsets increases. This gives a real challenge for the data mining process.

3. PROPOSED SYSTEM FLOW

The authors of [7][8][9][10][11] used support and confidence measures to hide sensitive association rules. In our proposed system the authors are also using these two measures to hide sensitive association rules.

Association rules are generated using the data mining tool from the original scalable database. The selected association rules based on given threshold values of 'Support' and 'Confidence', eventually called Sensitive Association Rules (SARs) are considered while modifying the original scalable database 'D' to draft copy of sanitized database '$D^{S`}$'. After enhancing various parameters of association rule hiding criterions, the sanitized database will get finalized as 'D^{S}'. The embedding function will use a secret key 'K' to convert 'D' to 'D^{S}'.

The embedding secret key 'K' will be function of embedding methodology and going to reflect the side information used at the time of embedding and necessary while extracting the original database D. This embedding secret key has to be shared through secured channel for protecting the privacy amongst the trusted organizations.

The intended user will use the same secret key 'K' and extraction function to get the original database 'D'.

The flow of the proposed system as shown in Fig. 3computes the inverted index of the transaction database. If the support of antecedent element of sensitive association rule is greater than the support of consequent element, hide the consequent element. The process will be repeated until all sensitive association rules are hidden to get sanitized database.

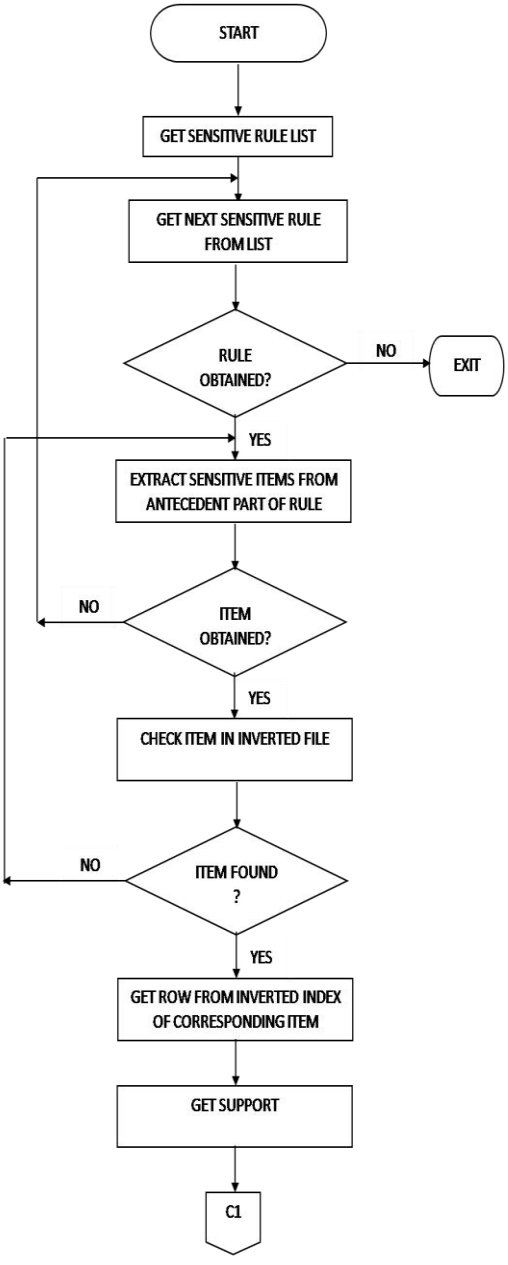

Fig 3: Flow of the Proposed System

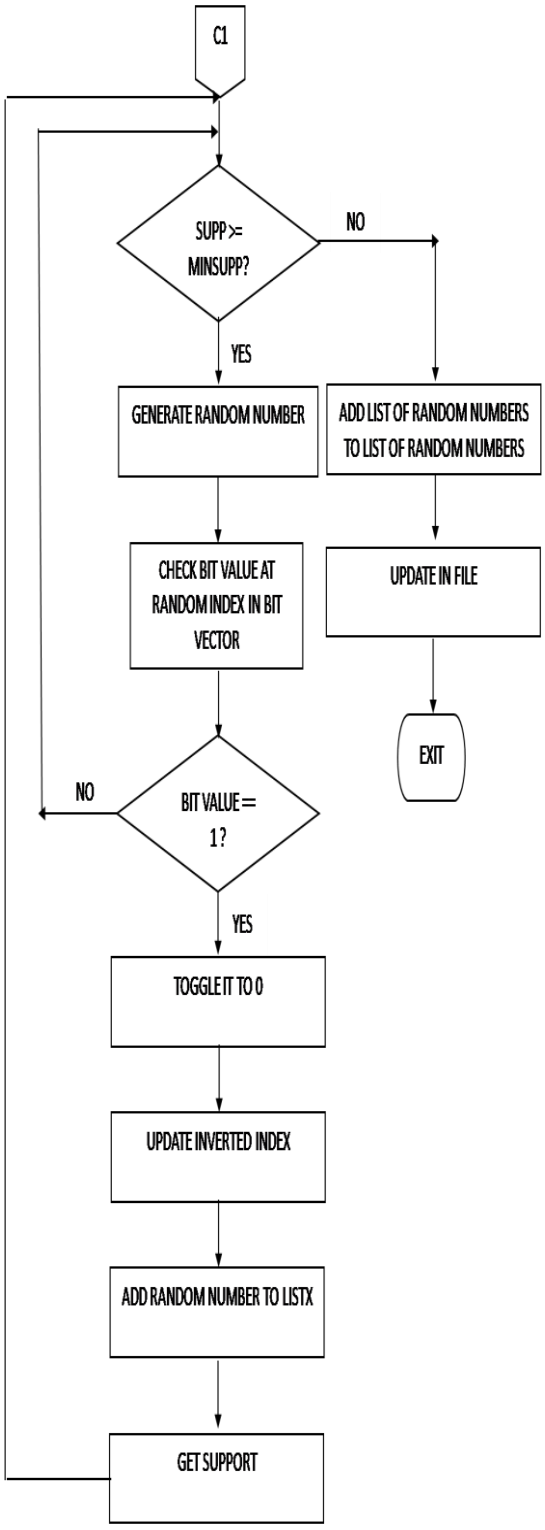

Fig 3: Flow of the Proposed System

4. RESULTS

The proposed method is implemented on CPU-i74710HQ, RAM-16 GB DDR3, SSD-256 GB. The dataset that the authors have used is Mushroom. Table 2 shows the number of rules to be hidden and hiding failure. Hiding failure is nothing but amount of sensitiverules that can beaccessed even afterapplying hiding process from sanitized database.The amount of sensitive rules that can be discovered after applying hiding process i.e. hiding failure must be zero. The proposed method is implemented to hide 1, 3, 5, 7 and 9 sensitive rules and checked for hiding failure.

Table 2.Hiding Failure

Number of Rules to be Hidden	Hiding Failure
1	0
3	0
5	0
7	0
9	0

5. CONCLUSION AND FUTURE WORK

Ensuring security in data mining activities is an incredibly basic issue in various applications.The authors presented the need for privacy preserving in data mining and description for the proposed system to hide sensitive association rules.

In future the implemented method will be checked to achieve better optimizationwith minimum side effects.

6. ACKNOWLEDGMENTS

The authors thanks to Sinhgad Institute of Technology and Science, Pune as well as to Smt. Kashibai Navale College of Engineering, Pune for technical support.

7. REFERENCES

[1] Agrawal, R., Imielinski, T., Swami, "A.: Mining Association Rules between Sets of Items in Large Databases," ACM SIGMOD International Conference on Management of Data (SIGMOD'93), Washington D.C., USA , pp. 207–216, May 1993

[2] Geng L., Hamilton, H.J., "Interestingness Measures For Data Mining:A Survey," ACM Comput. Surv. (CSUR) 38(3), 9, 2006.

[3] McGarry, K, "A survey of Interestingness Measures for Knowledge Discovery," Knowl. Eng. Rev. 20(1), pp. 39–61, 2005.

[4] G. Dong and J. Li, "Interestingness of Discovered Association Rules in terms of Neighbourhood-Based Unexpectedness," Second Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD'98), pp. 72-86, 1998.

[5] Lallich, S., Teytaud, O., Prudhomme, E., "Association Rule Interestingness: Measure and Statistical Validation," Qual. Measur. Data Min. 43, pp. 251–276, 2006.

[6] A.A. Freitas, "On Rule Interestingness Measures," Elsevier, Knowledge-Based Systems 12, pp. 309–315,1999.

[7] Dasseni E, Verykios VS, Elmagarmid AK, Bertino E., "Hiding Association Rules by Using Confidence and Support," In: Proceedings of the 4th International Workshop on Information Hiding, pp. 369–383, 2001.

[8] Verykios VS, Pontikakis ED, Theodoridis Y, Chang L., "Efficient Algorithms for Distortion and Blocking

Techniques in Association Rule Hiding," Distributed Parallel Databases, pp. 22:85–104, 2007.

[9] Wu Y-H, Chiang C-M, Chen ALP., "Hiding Sensitive Association Rules with Limited Side Effects," IEEE Trans Knowledge Data Eng, pp. 19:29–42, 2007.

[10] Gkoulalas-Divanis A, Verykios VS., "Hiding Sensitive Knowledge Without Side Effects," Knowledge Inf Syst, pp. 20:263–299, 2009.

[11] Gkoulalas-Divanis A, Verykios VS., "Exact Knowledge Hiding Through Database Extension," IEEE Trans Knowledge Data Eng, pp. 21:699–713, 2009.