INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH & DEVELOPMENT

ISSN: 2278 – 0211 (Online)

Privacy Preserving Large Scale Transactional Data Set Using State Transitional Matrix

Vijaykumar
Research Scholar, Department of Computer Science
Bharathiyar University, Coimbatore, Tamil Nadu, India

Dr. R.Manicka Chezian
Associate Professor, Department of Computer Science
NGM College, Pollachi, Tamil Nadu, India

Abstract:

Privacy preservation becomes more important task in various areas like business and social environment. In both the areas hiding privacy information is necessary to hiding personal information of users from others. We propose a new knowledge hiding and privacy preserving methodology using state transitional matrix. We generate transitional matrix from the large transaction data set, which will be published for other users of the business or social environment. Published data could be used to infer some knowledge but they could not back track the personal information from the sanitized data base. The proposed method is a simpler one which reduces the time and space complexity.

Key words: *Privacy preservation, state transition matrix, transactional data set, knowledge hiding.*

1.Introduction

Knowledge hiding is a process of sanitizing the original data base with loss of originality for the view of others. While sanitizing the originality of the data set has to be retained and also should be useful for other to make some inference from the data set published.

Most organizations publish their micro data to others for various purposes. For example hospitals publish their data for research purpose and online shopping companies publish their data for analysis of product for marketing improvement and so on. While publishing there will be personal information, which has to be hidden in order to provide privacy for the personal information. Earlier method hides certain attributes but the attackers could easily retrace the record to identify the person whose details it is, but anonymization of personal data is not sufficient in some applications.

Consider, for instance, the example of a large retail company which sells thousands of different products, and has numerous daily purchase transactions. The large amount of transactional data may contain customer spending patterns and trends that are essential for marketing and planning purposes. The company may wish to make the data available to a third party who can process the data and extract interesting patterns (e.g., perform data mining tasks). Since the most likely purpose of the data is to infer certain purchasing trends, characterized by correlations among purchased products, the personal details of the customers are not relevant, and are altogether suppressed. Instead, only the contents of the shopping cart are published for each transaction. Still, there may be particular purchasing habits that disclose customer identity and expose sensitive customer information.

	Wine	**Strawberry**	**Meat**	**Cream**	**Pregnancy Test**	**Viagra**
Raja	×		×			×
Mahesh	×		×			
Siva		×		×	×	
Kumar		×	×			
Prabu	×		×	×		

Table 1

Consider the example in Fig. 1a, which shows the contents of five purchase transactions. The sensitive products (items), which are considered to be a privacy breach if associated

with a certain individual, are shown shaded. The rest of the items, which are non sensitive, can be used by an attacker to re identify individual transactions, similarly to a quasi-identifier, with the distinctive characteristic that the number of potentially identifying items is very large in practice (hence, the QID has very high dimensionality). Consider the transaction of Claire, who bought a pregnancy test. An attacker (Eve) may easily learn about some of the items purchased by Claire on a certain day, possibly from a conversation with her, or from knowing some of her personal preferences. For instance, Claire may treat her guests, including Eve, with fresh cream and strawberries, and Eve can, therefore, infer that Claire must have purchased these items recently. By joining this information with the purchase transaction log, Eve can re identify Claire's transaction and find out that Claire may be pregnant.

To overcome this there must be a suitable method which should not disclose the privacy items and personal information, so that an attacker could not be able to identify or infer others purchase and so on. Also in case of business point of view the people's identity should not be visible; otherwise the product goal will not reach all the people.

In this paper we consider the purchase of an as a state and there will be many state in the transactional data set. Each row in the transactional data set has many numbers of states and end up with a state. We make it as transition matrix for the use of sanitization which will not show any direct dataset and will show a transition one which is a probabilistic and could be used by other for inference but a difficult one.

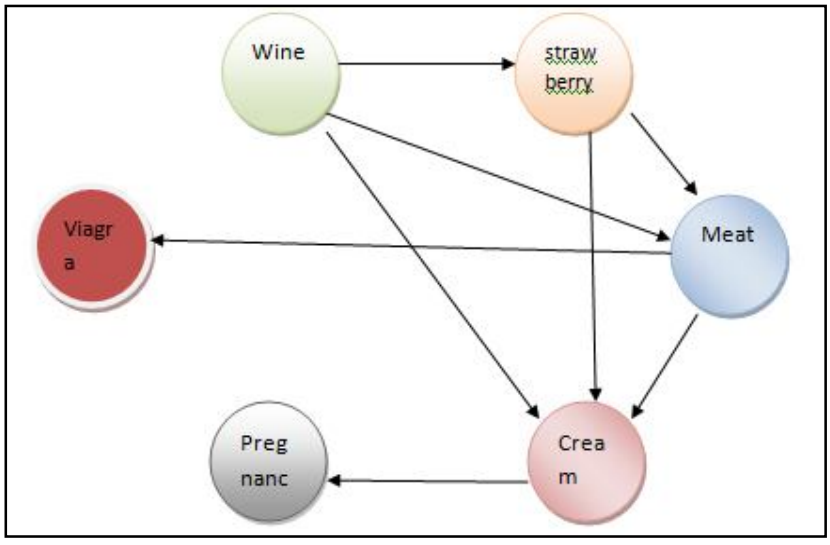

Figure 2: shows the state transition diagram

The fig2 shows the state transition diagram and it shows that there are various states and transitions between different states or items. It is very clear that there is only on transition from Meat to Viagra and Cream to Pregnancy. So that the state transition diagram and matrix is useful and easier for representation. It is not possible to find out the pattern easier and generates complete sanitized data set.

2.Background

There are many implementations for privacy preservation; here we discuss few of them. In [1] they proposed a random perturbation method to prevent re identification of records, by adding noise to the data. By adding noise to the data it loses the originality and it becomes incomplete one. In [2], it is shown that an attacker could filter the random noise, and hence, breach data privacy, unless the noise is correlated with the data. However, randomly perturbed data are not "truthful" [3] in the sense that it contains records which do not exist in the original data. Furthermore, random perturbation may expose privacy of outliers when an attacker has access to external knowledge.

Using quasi identifier to the records used in many methods, but in that case the original record can be matched with the few attributes and can be identified. To overcome this difficulty Samarthi in [4] introduced k-anonymity, a privacy-preserving paradigm which requires each record to be indistinguishable among at least k _ 1 other records with respect to the set of QID attributes. Records with identical QID values form an anonymized group. K-anonymity can be achieved through generalization, which maps detailed attribute values to value ranges, and suppression, which removes certain attribute values or records from the microdata.

LeFevre in [3], proposed optimal k-anonymity solutions for single-dimensional recoding, in that they used independent mapping for each attribute and in [4], he proposed a multidimensional recoding algorithm which maps the Cartesian product of multiple attributes. Mondrian outperforms optimal single-dimensional solutions, due to its increased flexibility in forming anonymized groups. Methods discussed so far perform global recoding, where a particular detailed value is always mapped to the same generalized value.

Bucketization [5, 6, 7] first partitions tuples in the table into buckets and then separates the quasi-identifiers with the sensitive attribute by randomly permuting the sensitive attribute values in each bucket. The anonymized data consists of a set of buckets with

permuted sensitive attribute values. In particular, bucketization has been used for anonymizing high-dimensional data [8].

In contrast, local recoding allows distinct mappings across different groups. Clustering-based local recoding methods are proposed in [9]. Anatomy [14] introduced a novel approach to achieve diversity: instead of generalizing QID values, it decouples the SA from its associated QID and permutes the SA values among records. Since QIDs are published directly, the information loss is reduced. A similar approach is taken in [11]. However, neither of these methods account for correlation between the QID and the SA when forming anonymized groups. We also adopt a permutation approach for transactional data, but we create anonymized groups in a QID-centric fashion, therefore preserving correlation and increasing data utility. Furthermore, our novel data representation helps us tackle the challenge of high-dimensional QID.

Privacy preservation of transactional data has been acknowledged as an important problem in the data mining literature. However, existing work [12], [13] focuses on publishing patterns, and not data. The patterns (or rules) are mined directly from the original data, and the resulting set of rules is sanitized to prevent privacy breaches. Such an approach has two limitations: 1) the usability of the data is constrained by the rules that the owner decides to disclose and 2) it is assumed that the data owner has the resources and expertise to perform advanced data mining tasks, which may not be the case in practice. We choose to publish the data, instead of patterns; this gives the recipient flexibility in choosing what rules to mine, and also allows for other types of data analysis, such as clustering. Furthermore, the processing cost need not be bared by the data owner.

All these methods have the problem of maintaining privacy and disclose privacy little bit and produces inefficient results. We propose a new transition matrix based privacy preservation technique for the problem of privacy preservation.

3.Proposed Method

The proposed method has three phases preprocessing, state transition matrix computation, matrix publishing.

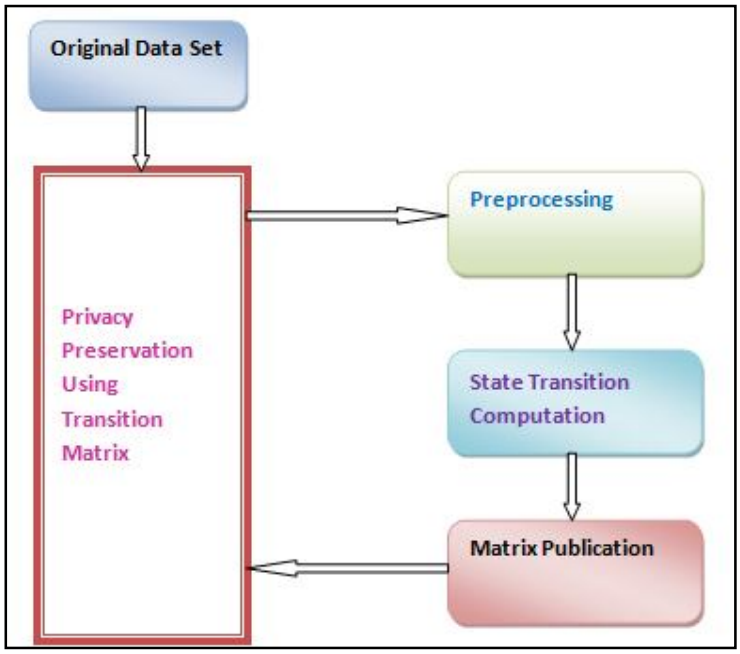

Figure 3: Proposed system Architecture

4.Preprocessing

At preprocessing stage the transactional dataset is loaded and we identify the schema of the data set. From the schema S we collect the item set I, present in the data set D_s. once we identified the item set , we initialize the occurrence matrix OM with the number of columns m with the size of item set I and column with the size of item set I. We compute the number of occurrence of item from the item set with the data set Ds and store them in the occurrence matrix, which will be used for transition matrix computation later.

5.Algorithm

- Step1: load transactional dataset D_s.
- Step2: Identify No of items N.
- Step3: Initialize Occurrence matrix OM.
- Step4: count number of records T in D_s.
- Step5: for each item I_i in D_s
  - Noc= Count Number of occurrence in all records T.
  - OM[I_i]=Noc.
  - Store Noc in Occurrence matrix.
  - End.
- Step6: End.

6.State Transition Matrix Computation

We compute the transition of states using the occurrence matrix and we compute the probability of transition using the occurrence matrix. Initialize the transition matrix with the size of I×I. For each item I_i in the occurrence matrix , compute the transition probability and store in the transition matrix.

7.Algorithm

- Step1: Read total No of items N.
- Step2: Initialize Transition matrix TR with size $TR_{m\times n.}$
- Step3: Read occurrence matrix OM.
- Step4:for each item I_i in item set I

    For each item I_j in item set I

        Identify number of occurrence of I_i from OM.

            Q= OM(I_i);

        Identify number of occurrence of I_j from OM.

            R= OM(I_j);

        Compute probability of transition from Q to R as follows.

            Pb = Q/R.

        Assign Pb in the transition matrix TR.

        - $TR_{(I,j)}$=pb.

    End

End

- Step5: End.

8.Matrix Publication

Generated state transition matrix for the transactional data base contains only the probability values. One can easily understand and infer some knowledge for useful purposes. The generated transition matrix is completely sanitized and there is no need to hide anything. The originality of the data set is maintained because, we will publish only the transition matrix not the original one. The user can infer and compute sample matrixes which represent the transition matrix but they cannot identify the exact one.

9.Result and Discussion

The proposed method generates good results compare to all other algorithms discussed earlier in this area. We used state transition matrix which contains only the probabilistic values not the original records, but still the user can generate a data set which resembles the transition matrix but may not be the original one. The proposed method retained the originality and also it preserves the private information. At this stage even if we declare the names with the matrix the user cannot identify which record belongs to one and it is not possible to identify the purchase pattern of the user.

	Wine	**Strawberry**	**Meat**	**Cream**	**Pregnancy test**	**Viagra**
Raja	×		×			×
Mahesh	×		×			
Siva		×		×	×	
Kumar		×	×			
Prabu	×		×	×		

Table 2: shows the original data set

The table 2 shows the example data set used in our method and it contains various purchase patterns and has privacy attributes which has to be hidden.

Wine	**Strawberry**	**Meat**	**Cream**	**Pregnancy Test**	**Viagra**
3	2	4	2	1	1

Table 3: shows the Occurrence Matrix

Table 3 shows the occurrence matrix generated by our algorithm and it represents the occurrence of items throughout the transactional data set. It shows the frequency of all items in the transactional data set.

Now using transition matrix generation method produces the matrix with both the transaction data set and occurrence matrix. It computes the probability of transition which represents the probability of purchasing items in one to one and one to many.

	Wine	**Strawberry**	**Meat**	**Cream**	**Pregnancy test**	**viagra**
Wine		0.0	1.0	0.3	0.0	0.3
Strawberry	0.0		0.5	0.5	0.5	0.0
Meat	1.0	0.25		0.25	0.0	0.25
Cream	0.5	0.5	0.5		0.5	0.0
Pregnancy	0.2					
viagra	0.2					

Table 4: Sanitized data set

Table 4 shows the output of state transition matrix with probability values. It shows clearly that each item have probability with others and but for the sensitive items we have given common probability and not associated with each item. So that the user can regenerate the data set but not the original one.

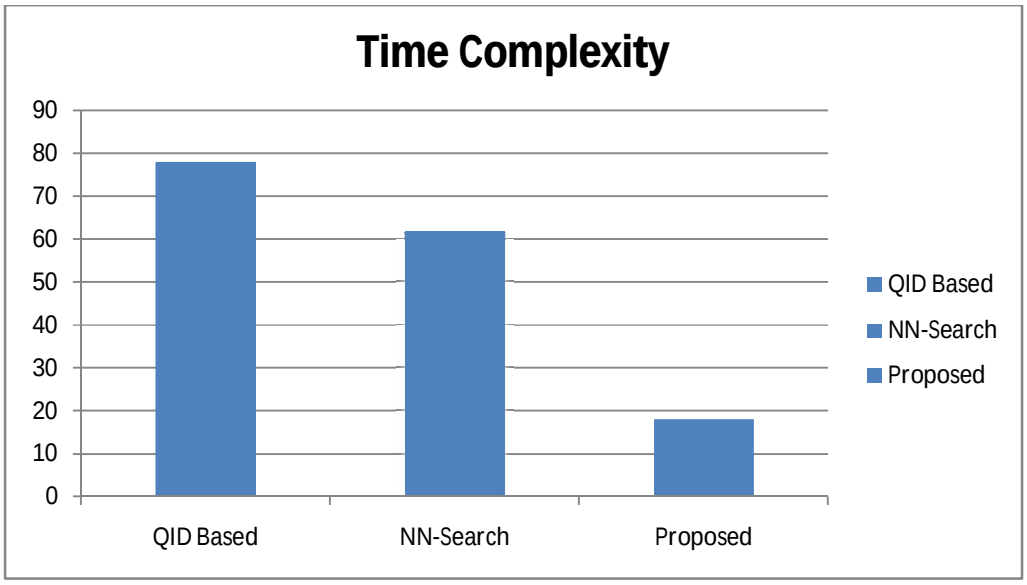

Figure 4: shows the time complexity between other methods.

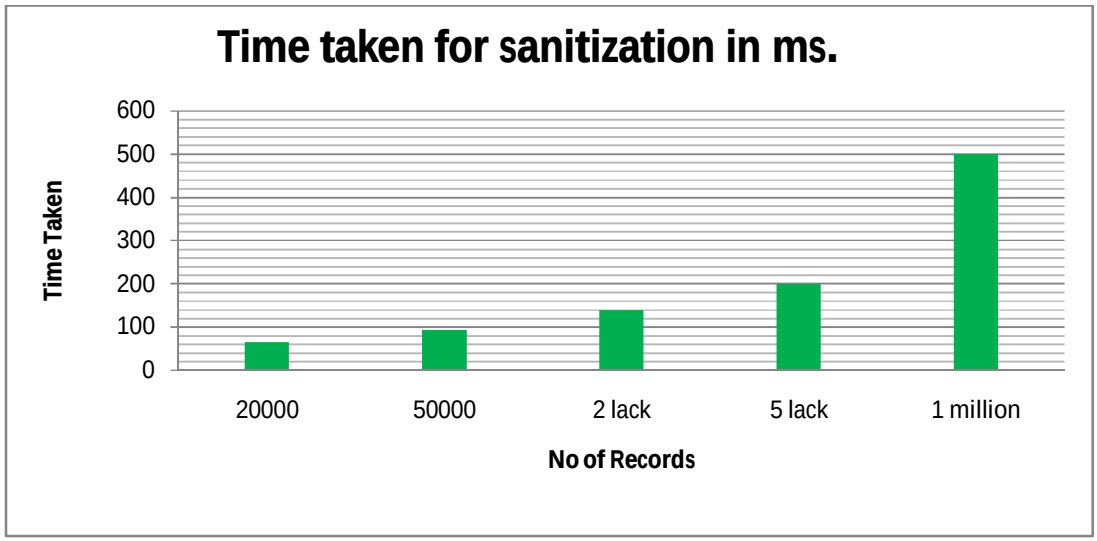

Figure 5 : shows the overall time taken for sanitization process

10.Conclusion

We proposed a new state transition based sanitization process for knowledge hiding and privacy preservation. The proposed method produces good results and it is efficient in time and space complexity. The data set published contains only the probability values not the original one, still the user can infer some knowledge but they cannot re generate the data set which is original.

11.Reference

1. G. Ghinita, Y. Tao, and P. Kalnis, "On the Anonymization of Sparse, High-Dimensional Data," Proc. IEEE Int'l Conf. Data Eng. (ICDE), pp. 715-724, 2008.
2. A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, "l-Diversity: Privacy beyond k-Anonymity," Proc. IEEE Int'l Conf. Data Eng. (ICDE), 2006.
3. K. LeFevre, D.J. DeWitt, and R. Ramakrishnan, "Incognito: Efficient Full-Domain k-Anonymity," Proc. ACM SIGMOD, pp. 49-60, 2005.
4. K. LeFevre, D.J. DeWitt, and R. Ramakrishnan, "Mondrian Multidimensional k-Anonymity," Proc. IEEE Int'l Conf. Data Eng.(ICDE), 2006.
5. X. Xiao and Y. Tao. Anatomy: simple and effective privacy preservation. In VLDB, pages 139–150, 2006.
6. D. J. Martin, D. Kifer, A. Machanavajjhala, J. Gehrke, and J. Y. Halpern. Worst-case background knowledge for privacy-preserving data publishing. In ICDE, pages 126–135, 2007
7. N. Koudas, D. Srivastava, T. Yu, and Q. Zhang. Aggregate query answering on anonymized tables. In ICDE, pages 116–125, 2007.
8. G. Ghinita, Y. Tao, and P. Kalnis. On the anonymization of sparse high-dimensional data. In ICDE, pages 715–724, 2008.
9. G. Aggarwal, T. Feder, K. Kenthapadi, S. Khuller, R. Panigrahy, D. Thomas, and A. Zhu, "Achieving Anonymity via Clustering," Proc. ACM Symp. Principles of Database Systems (PODS), pp. 153- 162, 2006.
10. G. Ghinita, P. Karras, P. Kalnis, and N. Mamoulis, "Fast Data Anonymization with Low Information Loss," Proc. Int'l Conf. Very Large Data Bases (VLDB), pp. 758-769, 2007.
11. Q. Zhang, N. Koudas, D. Srivastava, and T. Yu, "Aggregate Query Answering on Anonymized Tables," Proc. IEEE Int'l Conf. Data Eng. (ICDE), pp. 116-125, 2007.
12. M. Atzori, F. Bonchi, F. Giannotti, and D. Pedreschi, "Anonymity Preserving Pattern Discovery," VLDB J., vol. 17, pp. 703-727, 2008.
13. V. Verykios, A. Elmagarmid, E. Bertino, Y. Saygin, and E. Dasseni, "Association Rule Hiding," IEEE Trans. Knowledge and Data Eng., vol. 16, no. 4, pp. 434-447, Apr. 2004.

14. X. Xiao and Y. Tao, "Anatomy: Simple and Effective Privacy Preservation," Proc. Int'l Conf. Very Large Data Bases (VLDB), pp. 139-150, 2006.