The International Journal Of Engineering And Science (IJES)
||Volume||2 ||Issue|| 4 ||Pages|| 53-57 ||2013||
ISSN(e): 2319 – 1813 ISSN(p)**:** 2319 – 1805

Privacy Preservation Using Randomized Attribute Selection Based On Knowledge Hiding

[1,]P.Vijayakumar, [2,]Dr. R.Manicka chezian
[1,]*Research Scholar, Department of Computer Science,Bharathiar University Arts and Science College,Valparai, Coimbatore, Tamil Nadu, India*
[2,]*Associate Professor, Department of Computer Science NGM College, Pollachi, Tamil Nadu, India*

----------ABSTRACT----------
Privacy preservation of user transactions has highly important while publishing the transactional data sets for further usage. There have been various methods to generate synthetic data sets which maintain the originality of the data set and hide the personal information's. In order to preserve personal information's in multi-dimensional transactional data sets we propose a data hiding technique which identifies the item to be hidden and generates the synthetic publishing data where the sensitive pattern is unknown. Among the proposed technique sensitive patterns are identified using IP-Search in the data set. At the second stage the proposed method generate the dot-matrix which is the publishing data to the external world. The proposed system takes the raw transactional data set as input and we identify the frequent patterns of purchase from the original data set based on the frequency of purchase pattern, we identify the interested patterns for a specific user and those identified interested patterns are sanitized using dot matrix operation. Randomized selection is applied to select the attribute to be sanitized in the dot matrix operation.

Keywords: *privacy preservation, knowledge hiding, frequent pattern.*

Date Of Submission: 26 march 2013  Date Of Publication: 20, April.2013

I. INTRODUCTION

The rapid growth of technology in data processing and knowledge mining, there is a huge threat to the privacy information which can be identified and misused. Organizations share information's for better growth, on publishing own transactional dataset the privacy information has to be sanitized. For example a food item manufacturer publishes its transactional dataset of its own buyers for a joint venture with a cosmetic manufacture. Upon publishing the food manufacture has to hide little sensitive or privacy information. The food manufacturer's data set contains personal information's like name, age, salary, and employer and so on. So that he has to identify what are the privacy information has to be sanitized before publishing. If those personal information's are not identified and preserved those personal information's can be wrongly used by other peoples and lead to serious damages. The same purchase data set is used in various fields like finance, marketing, medical, and statistical. While used in medical field the food habit of people is used to identify the related disease may be affected, or to infer the food habit which becomes the reason for a particular disease. On this way if the personal information of the purchases is disclosed then it affects the personalization of the purchaser and the dataset user can easily reach the persons which breach the personalization.

II. BACKGROUND

There exists various researches in this area and we discuss few of them here. In Hiding Sensitive Association Rules with Limited Side Effects [1], a limited side-effect approach is proposed that modifies the original database to hide sensitive rules by decreasing their supports or confidences. At the stage one it classifies all the valid modifications related to the sensitive rules, the non-sensitive rules, and the spurious rules that they can affect when applied. At stage two, it uses heuristic methods to modify the transactions in an order that increases the number of hidden sensitive rules, while reducing the number of modified entries.

In "Dare to Share: Protecting Sensitive Knowledge with Data Sanitization," A.Amiri [2] has proposed three data sanitization heuristics which increases the data utility and increases the computational speed. The first method deletes few item sets which increases the support of item sets until the support is reduced. The second method increases few fake records to reduce the support of the sensitive items.

A Border-Based Approach for Hiding Sensitive Frequent Itemsets [3] introduces a border-based approach (BBA) for frequent item set hiding. It's a greedy approach in nature and focuses on preserving the quality of the border constructed by the non-sensitive frequent item sets in the item set lattice. The authors use the positive border, after it has been shaped up with the removal of the sensitive item sets, to keep track of the impact of altering transactions in the database.

In Maximizing Accuracy of Shared Databases, Menon [4] present an integer programming approach for the hiding of sensitive item sets. The algorithm treats the hiding process as a CSP that identifies the minimum number of transactions to be sanitized. The authors first reduce the size of the CSP by using constraints involving only the sensitive item sets and then solve it by using integer programming. A heuristic is then enforced to identify the actual transactions and sanitize them.

Adistortion approach that is also based on integer programming is presented in "An Integer Programming Approach for Frequent Itemset Hiding," in the work of Gkoulalas- Divanis and Verykios [5]. The authors propose an exact methodology that relies on the process of border revision to identify the least amount of candidate items for sanitization.As a consequence, the provided hiding solution is guaranteed to minimally distort the original transactions (as opposed to appending new transactions like it is proposed in this paper) to accommodate for knowledge hiding.

Privacy-preserving data publishing has enforced in various ways in [6] random perturbation to prevent re-identification of records, by adding noise to the data is proposed. An attacker could filter the random noise, and hence, breach data privacy, unless the noise is correlated with the data. Randomly perturbed data generated using [7] contains records which do not exist in the original data. Random perturbation may expose privacy of outliers when an attacker has access to external knowledge. Published data about individuals (microdata) contain QID, such as age, or zip code, which can be joined with public databases (e.g., voting registration lists) to re-identify individual records. To prevent this threat, 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 is proposed by samarati. Records with identical QID values for outcome 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 micro data. The process of data anonymization is called recoding, and it inadvertently results in information loss.

Several privacy-preserving techniques have been proposed, which attempt to minimize information loss, i.e., maximize utility of the data. LeFevre et al. [8] proposed optimal k-anonymity solutions for single-dimensional recoding, which performs value mapping independently for each attribute. In [9], the same authors introduced Mondrian, a heuristic solution for multidimensional recoding, 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. In contrast, local recoding allows distinct mappings across different groups. Clustering-based local recoding methods are proposed in [1].

K-anonymity prevents reidentification of individual records, but it is vulnerable to homogeneity attacks, where many of the records in an anonymized group share the same sensitive attribute (SA) value. '-diversity [2] addresses this vulnerability and creates anonymized groups in which at least 'SA values are well represented. Any k-anonymity technique can be adapted for '-diversity; however, this approach typically causes high information loss. The work in [15] proposes a framework based on dimensionality mapping, which can be tailored for k-anonymity and '-diversity, and outperforms other generalization techniques. However, dimensionality mapping is only effective for low-dimensional QIDs; hence, the method is not suitable for transactional data. Furthermore, existing '-diversity methods work for a single sensitive attribute, whereas in our problem, we need to consider a larger number of sensitive items. The work in [9] considers that external knowledge is available to an adversary, in the form of logical constraints on data records.

However, the solution proposed targets relational (i.e., low dimensional) data. Anatomy [6] 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 [16]. 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. Privacy preservation through association rule mining is proposed in [30], which modifies the generated support and count values to generate rules.In Slicing: A New Approach to Privacy Preserving Data Publishing [31] , selection of attribute is difficult because the system does not know which is the sensitive attribute. To solve this we propose a new technique to select the attribute to be hidden using randomized technique.

III. PROPOSED METHODOLOGY

We propose a new hiding technique which reduces the anonymity. The proposed technique involves in a two stage process. At the stage one it performs IP-search, which identifies interest patterns based on the occurrence of items in the transactional data sets. From the generated interested patterns, few privacy items are identified. At the second stage of the process sanitized dataset is generated using dot-matrix technique.

Table I: view of original data set

Name	wine	cream	smoking	cancer
raj	1	0	0	0
siva	1	1	1	1
kumar	0	1	1	0
david	0	0	1	1
ravi	1	0	1	1
muthu	0	0	1	1
selva	1	1	0	0
bharat	0	1	1	0
prabu	0	1	0	0

For example Table I. Shows the original the transactional data set, using which medical industries can identify the habits which generates cancer. While publishing the dataset the privacy information like name and item cancer has to be hidden in order to preserve the privacy of the persons.

3.1 IP-Search

Identifying the sensitive item and patterns are done using IP-search. The original data set is used as input and interested patterns are returned as output. For each Transaction T_t in the data set D_t, we calculate Number of positive values for items in the transaction T_t as N_{pt}. Sort each transaction T_t in dataset D_t using N_{pt}. From the set of transaction in D_t calculate, for each transaction item T_i calculate Number of occurrence N_c and Frequency of occurrence F_c.

Number of occurrence N_c=Total Number of positive values in the transactional dataset.

Frequency of occurrence F_c:- N_c/Total number of rows T_r.

3.2 Algorithm

Input:- Original data set

Output:- Interested pattern I_p and privacy item P_t.

Step1:- Read input data set

Step2:- Select all records for processing.

Step3:- Sort records by number of positive occurrence of items N_c.

Step4:- Calculate F_c.

Step4:- Select the item which has more occurrence frequency value F_c.

Step5:- Select name as one of the private item.

Step6:- select private items which have fewer occurrences F_c.

Step7:- assign more occurring item to interested patterns.

Step8:- return privacy item P_t and interest pattern I_p.

Table II: Result of occurrence

Field name	wine	cream	Smoking	cancer
siva	1	1	1	1
selva	1	1	0	0
ravi	1	0	1	1
bharat	0	1	1	0
kumar	0	1	1	0
david	0	0	1	1
raj	1	0	0	0
muthu	0	0	1	1
prabu	0	1	0	0

Table II. Displays the result after sorting the items according to the occurrence. From this data set the item which has high occurrence value is identified, identified item is selected as interested pattern and which has low occurrence has selected as private item which has to be sanitized.

3.3 Dot-Matrix

After generating the interested pattern and identifying privacy items, we generate the dot matrix. The dot matrixes represent the whole transactional data set with originality preserved. The dot matrix is generated as follows:From the sorted matrix from IP-Search, we insert few random rows and we generate dot values for the items which are identified as private item by the IP-search.

3.4 Algorithm

Step1:- read all rows from the data set D_t.
Step2:- generate random rows R_t with random values for the items.
Step3:- merge both data sets(D_t and R_t).
Step4: for each row in the data set M_t
Select a random Number.
Identify whether it is a prime or not.
If (prime)
Assign dot to the private item
Else
Assign true value to the private item.
Step 5:- release data set M_t.

Table III: result of dot matrix

Name	wine	cream	smoking	cancer
Siva	1	1	1	
Selva	1	1	*	
Ravi	1	0	*	
Bharat	0	1	1	
Kumar	0	1	*	4
Picaso	0	0	1	
Merlin	1	0	*	
Jacob	0	0	*	
Jackson	0	1	*	

Table IV: final result

Wine	cream	smoking	Cancer
1	1	1	
1	1	*	
1	0	*	
0	1	1	
0	1	*	4
0	0	1	
1	0	*	
0	0	*	
0	1	*	

The Table III displays the result of dot-matrix operation and Table IV displays the final data set for publishing. It is clear that the first five rows in the table is from the original data set and rest are added by our dot matrix procedure. And the attribute smoking is identified as the main private item so that in our dot matrix procedure the attribute values are modified with '*' value.

IV. RESULT AND DISCUSSION

The table 3 shows the result produced by our methodology, which maintains the originality of the data set and also hides the privacy information. So that the third party can use the published dataset, still they cannot identify the personal information. The proposed randomized attribute selection process reduces the probability of guessing the attribute and the person who have interested pattern like from the sanitized data. This increases the factor of privacy preservation and the efficiency of the system. This can be further optimized using multiple attribute selection based on randomization technique which can further increase the efficiency of the proposed methodology.

REFERENCES

[1] R. Agrawal and R. Srikant, "Privacy-Preserving Data Mining," Proc. ACM SIGMOD '00, pp. 439-450, 2000.

[2] C.C. Aggarwal and P.S. Yu, "On Variable Constraints in Privacy Preserving Data Mining," Proc. SIAM Int'l Conf. Data Mining(SDM), 2005.

[3] C. Clifton and D. Marks, "Security and Privacy Implications of Data Mining," Proc. 1996 ACM SIGMOD Int'l Workshop Data Mining and Knowledge Discovery, pp. 15-19, Feb. 1996.

[4] C. Clifton, M. Kantarcioglu, J. Vaidya, X. Lin, and M.Y. Zhu, "Tools for Privacy Preserving Distributed Data Mining," ACM SIGKDD Exploration Newsletter, vol. 4, no. 2, pp. 28-34, 2002.

[5] Y. Saygin, V.S. Verykios, and C. Clifton, "Using Unknowns to Prevent Discovery of Association Rules," ACM SIGMOD Record, vol. 30, no. 4, pp. 45-54, 2001.

[6] C.C. Aggarwal and P.S. Yu, Privacy Preserving Data Mining: Models and Algorithms (Advances in Database Systems). Springer-Verlag,2008.

[7] V.S. Verykios, A.K. Emagarmid, 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.

[8] M. Atallah, E. Bertino, A. Elmagarmid, M. Ibrahim, and V.S. Verykios, "Disclosure Limitation of Sensitive Rules," Proc. IEEE Knowledge and Data Eng. Exchange Workshop (KDEX '99), pp. 45-52, 1999.

[9] B.-C. Chen, K. LeFevre, and R. Ramakrishnan, "Privacy Skyline:Privacy with Multidimensional Adversarial Knowledge," Proc.Int'l Conf. Very Large Data Bases (VLDB), pp. 770-781, 2007.

[11] Z. Huang, W. Du, and B. Chen, "Deriving Private Information from Randomized Data," Proc. ACM SIGMOD, pp. 37-48, 2005.

[12] K. LeFevre, D.J. DeWitt, and R. Ramakrishnan, "Workload-Aware Anonymization," Proc. Int'l Conf. Knowledge Discovery and Data Mining (KDD), pp. 277-286, 2006.

[13] P. Samarati, "Protecting Respondents' Identities in Microdata Release," IEEE Trans. Knowledge and Data Eng., vol. 13, no. 6,pp. 1010-1027, Nov./Dec. 2001.

[14] 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.

[15] 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.

[16] 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.

[17] M. Atzori, F. Bonchi, F. Giannotti, and D. Pedreschi, "Anonymity Preserving Pattern Discovery," VLDB J., vol. 17, pp. 703-727, 2008.

[18] 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.

[19] C.C. Aggarwal and P.S. Yu, "On Privacy-Preservation of Text and Sparse Binary Data with Sketches," Proc. SIAM Conf. Data Mining, 2007.

[20] M. Terrovitis, N. Mamoulis, and P. Kalnis, "Privacy-Preserving Anonymization of Set-Valued Data," Proc. Int'l Conf. Very Large Data Bases (VLDB), 2008.

[21] Y. Xu, K. Wang, A.W.-C. Fu, and P.S. Yu, "Anonymizing Transaction Databases for Publication," Proc. SIGKDD, pp. 767-775, 2008.

[22] D. Richards, "Data Compression and Gray-Code Sorting,"Information Processing Letters, vol. 22, pp. 201-205, 1986.

[23] A. Pinar, T. Tao, and H. Ferhatosmanoglu, "Compressing Bitmap Indices by Data Reorganization," Proc. IEEE Int'l Conf. Data Eng.(ICDE), pp. 310-321, 2005.

[24] D. Kifer and J. Gehrke, "Injecting Utility into Anonymized Datasets," Proc. ACM SIGMOD, pp. 217-228, 2006.

[25] A. Andoni, M. Datar, N. Immorlica, P. Indyk, and V. Mirrokni, Nearest Neighbor Methods in Learning and Vision: Theory and Practice. MIT Press, 2006.

[26] J.K. Reid and J.A. Scott, "Reducing the Total Bandwidth of a Sparse Unsymmetric Matrix," SIAM J. Matrix Analysis and Applications, vol. 28, no. 3, pp. 805-821, 2006.

[27] C. Papadimitriou, "The NP-Completeness of the Bandwidth Minimization Problem," Computing, vol. 16, pp. 263-270, 1976.

[28] Z. Zheng, R. Kohavi, and L. Mason, "Real World Performance of Association Rule Algorithms," Proc. Int'l Conf. Knowledge Discovery and Data Mining (KDD), pp. 401-406, 2001.

[29] A. Narayanan and V. Shmatikov, "How to Break Anonymity of the Netflix Prize Dataset," http://arxiv.org/abs/cs/0610105,2010.

[30] A pattern based framework for privacy preservation through association rule mining (2012).