International Journal of Security and Its Applications
Vol.7, No.6 (2013), pp.109-116
http://dx.doi.org/10.14257/ijsia.2013.7.6.11

A Method for Individualized Privacy Preservation

Song Yang, Li Lijie, Zhang Jianpei and Yang Jing

College of Computer Science and Technology, Harbin Engineering University, Harbin 150001, China
songyang@hrbeu.edu.cn

Abstract

This paper is based on the (α, k)-anonymity models by introducing the individualized privacy-sensitive factor then computing the individualized privacy preservation demand degree of each sensitive value to realize the individualized service of each sensitive value. It combining with the top-down local codes and the sensitive attribute generalization anonymous technology proposes a method based on the attribute classification tree to realize the individualized (α, k)- anonymity models. At last, the results of simulation show that this method could complete the anonymilization and meet the demand of individualized privacy preservation, and the information loss is lower than the original one.

Keywords: *Privacy-preservation, Individualized, Data anonymity, Information loss*

1. Introduction

With the fast development of the internet technology, the data collections and analyzing operations are becoming easier and easier. The data after being excavated which are to be released is still containing a lot of personal private data during the releasing process, and the attackers can get the private data of the respondents through other ways.

Anonymous technology is an effective method of private preservation. The concept of anonymous is first proposed by Samarati and Sweeney [1]. As deepened research, researchers get a lot of results of this area. At the same time there are a lot of summarize about methods of anonymous privacy-preservation [2-4]. Even though there are many ways, most of them didn't concern about the autonomy of individual privacy and the protecting demand of different interviewees to different sensitive value.

2. Related Work

The individual privacy-preservation mainly emphasizes the individualization compared with the traditional methods which just concern about the different conditions of different sensitive values. The individualized constraint is to provide protection of different granularity with different privacy levels so to meet the individual demand.

The conception of Individualized anonymity is first put forward by Xiao [5]. Its main idea is that the degree of anonymous granulation specifying for individuals. It provides individual with different sensitive attribute constraint to realize the individualized anonymity. The authors of literature [6] propose a clustering-based individualized model which is based on the model described above. In literature [7], it completes the (k,l)-anonymity model through setting the anonymous degree of each record and the sensitive attribute variety degree. The model uses k and l for the constraint condition of generalization. Ye use generalization of sensitive attributes to implement the (α,k)-anonymity model to protect the individualization

ISSN: 1738-9976 IJSIA

of sensitive values [8]. In literature [9], authors combining granulation, Rough Set Theory and k-anonymity principle propose a k-anonymity based individualized preservation method.

This paper proposes the individualized (α,k) -anonymity model based on the traditional (α,k)-anonymity model, and the individualized non-correlating constraint. The constraint is set by individuals according to the different sensitive degree which is reflected by different sensitive values. It defines an individualized (α,k)-anonymity model.

3. Definitions

Definition 3.1 k-anonymity model: D is a non-empty data set, Π_{AS} is the equivalence partition of the attribute set AS on the data set D, $\Pi_{AS}=\{\varphi_1,\varphi_2,\ldots,\varphi_k\}$. If $\forall x\in\varphi_i$, $\exists\Delta=\{y\mid y\in\varphi_i\wedge y\neq x\}$, and $|\Delta|\geq k-1$, then the equivalence partition Π_{AS} is k-anonymity, the data set D with k-anonymity model, where, $1\leq i\leq k$, $|\Delta|$ is the number of tuple in data Δ.

Definition 3.2 α -non-correlating: its quasi-identifier is QI, the sensitive attribute is $SA(SA\notin QI)$, the value of SA is s. $|(E,s)|$ is the occurrence number of s in equivalence class of anonymous table T', $|E|$ is the number of records in equivalence class E. If $\forall E$ satisfied with condition $(E,s)/|E|\leq\alpha\quad(0<\alpha<1)$, where α is the threshold set by data publishers, then table T' about quasi-identifier QI and sensitive value s is α -non-correlating.

Definition 3.3 (α,k) -anonymity: the original data table is T, the table after anonymilization is T', if T' both satisfied with k -anonymity model and the sensitive value s which is α -non-correlating, then the anonymous table T' is the (α,k)-anonymity of quasi-identifier and sensitive value s, also called "special (α,k)-anonymity". If the equivalence class of T' is α -non-correlating to each sensitive attribute value and also k -anonymous, so the anonymous table T' is (α,k)-anonymous on quasi-identifier and sensitive attribute set $SA(SA\notin QI)$, also called "normal (α,k)-anonymity".

A record is called a tuple. It includes a lot of attribute such as quasi-identifier and sensitive attribute and so on. Then the generalization information loss of the tuple is the sum of each attribute generalization information loss. Its definition below:

Definition 3.4 Tuple generalization information loss: defined tuple t and the tuple after generalization t', $t=\{A_1,A_2,\ldots,A_m\}$, $t'=\{A_1',A_2',\ldots A_m'\}$, A_i' is the generalization attribute of $A_i, 1\leq i\leq m$, so the generalization information loss of tuple t is defined as:

$$GIL(t,t')=\sum_{i=1}^{m}WDS(A_i,A_i') \tag{1}$$

So the generalization information loss of a whole data table is the sum of tuples' generalization information loss included by the table.

4. Individualized (α, k)-anonymous Model

According to the definition of non-correlation in literature [10] and the method of individualized anonymity in literature [6], we extend the concept from the whole data sets to anonymous equivalence class, the frequency of each sensitive attribute value should equal or greater than the threshold α_{s_i}.

Definition 4.1 α-non-correlation constraint: as to anonymous equivalence class G and sensitive value s, if G satisfies with α-non-correlation, then $freq(G,s)/|G| \le \alpha$, where $freq(G,s)$ is the record number of sensitive value s in equivalence class G, and α is the threshold defined by data owners $(0 \le \alpha \le 1)$, $freq(G,s)/|G|$ is the correlation between each record in equivalence class G and sensitive attribute value s, also the frequency of sensitive s in equivalence class G.

Definition 4.2 Individualized non-correlating constraints: r is the record in data set, the value set of sensitive value s is defined as $F(S)=\{s_1, s_2, \cdots\}$, and sensitive value s is the attribute that record r correlating with the attributes in equivalence class G the interviewee wanted to protect. R_r is the individualized non-correlating constraint, r is a two-tuple represented as (α_{s_i}, s_i). The factor (α_{s_i}, s_i) in R_r is the individualized non-correlating constraint $freq(G(r), s_i)/|G(r)| \le \alpha_{s_i}$ needed by record r and correlating sensitive attribute s_i. In other words, the correlating constraint between single record r and any sensitive value s_i in $G(r)$ must be less than the α_{s_i} defined by data interviewees.

During the process of data collecting, the individualized non-correlating constraint (R_r) permits persons to set the non-correlating constraints of each sensitive value needed to be protected according to their hobbies. There are two extreme classes of α_{s_i} below:

1. $\alpha_{s_i} = 1$, it indicates that the correlating value between record r and sensitive attribute value s_i is less than 1.

2. $\alpha_{s_i} = 0$, it means the interviewee isn't correlating to sensitive attribute value s_i.

Definition 4.3 individualized (α, k)-anonymity model: as to dataset D, D' is the set after anonymilization, if D' satisfies with the (α, k)-anonymity constraints, then D' is satisfies with k-anonymity constraints and records in D' of their own equivalence class meets the individualized non-correlating constraints.

In order to minimize the loss of information, the Top-down greedy algorithm is used to find an effective solution. Firstly, the quasi-identifiers are generalized. All of the records are in the same anonymous group. Secondly, step by step the anonymous group is divided into many smaller ones, until the scale of new anonymous is less than 2k or personalized correlation constraint will not be able to meet.

The pseudo-code description of algorithm is as follows:

Input: Non-anonymous group GD of dataset D

Output: Anonymous equivalence class set SG'

1. $SG = G_D$, assign a anonymous group of dataset D including all records to SG

2. While $(SG \neq \phi)$, select a anonymous group G from SG

3. $Cost_{best} = Cost_G$

4. for i:$1 \to m$

5. $G_{subs} = QI\_Specialize(G,i)$

6. If $\forall g \in G_{subs}$

7. $g' = Sensitive\_Generlize(g)$&& g' satisfy personalized (α, k)-anonymity

8. && $\sum_{G_{subs}} Cost(g') < Cost_{best}$

9. Then $Cost_{best} = \sum_{G_{subs}} Cost(g')$

10. If there exists α better detailed scheme in anonymous group G

11. Then $SG \leftarrow SG \bigcup G_{subs}$

12. Else $SG' \leftarrow SG \bigcup G_{subs}$

13. End for , for loop end

14. End while, while loop end

15. Programmer end and output SG'

We adopt standardizing degree of distortion to estimate loss of information in the process of anonymization. The standardizing degree of distortion based on the attribute hierarchy tree is as follows:

$$DR(v) = height(v, v') / height(taxnomytree) \tag{2}$$

Where v' is the generalized value of v, $height(v, v')$ represents the number of generalization of attribute v to v', or the hierarchy distance of v and v' in the attribute hierarchy tree? The $height(v, v') = 0$ if there is not generalization.

To record $r(q_1, \cdots, q_n, s)$, define its degree of distortion of all attributes as:

$$DR(r) = \sum_{i=1}^{n} DR(q_i, q_i') + p * DR(s, s') \tag{3}$$

Where P is sensitive attribute generalized penalty coefficient? P is defined based on application requirements and semantic of datasets. In the paper, P is simply defined as:

$$p = \sum_{i=1}^{n} height(q_i) / height(s) \tag{4}$$

Intuitive sensitive attribute generalization may lose more information than quasi-identifier attributes generalization [11]. So, we lucubrate sensitive attribute generalization techniques after quasi-identifier k -anonymization to control attributes leak.

5. Experiment and Result Analysis

Simulation is executed according to following three aspects: applicability of anonymous model, information loss of anonymous model and efficacy of corresponding algorithm. The algorithms used in experiment are personalized (α, k)-anonymity algorithm and traditional (α, k)-anonymity algorithm. Experiment adopts IPUMS dataset.

5.1. Applicability Analysis

To applicability analysis, sensitive attribute generalization is feasible. We use 50K records and add lots of different non-relevant constraints to data. Figure 1 shows comparative results. In Fig.1, traditional (α, k)-anonymity algorithm cannot get a proper anonymous dataset when $\alpha < 0.2$. The personalized algorithm can get one when $\alpha < 0.15$. The average distortion measure is greater than 1 when $\alpha = 0.15$. Hence, the personalized algorithm can deal with more rigid non-relevant constraint than traditional (α, k)-anonymity algorithm.

5.2. Information Loss Analysis

Figure 2 shows information loss of non-relevant constraint. What the experiments simulate is three different distributions: unified constraint, normal distribution and uniform distribution when constraint value is 0.25. Simulations execute 100 times. The average information loss shows in Figure 2. From the Figure 2 we can know that the greater the difference, the more information loss after personalizing non-relevant constraint. The strongest personalization constraint in a anonymous group determines the non-relevant constraint of this group. So the greater the constraint difference, the greater the anonymous group need to be generalized.

The result shows that personalized (α, k)-anonymity model support specified anonymous strategy. Local recoding technique is more flexible when dealing with non-uniform distribution non-relevant constraint.

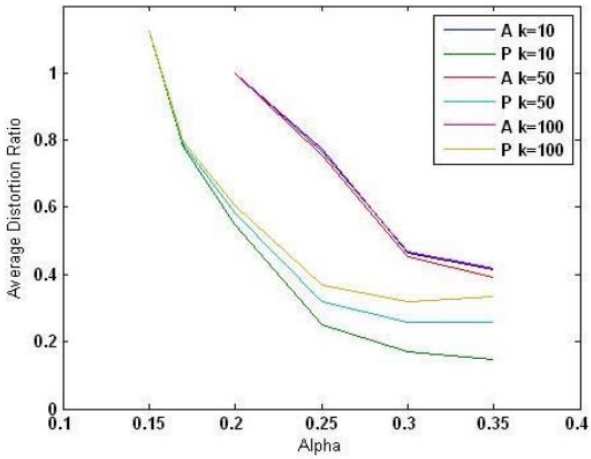

Figure 1. The Scopes of Application of Algorithms Compare and Analyse

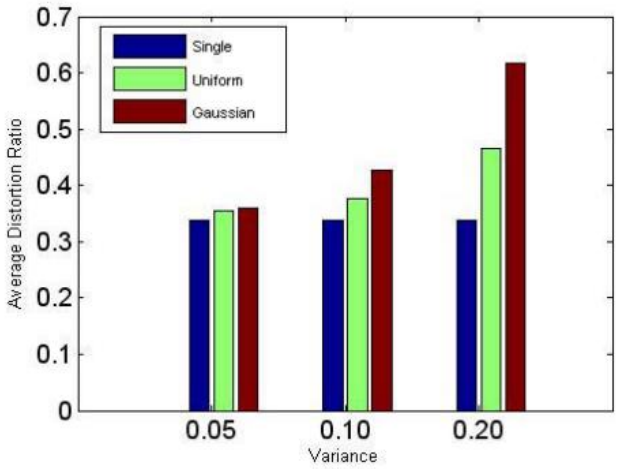

Figure 2. The Non-relevant Constraint Information Loss of Difference Distribution

5.3. Algorithm Efficiency Analysis

Figure 3 shows that the two algorithms' time cost under different dataset's scale. Time cost of personalized algorithm is about 2-5 times than traditional (α, k)-anonymity algorithm. The scope of parameter α is 0.25 to 0.4 and scale of datasets is 5K to 10K. Personalized algorithm finds more possible solution space than traditional (α, k)-anonymity algorithm. In every detailed procedure, the former computes group non-relevant constraint and sensitive attribute generalization.

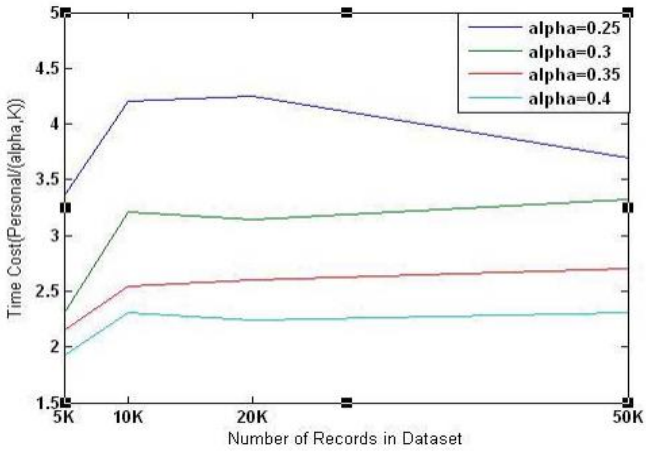

Figure 3. The Non-relevant Constraint Information Loss of Difference Distribution

6. Conclusion

In recent years, privacy protection has been an attractive field. The paper proposed personalized (α, k)-anonymity algorithm which supports ones' specific preference to protect sensitive attribute from leaking. Simulation shows that our personalized (α, k)-anonymity

algorithm is efficient. In contrast with traditional (α,k)-anonymity algorithm, this algorithm can reduce information loss in the anonymous procedure.

Acknowledgements

This paper is supported by Heilongjiang Province Science Foundation for Youths (No. QC2011C012), Harbin Foundation for the Talents of Technology Innovation (No. 2012RFQXG095), Fundamental Research Funds for the Central Universities (No.HEUCF100605).

References

[1] P. Samarati and L. Sweeney, "Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression", Technical report, SRI International, **(1998)**.
[2] L. Sweeney, "K-anonymity: A model for protecting privacy", International Journal on Uncertainty, Fuzziness and Knowledge-Based Systems, vol.5, no. 10, **(2002)**.
[3] K. LeFevre, D. J. DeWitt, R. Ramakrishnan, "Incognito: Efficient full-domain k-anonymity", ACM SIGMOD International Conference on Management of Data, Baltimore, USA, **(2005)**.
[4] K. LeFevre, D. J. DeWitt and R. Ramakrishnan, "Mondrian multidimensional k-anonymity", IEEE International Conference on Data Engineering, **(2006)**, Atlanta, USA.
[5] L. Willenborg and T. De Waal, "Elements of statistical disclosure control", Springer Verlag, vol. 155, **(2001)**.
[6] X. Xiao and Y. Tao, "Personalized privacy preservation", Proceedings of the 2006 ACM SIGMOD international conference on Management of data, Chicago, USA, **(2006)**.
[7] L. Ninghui, L. Tiancheng and S. Venkatasubramanian, "Closeness: A New Privacy Measure for Data Publishing", IEEE Transactions on Knowledge and Data Engineering, vol. 7, no. 22, **(2010)**.
[8] B. Zhou, J. Pei and W. S. Luk, "A brief survey on anonymization techniques for privacy preserving publishing of social network data", ACM SIGKDD Explorations Newsletter, vol. 2, no. 10, **(2008)**.
[9] D. A. Gkoulalas and V. S. Verykios, "A free terrain model for trajectory k-anonymity", Proceedings of the 19th International Conference on Database and Expert Systems for Applications, Berlin, Germany, **(2008)**.
[10] R. C. W. Wong, J. Li, A. W. C. Fu and K. Wang, "(alpha, k)-anonymity: an enhanced k-anonymity model for privacy preserving data publishing", Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, **(2006)**, Philadelphia, USA.
[11] A. Machanavajjhala, J. Gehrke, D. Kifer and M. Venkita, "Subramaniam.l-diversity:Privacy beyond k-anonymity", ACM Transactions on Knowledge Discovery from Data, vol. 1, no. 1, **(2007)**.

