Research Journal of Applied Sciences, Engineering and Technology 4(22): 4923-4927, 2012
ISSN: 2040-7467

Submitted: June 13, 2012 Accepted: July 18, 2012 Published: November 15, 2012

A Novel Anonymity Algorithm for Privacy Preserving in Publishing Multiple Sensitive Attributes

Jian Wang
College of Computer and Information Engineering, Henan University of Economics and Law, Zhengzhou, China

Abstract: Publishing the data with multiple sensitive attributes brings us greater challenge than publishing the data with single sensitive attribute in the area of privacy preserving. In this study, we propose a novel privacy preserving model based on *k*-anonymity called (α, β, *k*)-anonymity for databases. (α, β, *k*)-anonymity can be used to protect data with multiple sensitive attributes in data publishing. Then, we set a hierarchy sensitive attribute rule to achieve (α, β, *k*)-anonymity model and develop the corresponding algorithm to anonymize the micro data by using generalization and hierarchy. We also design experiments to show the application and performance of the proposed algorithm.

Keywords: Data publishing, *k*-anonymity, multiple sensitive attributes, privacy preserving

INTRODUCTION

Privacy preserving in data publishing has received considerable attention from the database security researchers and become a serious concern in publishing of personal data with the rapid development of computer technology and internet technology in the past few years. Nowadays a lot of work had been done in the research field of privacy preserving data publishing (Jian *et al*., 2009). Sweeney presents *k*-anonymity model (Sweeney *et al*., 2002) in order to protect privacy information from such linking attack. To avoid background knowledge attack and homogeneity attack, ℓ-Diversity (Machanavajjhala *et al*., 2006) algorithm and *p*-sensitive *k*-anonymity model (Truta *et al.*, 2006) are proposed.

In some applications of privacy preserving data publishing, a practical demand is to publish a data set on multiple quasi-identifiers for multiple users simultaneously, which poses several challenges. Nowadays there is little work focusing on this. So my paper proposes a novel anonymous (α, β, *k*)-anonymity model oriented multiple sensitive attributes privacy preserving. In order to avoid the disclosure of sensitive information in the publication datasets with multiple sensitive attributes, this paper carries out a detailed and deep analysis to homogeneous attack and background knowledge attack, builds proposed model by utilizing the rule of classified sensitive attributes and ensures the diversity among the multiple sensitive attributes values. Then we propose (α, β, *k*)-anonymity algorithm which implementing the proposed model by utilizing top-down multidimensional division method and single dimension sequel sets division method. At last a series of experiments are designed to compare the proposed algorithm with other algorithms. The results of experiments show the superiority in the aspects of information loss degree and privacy preserving degree for the proposed algorithm.

METHODOLOGY

Implement:

Definition 1: Multi-Attribute ℓ-Diversity
Let T be a table with nonsensitive attributes $Q_1, \ldots, Q_{m1}$ and sensitive attributes $S_1, \ldots, S_{m2}$. We say that T is ℓ-diverse if for all $i = 1 \ldots m_2$, the table T is ℓ-diverse when S_i is treated as the sole sensitive attribute and $\{Q_1, \ldots, Q_{m1}, S_1, \ldots, S_{i-1}, S_{i+1}, \ldots, S_{m2}\}$ is treated as the quasi-identifier (Machanavajjhala *et al.*, 2006).

Definition 2: Quasi-identifier
A set of nonsensitive attributes $\{Q_1, \ldots, Q_i\}$ of a table is called a quasi-identifier if these attributes can be linked with external data to uniquely identify at least one individual in the general population (Sweeney *et al.*, 2002).

Definition 3: *k*-anonymity
Let RT (A_1, ..., A_n) be a table and QI_{RT} be the quasi-identifier associated with it. RT is said to satisfy *k*-anonymity if and only if each sequence of values in RT [QI_{RT}] appears with at least *k* occurrences in RT [QI_{RT}] (Sweeney *et al.*, 2002).

Definition 4: Local recoding
Given a data set D of tuples, a function F that convert each tuple *t* in D to F (*t*) is a local recoding for D (Wong *et al.*, 2006).

In global recoding, all values of an attribute come from the same domain level in the hierarchy. For example, all values in Birth date are in years, or all are in both months and years. One advantage is that an anonymous view has uniform domains but it may lose more information and suffer from over-generalization (LeFevre *et al.*, 2005). With local recoding, values may be generalized to different levels in the domain, so local recoding is more flexible than global recoding.

Definition 5: (α, β, *k*)-anonymity
Let T (A_1, ..., A_n) is a table with nonsensitive attributes Q_1, …, Q_m and sensitive attributes S_1,…, S_i. The number of tuples is QI_n in QI-group. The number of distinct values of sensitive attribute S_i is nS_i, the corresponding number of distinct sensitive attribute values is nS'_i in S_i of all the same sensitive attribute values in S_{i-1}. T is said to satisfy (α, β, *k*)-anonymity if and only if T satisfies *k*-anonymity and the number of distinct values for each sensitive attribute occur at least β times ($2 \le \beta \le k$) within the same QI-group and $\alpha = nS_i - nS'_1 \neq 1$ in each QI-group of tuples.

To illustrate this anonymity approach, we analyze the data from Table V that satisfies 4-anonymity with respect to Sex, Age and Zip code and includes two QI-groups. The first group has three different diseases and three different salaries, the second group also has three different diseases and three different salaries, therefore β = 3. In the first group, $nS_1 = nS_2 = 3$, $nS'_2 = 2$ because the corresponding distinct Disease attribute values are Headache and Paranoia of the same Salary attribute values {6000, 6000} in the Salary attributes. Thus, $\alpha = nS_2 - nS'_2 = 3-2 = 1$, it is not satisfies (α, β, *k*)-anonymity. From the analyses above, we know that Table V will lead to a leakage of privacy information, that is, if α = 1, it will cause a leakage if an attacker has background knowledge.

We adopt hierarchy tree technology to solve above problem. A publicly known hierarchy on the sensitive attributes (Salary, Disease) can be set.

Definition 6: Hierarchy sensitive attribute rule
For the last same sensitive attribute values in S_i, the parent node of the sensitive attribute values will instead of at least one half of these values if α = 1.

For example, we will use its parent node Affective disorder instead of Depression for the last same Disease attribute values Depression in the first QI-group. According to the Hierarchy sensitive attribute rule, at least one Depression is replaced by Affective disorder, the result will become {Depression, Affective disorder}. Then $\alpha = 2 \neq 1$. Likewise, at least one Catatonia is replaced by Schizotypal disorder in the second QI-group, the result will become {Catatonia, Schizotypal disorder}. The new anonymity table is satisfies (α, β, *k*)-anonymity, it can efficiently prevent background knowledge attack.

Algorithm:

The algorithm is used in the process of changing a microdata into a masked microdata that satisfies (α, β, *k*)-anonymity, achieving secure publishing data that contains multiple sensitive attributes and resisting the background knowledge attack, which is illustrated as follows.

Input: The microdata T, α, β, *k*, *i*, $2 \le \beta \le k$ and hierarchies on categorical attributes
Output: The publishable relation T*
Procedure:

- Anonymizing microdata T to T′ using *k*-anonymity
- If T′ has (α, β, *k*)-anonymity property then for each sensitive attribute in the same QI-group
- Let nS_i is the number of distinct values of sensitive attribute S_i
- Let nS'_i is the corresponding number of distinct sensitive attribute values in S_i of all the same sensitive attribute values in S_{i-1}
- If ($nS_i - nS'_i ! = 1$)
  {T′ satisfies (α, β, *k*)-anonymity
  Return T^*}
- Else {
  T′ doesn't satisfy (α, β, *k*)-anonymity
  at least one half of the last same sensitive attribute values in S_i are replaced by its parent nodes
  return T^*
  }

EXPERIMENTAL EVALUATION

In this section we design experiments to show the application and performance of our proposed (α, β, *k*)-anonymity algorithm. In the following experiments we use IBM Quest Synthetic Data Generator to generate the experimental data. The responsible computer for the implementation is equipped with Intel Core i3 2.13 GHZ, 4 GB memory and 500 GB hard disk. Throughout the experiments, we use Java programming language to complete the work.

The changes of number for multiple sensitive attributes which are discussed in the experiments are described in Table 1. We can adjust the data set used in the experiments according to variation rule of number for multiple sensitive attributes in Table 1.

Assume C_{DM} means the punishment degrees for multiple sensitive attributes table and each tuple of the table has a punishment degree. C_{DM} is calculated by:

$$C_{DM} = \sum_{\forall Es.t.|E|>k} |E|^2 = \sum_{j=1}^{i} |QI_j|^2 \quad (1)$$

where, |E| means the granularity of equivalent group QI_j.

Calculation formula for average granularity of equivalent group is below:

Table 1: Multiple sensitive attributes used in the experiment

Number of sensitive attributes	Sensitive attributes
i = 1	<tumor topography>
i = 2	<tumor topography, histology>
i = 3	<tumor topography, histology, survival years>
i = 4	<tumor topography, histology, survival years, diagnose date,>
i = 5	<tumor topography, histology, survival years, diagnose date, age at diagnosis>
i = 6	<tumor topography, histology, survival years, diagnose date, age at diagnosis, turnor size>

$$C_{AVG} = \frac{count(t_i)}{k \times count(|QI_j|)} \quad (2)$$

where, count (t_i) means the number of all tuples and count ($|QI_j|$) indicates the number of all equivalent group.

Assume count (t_i) = r, minimal discernability of multiple sensitive attributes set which contains *r* tuples is calculated by:

$$\min C_{DM} \geq \min\left(\sum_{|E|\geq k} |E|^2\right) \geq \sum_{i=1}^{n} \min(|E_i|) \geq \sum_{i=1}^{n} = count(t_i) \times k \quad (3)$$

Calculation formula for information loss is below:

$$loss(v^*) = \frac{Vnum(v^*) - 1}{Vnum.Dom(A)} \quad (4)$$

where, $v \in Dom(A)$, loss(v*) indicating the information loss caused by the generalization form v to v*. Vnum(v*) means the number of siblings to v*. The number of all the values for attribute A is named Vnum.Dom (A).

Definition 7: Multiple sensitive attributes discernability. We call all the information loss in the process of anonymizing data table *T* which contains multiple sensitive attributes into T* as discernability of multiple sensitive attributes. Multiple sensitive attributes discernability incorporates identifier attribute information loss named QI.DisR(T*) and sensitive attribute information loss named SA. DisR(T*):

$$T.DisR(T^*) = QI.DisR(T^*) + SA.DisR(T^*) \quad (5)$$

$$QI.DisR(T^*) = \frac{\sum_{\forall t^* in T^*} Loss_{tuple}(t_{qi}^*)}{\sum_{\forall t^* in T^*} \sum_{i=1}^{e+f} 1} \quad (6)$$

$$QI.DisR(T^*) = \frac{\sum_{\forall t^* in T^*} Loss_{tuple}(t_{qi}^*)}{\sum_{\forall t^* in T^*} \sum_{i=1}^{e+f} 1} \quad (7)$$

In the aspect of information loss, we compare (α, β, *k*)-anonymity algorithm with Exponential-L algorithm

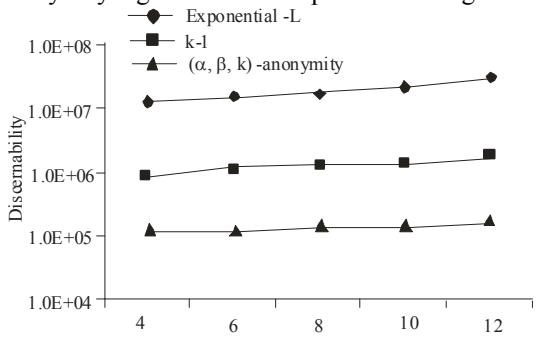

Fig.1: Discernability changes depending on *l*-value changes (k = 40, i = 3)

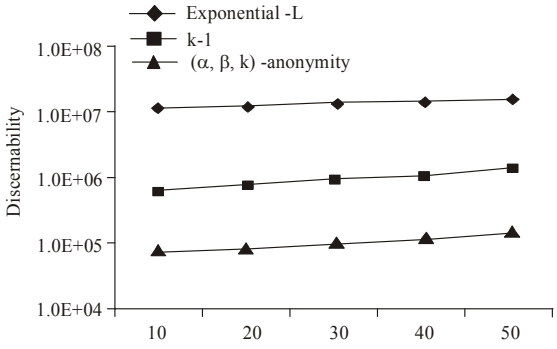

Fig. 2: Discernability changes depending on *k*-value changes (l = 5, i = 3)

and k-l algorithm. Firstly test the discernability. The experiment has three important parameters, namely *k*, *l*, *i*. In each experiment we fix two parameters and make the third parameter changeable. Figure 1 shows that when k = 40 and i = 3, using the top three sensitive properties, discernability changes depending on *l*-value changes. As is shown from the figure, the discernability of these three algorithms increase as the *l*-value increasing. But the discernability of (α, β, *k*)-anonymity algorithm is much lower than that of the other two algorithms. The reason is that anonymous rules make more equivalent groups cannot be further decomposed, resulting in the increase of *l*-value cause little influence to equivalent group.

Figure 2 shows that when l = 5 and i = 3, discernability changes depending on *k*-value changes. As is shown from the figure, the discernability of these three algorithms increase as the *k*-value increasing. And (α, β, *k*)-anonymity algorithm has a significantly lower

information loss than Exponential-L algorithm and k-l algorithm. The reason is that (α, β, *k*)-anonymity

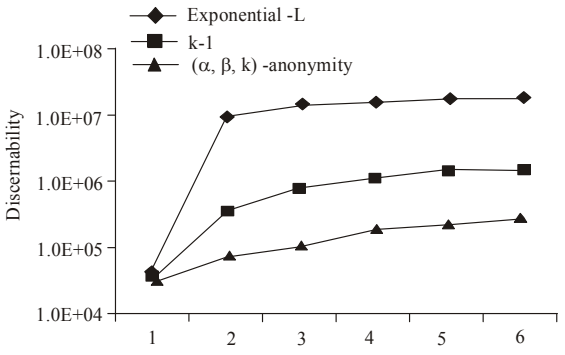

Fig. 3: Discernability changes depending on *i*-value changes (k = 20, l = 8)

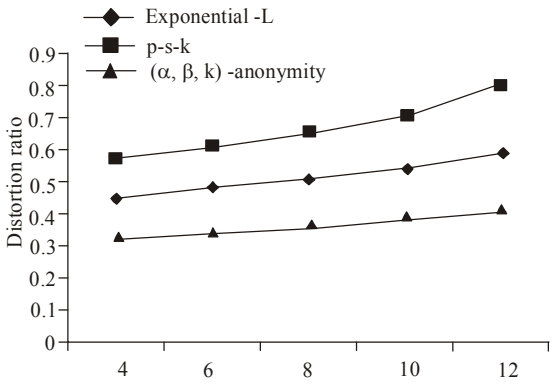

Fig. 4: Distortion ratio changes depending on *l*-value changes (k = 40, i = 3)

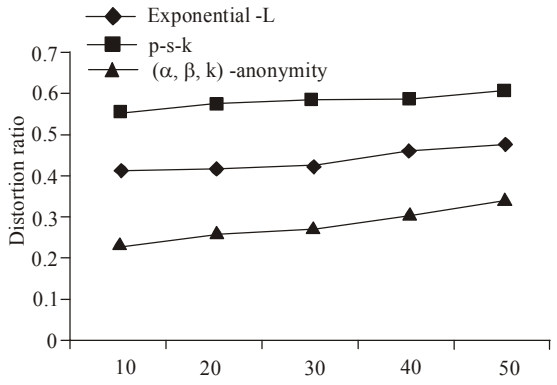

Fig. 5: Distortion ratio changes depending on *k*-value changes (l = 5, i = 3)

algorithm adopts different generalization method for numerical attributes and categorized attributes in order to enhance the flexibility of data generalization.

Figure 3 shows that when $k = 20$ and $l = 8$, discernability changes depending on *i*-value changes, where *i* means the number of multiple sensitive attributes. (α, β, *k*)-anonymity algorithm has a significantly lower discernability than Exponential-L algorithm and k-l algorithm. At the same time, when $i = 1$, the discernability of the anonymous tables generated

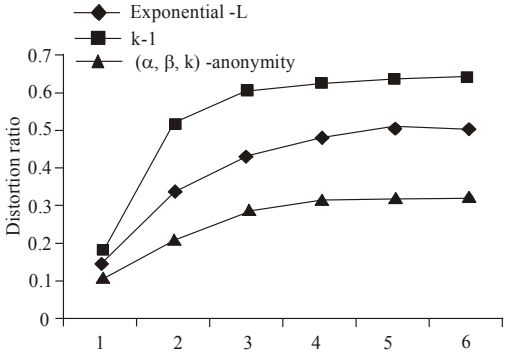

Fig. 6: Distortion ratio changes depending on *i*-value changes ($k = 20$, $l = 8$)

by these three algorithms are similar. Discernability value increase as the number of sensitive properties increases. The information loss degree of data set which contains multiple sensitive attributes is larger than that of data set which contains single sensitive attribute. The reason is that as the number of sensitive properties increasing, the possibility that any two tuples have the same values in sensitive attributes increases. Therefore, there will be more equivalent groups which contain more tuples need to achieve *l*-diversity, thus leading to the degree of generalization intensified. Besides, with the *i*-value, number of sensitive attributes, increasing, generalization algorithm will generate more constraints for multiple sensitive attributes, which is the reason why discernability value increases.

Figure 4, 5 and 6 respectively illustrate that distortion ratio changes depending on the variation of *l*-value, *k*-value and *i*-value. When *l*-value increases, privacy protection degree increases. The number of different sensitive attributes within the same equivalent group increases, resulting in the generalization level of identifier attributes increased, then leading to substantial information loss from the original data increased, ultimately causing distortion ratio of multiple sensitive attributes set increased. When other conditions are identical, *k*-value increased, the group size becomes larger. Because the grouped data should meet diverse, attributes in equivalent group increase. Then generalization levels increase and data distortion rate also increases. When the number of data sets and the parameters are equal, the dimension of sensitive attribute *i* increases, the distortion rate of multiple sensitive attributes is higher, which is caused by sensitive property diversity in each dimension. By contrasting the following three figures, we can conclude that (α, β, *k*) algorithm by using multiple sensitive attributes generalization can avoid excessive generalization of the identifier attributes, so the total distortion rate is less than the other two methods.

Therefore the total distortion ratio of (α, β, *k*)-anonymity algorithm is smaller than other two algorithms and compared to the *l*-diversity, *k*-*l*-sensitive rules has a relatively low level of data loss. This experiment suggests that (α, β, *k*)-anonymity algorithm has lower information loss than *l*-diversity algorithm and *k*-*l* algorithm. According to the above experimental results analysis, we can conclude that (α, β, *k*)-anonymity algorithm can produce better generalization outcome than other algorithms under the same conditions.

CONCLUSION

In this study, our main contributions are that we propose (α, β, *k*)-anonymity model and related algorithm in order to solve the problem of privacy information leakage in publishing the data with multiple sensitive attributes. We adopt hierarchy tree technology to make sensitive attributes with diversity. On the one hand, (α, β, *k*)-anonymity model reduces the amount of information loss. Besides, it can make anonymous data effectively resist background knowledge attack. In the future, we will extend our ideas for handling how to solve privacy information leakage problem caused by re-publication the data with multiple sensitive attributes.

ACKNOWLEDGMENT

This research is supported by the National Natural Science Foundation of China under the grant No. 60773100 and by the Education Science and Technology Project of Henan Province under the grant No. 12A520005.

REFERENCES

Jian, W., Y. Luo, S. Jiang and J. Le, 2009. A survey on anonymity-based privacy preserving. International Conference on E-Business and Information System Security, pp: 721-724.

LeFevre, K., D.J. DeWitt and R. Ramakrishnan, 2005. Incognito: Efficient full-domain k-anonymity. Proceeding of ACM SIGMOD Conference on Management of Data, pp: 49-60.

Machanavajjhala, A., J. Gehrke and D. Kifer, M. Venkitasubramaniam, 2006. ℓ-diversity: Privacy beyond k-anonymity. Proceeding of ICDE.

Sweeney, L., 2002. K-anonymity: A model for protecting privacy. Int. J. Uncertainty, Fuzziness Knowl-Based Syst., 10(5): 557-570.

Truta, T.M. and B. Vinay, 2006. Privacy protection: P-sensitive k-anonymity property. Proceeding of the 22nd IEEE International Conference on Data Engineering Workshops, pp: 94.

Wong, R.C., J. Li, A.W. Fu and K. Wang, 2006. (A, k)-Anonymity: An enhaned k-anonymity model for privacy-preserving data publishing. Proceeding of the 12th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp: 754-759.