Data publishing Anonymity Algorithm Research Based on Clustering

Yu Yang[1, a], Longjun Zhang[1]

[1]Department of Communication Engineering Engineering College of Armed Police Force, Xi'an, China

[a]miaoyude@163.com

Keywords: (*l*, *c*) anonymous algorithm; Data publishing; Clustering; Personal privacy

Abstract. Data publishing provides convenience for data exchange and data sharing. But at the same time, the issue of personal privacy information leakage has become increasingly prominent. Anonymous algorithm is one of the main technologies in data publishing environment to realize privacy protection, but most anonymity algorithm of all sensitive attributes values are treated equally, without considering their sensitivity and specific distribution. It is vulnerable to similar attacks and deviation of attack. The equivalence classes are established by clustering technique, and the different levels of privacy protection are defined for each sensitive attribute value. Using local heavy coding scheme on the identifier to anonymous, anonymity algorithm (*l*, *c*) based on clustering is put forward. Experimental results show that the proposed algorithm improves the availability of published data while protecting privacy.

Introduction

With the development of Internet technology and data processing technology, a large number of individual related data are widely collected and published. However, these data may contain the individual's privacy information, so privacy protection becomes the key issue to be solved in the field of data publishing. Anonymity is one of the key technologies to realize privacy protection in data publishing application at present. Since Sweeney L[1-2] proposed k-anonymous model, privacy protection technology has become a hot topic in the field of domestic and foreign experts.

The study of literature [3] shows that the problem of optimal data anonymity is a NP problem. Around how to improve the efficiency of the implementation of the process of anonymity and reduce the loss of privacy protection of anonymity, there are a variety of heuristic data anonymous method is proposed. Due to the lack of k-model, it is easy to be subjected to homogeneous attack and background knowledge attack , which result in privacy leakage. The *l*- diversity model was improved in literature [6]. Full domain generalization strategy leads to higher information loss. Therefore, a sensitive attributes based on clustering personalized (*l*, *c*)-anonymity model is proposed. By clustering techniques for generating equivalence classes and the local recoding scheme implementation of anonymous treatment in order to reduce the information loss is presented in this paper. Theoretical analysis and experimental results show that the proposed method is effective and feasible.

Related Notion

***k*–anonymity.** *k*-anonymity publishes lower accuracy data through generalization and concealment techniques. It makes each record in anonymous table $T(id,q_1,q_2,\ldots,q_m,s)$ at least with other *k*-1 records have the same value as the identifier attribute. For example, Table 2 is a 2- anonymous table for Table 1.

Table 1 Original data table

Name	Race	Sex	Birth	Zip	Disease
Alice	Black	M	1965-3-18	02141	Flu
Bob	Black	M	1965-5-1	02142	Cancer
David	Black	M	1966-6-10	02135	Obesity
Helen	Black	M	1966-7-15	02137	Gastritis
Jane	White	F	1968-3-20	02139	HIV
Paul	White	F	1968-4-1	02138	Cancer

Table 2 2- anonymous table

Race	Sex	Birth	Zip	Disease
Black	M	1965-3-18	0214*	Flu
Black	M	1965-5-1	0214*	Cancer
Black	M	1966-6-10	0213*	Obesity
Black	M	1966-7-15	0213*	Gastritis
White	F	1968-3-20	0213*	HIV
White	F	1968-4-1	0213*	Cancer

***l*- diversity.** Because the *k*-anonymity model is lack of the sensitive attribute value constraints, it is easy to be attacked by homogeneous attacks and background knowledge. *l*-diversity refers to the anonymous data table $T^*(q_1,q_2,\ldots,q_m,s)$ which satisfies the *k*-anonymity at the same time, the same equivalence class contains at least contain l well-represented sensitive attribute value. One of the most simple and direct interpretation is that the same equivalence class contains at least l different sensitive value, so that the attacker cannot be more than $1/l$ of the probability of reasoning out the sensitive value of the individual. However, the *l*-diversity anonymity model can produce large information loss in some data sets, and it is not enough to resist the similarity attacks and skew attacks.

(*l*, c)- anonymity algorithm based on Clustering

Personalized (*l* ,c)- anonymous. Definition 1 Anonymous data table $T^*(q_1,q_2,\ldots,q_m,s)$ is (l,c) - anonymous, if $T^*(q_1,q_2,\ldots,q_m,s)$ satisfies the *l*-diversity condition, and the ratio of sensitive attribute values in the same equivalence class does not exceed the threshold value of $c(0 < c < 1)$.

(l,c) - anonymously by limiting the equivalence class of sensitive attribute values in the highest ratio to prevent against sensitive attribute values of skewness attack, but there are still similarities between the risk of attack. Therefore, the sensitivity of each sensitive attribute value is defined, as shown in Table 3, the value of the higher sensitivity is distributed in the different equivalence classes, thus preventing the similarity attack to the value of the highly sensitive attribute.

Table 3 Sensitivity and l-value of sensitive attribute values

Disease	Sensitivity	*l*-value
Flu	0.10	3
Obesity	0.40	4
Gastritis	0.50	4
HIV	0.90	6
Cancer	0.95	7

Information loss metrics. Definition 2 (Information loss) Order $e=\{r_1,r_2,\ldots,r_k\}$ is a cluster, and its quasi identifier contains numeric attributes $N_1,N_2,\ldots,N_m$ and sub type properties $C_1,C_2,\ldots,C_n$, T_{Ci} is the classification tree of the C_i domain of the classification tree, Min_{Ni} and Max_{Ni} are the minimum and maximum values of the numerical attribute Ni in the cluster *e*. $\cup_{Ci}$ expresses different attribute set in

cluster e value type attribute C_i . The cluster e generalization processing information loss $IL(e)$ is defined.

$$IL(e) = |e| \left(\sum_{i=1}^{m} \frac{Max_{N_i} - Min_{N_i}}{|N_i|} + \sum_{j=1}^{n} \frac{W\left(\Lambda\left(\cup_{C_j}\right)\right)}{W\left(T_{C_j}\right)} \right) \tag{1}$$

$|e|$ represents the number of records in the cluster e . $|N_i|$ represents the size of the numerical domain N_i . $\Lambda(\cup_{Cj})$ represents subtree in the classification tree of all values of the least common ancestor. $W(T)$ represents the sum of the inter layer distance of the classification tree T.

Definition 3 (Total information loss) Make E a collection of equivalence classes for anonymous table T*. The total information loss of T* is defined.

$$Total - IL(T^*) = \sum_{e \in E} IL(e) \tag{2}$$

(*l*,*c*)- anonymity algorithm based on Clustering. In this paper, the algorithm can be divided into four steps. (1)The data set according to the sensitive attribute values is divided into a number of hash buckets, each hash bucket stored data recording of a sensitive attribute values, the hash bucket according to the sensitive attribute values of sensitivity descending arrangement. (2)In order to prevent similar attacks, from the high sensitivity of the hash bucket optionally a record as poly clusters in the seed records and by sensitivity from low to high from the corresponding hash bucket each take a record to establish l-diversity cluster, repeat the operation until no meet l-diversity condition records exist. (3)The remaining records to meet the minimum information loss and the sensitive attribute value of the highest frequency constraint c principles into existing clusters, if poly cluster record number reaches a certain conditions when performing cluster split operation, but after splitting the cluster still need to meet the(l ,c) - anonymity conditions. (4) An anonymous data table is generated for the alignment identifier to perform the generalization process on each cluster. Algorithm procedure is as follows.

Input Original data set T, Sensitivity of each sensitive value, diversity parameter l_i and maximum frequency constraint parameter c.

Output Anonymous data table T^* , Each equivalence class contains at least l different sensitive attribute values and satisfies the maximum frequency constraint c.

Experiment Analyses

Experimental operating environment: Intel® CoreTM 2 Duo CPU T5450 1.67 GHz，2.0 GB RAM，Windows XP，Visual C++6.0，MATLAB7.0.

Table 4 Sensitivity and *l*-value setting

No	Occupation	Sensitivity	*l*-Value
1	Armed-Forces	0.95	10
2	Exec-managerial	0.90	10
3	Prof-specialty	0.85	9
4	Protective-serv	0.80	8
5	Priv-house-serv	0.80	8
6	Transport-moving	0.70	8
7	Tech-support	0.65	7
8	Farming-fishing	0.60	7
9	Machine-op-inspct	0.50	6
10	Handlers-cleaners	0.40	5
11	Craft-repair	0.30	4
12	Adm-clerical	0.20	3
13	Other-service	0.10	2
14	Sales	0.10	2

Attack vulnerability. The vulnerability of the two algorithms to the vulnerability of the skewed attacks are compared in Fig.1 and Fig.2. As can be seen from the graph, the algorithm can effectively prevent the skew attack against sensitive attribute values. And *l*-diversity algorithm (assuming l = 7), the maximum frequency constraint *c* the smaller the value of the more recorded by the skewed attacks. The smaller the value of *l* (assuming C = 0.4), the more records are subjected to skewed attacks.

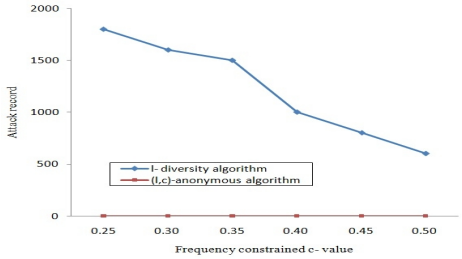

Fig.1 Attack vulnerability (*l*=7)

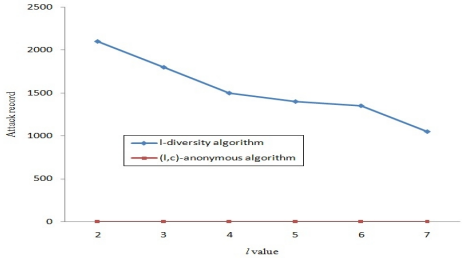

Fig.2 Attack vulnerability (*c*=0.4)

Information loss. Information losses of two algorithms are compared in Fig.3. As can be seen from the graph, the algorithm of this paper has less information loss compared to the *l*- diversity algorithm. The reason is that *l*-diversity algorithm uses a global generalization strategy, the information loss is usually much higher than the local heavy encoding scheme.

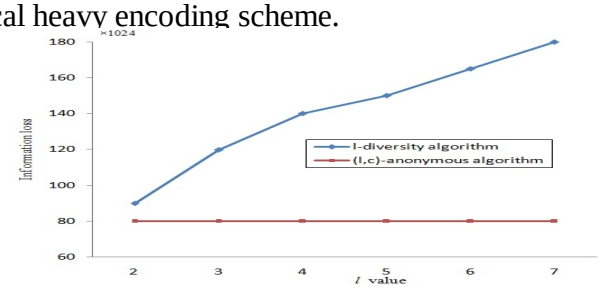

Fig.3 Information loss metrics（*c*=0.4）

Conclusion

In view of the similarity attack and skew attack of *l*-diversity anonymous model, this paper proposes a clustering based personalized (*l*,*c*) - anonymous model. Different sensitive attribute values are defined to define different levels of privacy protection, which endow different *l*- diversity parameter values, and limit the maximum frequency of sensitive attribute values in equivalent classes, which improves the security of the data. The equivalence classes are established by clustering technology, and the local weight of the scheme is used to deal with the identifier by using the method of local weight encoding, which reduces the loss of information. Theoretical analysis and experimental results show that the method is effective and feasible.

Acknowledgments

This work was financially supported by Engineering University of Armed Police Force basic research fund (WJY201307), Department of Communication Engineering background research fund(XJY201405).

Reference

[1]Sweeney L, INTERNATIONAL JOURANL OF UNCERTAINTY,J, Forum Vol 10(2002)557-570.

[2]Sweeney L, INTERNATIONAL JOURANL OF UNCERTAINTY,J, Forum Vol 10(2002)571-588.

[3]Meyerson A，Williams R.PROC OF THE 23RD ACM SIGACT-SIGMODSIGART SYMPOSIUM ON PRINCIPLES OF DATABASE SYSTEMS,J,(2004)223-228.

[4]Wang K，Yu P，Chakraborty S. *Proc of the 4th IEEE Int'l Conf on Data Mining.Washington*(IEEE Computer Society, Washington DC,2004).
[5]Fung B,Wang Ke,Yu P. *Proc of the 21st IEEE Int'l Conf on Data Engineering*(IEEE Computer Society, Washington DC,2005).
[6] Machanavajjhala A，Gehrke J，Kifer D. *Proc of the 22nd Int'l Conf on Data Engineering*(IEEE Computer Society, Los Alamitos,2006).