Research on the Anonymity Method Based-on k-anonymity for Electronic Commerce

Gaizhen YI
School of Information & Engineering
XianYang Normal University
XianYang, China, 712000
E-mail: yigaizhen@163.com

Zhenglong XIE
School of Information & Engineering
XianYang Normal University
XianYang, China, 71200
E-mail: xyncxiezl@126.com

***Abstract*—According to the model of K-anonymity, the attributes of information tables in Electronic Commerce have been divided into three parts: identifier, Quasi Identifier, sensitive information. To conceal privacy information for consumer, the paper adopted a recode method to protect the name and cell-phone number against revealing, used the generalization algorithm to conceal the quasi identifier for consumer. The data resource of paper derived from the Adult database. It has been analyzed on the correlations among of the generalization hierarchy, the ratio of information loss and K. The result provided an alternative scheme of data processing decision for data holder.**

Keywords- Electronic Commerce；K-anonymity; generalization; ratio of information loss

I. INTRODUCTION

With the rapid development of Internet technology, electronic commerce arises at the historic moment and has been got rapid development based on the Internet technology. Electronic commerce gradually replaced the traditional commercial operation mode as a kind of modern business model. Consumers don't want others to know their own consumption, such as buying something, how many did they buy, and how much did they pay and so on, thus the electronic commerce security transaction has gradually become the core of the development of e-commerce and the key problem. Especially with the users of Internet growing and with the technology increasing, information has been obtained by people through the legal or illegal means in various forms, especially the privacy information of others, this brings severe threat for the consumer of an electronic commerce. Furthermore, the commercial organizations collect and analysis the personal information of a consumer for the marketing purposes, and even the information will be released by some merchants. Thus, it has two problems that urgently needed to be solved by the electronic commerce: 1) to protect the personal privacy of a consumer against leaking out, and 2) to need to have enough information for a data analysis. This paper researched on the anonymity implementation of individual information for the end of electronic commerce.

II. CONCEPTS OF K-ANONYMITY

In order to solve the leakage problem on the privacy data when it was released, Samarati P and Sweeney L [1-3] proposed a technology model on K-anonymity, it required to exist the contain number of undistinguishable individual in the released data tables in order to disable an adversary to deduce the specific individual of privacy information, thus to protect the personal privacy against being leaked out.

In *k*-anonymity model, the information that identified an individual property is divided into three categories: 1) the individually identifying attributes, denoted by *ID*, it can unique determine an individual, such as Names, ID number, and Mobile Phone Number etc; 2) the quasi identifier attributes, denoted by *QI*, it is a set of attributes attacked by external data link, such as Sex, Age, and Zipcode; 3) the individual sensitive attributes, denoted by ST, it represents an individual privacy information, for example, what he bought, the numbers, and pay etc. In the *k*-anonymity model, to protect the individual privacy of the released data table against leakage, we don't conceal the privacy information, because this is disadvantage of data analysis, but only delete simply individually identifying attribute as individual can be deduced by linking two tables in order to find a relationship between individual and the value of a sensitive attribute, for example, the individual identifies can be unique determined by a non explicit identifier such as zip code, gender, and birthday, it has about 87% in U.S. according to statistics[4]. Thus, we generalize the non explicit identifier, and represent explicitly individual sensitive attribute, don't process.

Definition 1. Quasi-identifier

Given a set of entity *U*, entity table $T(A_1,...,A_n)$, f_c: $U \rightarrow T$, and f_g: $T \rightarrow U'$, where $U \subseteq U'$. A quasi-identifier of *T*, written *QI* (Quasi Identifier), is a set of attributes $\{A_i,\ldots,A_j\}$, and $\{A_i,\ldots,A_j\} \subseteq \{A_1,\ldots,A_n\}$,where $\exists p_i \in U$ such that $f_g(f_c(p_i)[Q_I]) = p_i$.

That is, Quasi-identifier is a set of attributes identifying individual information by using inference, and exists in privacy tables and outward tables at the same time. The definition of quasi-identifier depends on the information of outward tables achieved by attacker, namely the correlation attributes of an outward table. A privacy table may have different quasi-identifier toward different outward tables.

Definition 2. *k*-anonymity

Given data table T(*A1*,...,*An*), *QI* is the quasi-identifier associated with *T*. *T* is said to satisfy *k*-anonymity if and only if each sequence of values in *T[QI]* appears with at least k

occurrences in *T[QI]*. *T[QI]* represents the projection of tuples of T on the *QI*.

The definition meets the sufficient condition of k-anonymity. If a series of attribute groups of a outward table appear in quasi-identifier associated with *T* (where we assumes that *T* suffices the definition of k-anonymity). Thus the link between *T* table and outward table will not allow attacker deducing any less than K individual tuples.

Corollary 1. Given a data table $T(A_1,...,A_n)$, and a quasi-identifier $QI(A_i,\ldots, A_j)$ associated with *T*, where $A_i,\ldots,A_j \subseteq A_1,\ldots,A_n$. If T satisfies *k*-anonymity, then each sequence of values in $T[Ax]$ appears with at least *k* occurrences, where $x=i,\ldots,j$.

It can be trivially proven that if the released data RT satisfies *k*-anonymity, then the combination of the released data RT and the external sources, cannot link on QI_{PT} or a subset of its attributes to match fewer than *k* individuals. This feature can prevent the link reasoning between the table information and the external sources.

III. ANONYMOUS IMPLEMENTATION ABOUT E-COMMERCE PRIVACY INFORMATION

A. *The Structure of Individual Data Information Collected by E-commerce Sites*

The template is used to format your paper and style the text. All margins, column widths, line spaces, and text fonts are prescribed; please do not alter them. You may note peculiarities. For example, the head margin in this template measures proportionately more than is customary. This measurement and others are deliberate, using specifications that anticipate your paper as one part of the entire proceedings, and not as an independent document. Please do not revise any of the current designations.

When the consumers of the E-commerce go shopping at a certain E-commerce site first of all to register, in course of registration, one need to provide the information as following: Username, Sex, Birthdate, Address, Zipcode, Mobile Phone Number, and so on. When a consumer buys goods, businesses need to provide the information about commodity as following: trade name, the value of trade attribute, quantity, pricing etc.

The individual attribute information of E-commerce transaction consumer can be divided into three categories according to the *k*-anonymity model: 1) individually identifying attributes, they include name, mobile phone. 2) quasi-identifying attributes, they contain Sex, Birthdate, Address and Zipcode. 3) sensitive attributes, they include Commodity Name, Quantity, Pricing, and Pay.

B. *Anonymous Methods*

If we delete simply the attribute of individual identifier according to K-anonymity before the data can be released, this is not suitable for require of data mining and data analyzing, and has not approaches to analyze the relation among commodity, for example, one loves to buy what kind of goods, one also buy which some goods when he buy a certain goods; one often buy which goods, and he occasionally buy which goods. Therefore, we adopt the approach to comprehensive recode in order to neither change individual identifier attribute, and don't leak privacy. That is, we adopt an algorithm to form a new value of code for the attribute of individual identifier. However, we generalized the quasi-identifying attributes, and the sensitive attributes kept unchanged. Thus, we used this idea to achieve the anonymous of individual information.

1)Recoding algorithm for the individually identifying attributes

For the individually identifying attribute, we adopt a comprehensive recode that assigned a same code to two individuals with equaling the value of attribute about name and mobile phone. The algorithm can identify the different individual, and can not reference the relation between a code and an individual. The algorithm can be described as follows:

(1) Input a data table *T*;

(2) Creating an index by using name as the first tags, and mobile phone as the second tags;

(3) variable definition：

string: stored a string that combined name with mobile phone;

ID: this variable represented the different individual, its initial value is 1;

num.: a temp variable.

(4) *string*=Name+Mobile Phone Number

```
ID=1
num. ='t'+str(ID)
skip
while(!EOF)
 do
   if string == Name+Mobile Phone Number
     num.='t'+str(ID)
   else
       ID=ID+1
       num.='t'+str(ID)
   endif
   string=Name+Mobile Phone Number
   skip
 done
```

2) Generalizing the individually quasi-identifier attribute

At present, generalization method was divided into a local recoding and a global recoding. The local recoding does not require a value of attribute to being generalized to the same level. Wong et al. presented a top-down local recoding algorithm [5]. This method first would generalize all tuples in a data table into one equivalence class. Then, tuples are

specialized in iterations. During the specialization, we must maintain *k*-anonymity. The process continues until we cannot specialize the tuple anymore. However, the global recoding required generalizing a quasi-identifier attribute into the same level. The literature [6] presented a global recoding algorithm, called Incognito, it can be applied to the generalization method of maintaining identity. The algorithm was relatively simple, it only selected the quasi-identifier attribute to generalize in iterations until it met the demand of k-anonymity. This paper used a top-down global recoding algorithm.

Due to the purpose of data mining in E-commerce is to sale commodity. The businesses can procure and advertise the commodity according to analyzing the ability of consumer purchase in different areas, different types of consumer purchase. Based on this idea, we first selected Sex attribute to group all tuples; then generalized Zipcode to city, for example, if Zipcode is 712000, it can be generalized 712**; the third step, we generalized Age in terms of Age internal, for example, Age can be divided into Age≤25, 25<Age≤35, 35<Age≤50, and Age>50. Thus, this method formed a generalization tree, it achieved the purpose of anonymous as well as convenient for analyzing the data of E-commerce. This algorithm can be described as following:

input：source data table T；

output：released table PT；

（1）PT is a result table that recoded the individual identifying attribute in table T;

（2）creating a key index by Sex for PT;

（3）creating a second keyword index by Zipcode for PT, and generalizing all tuples in terms of the high four numbers , and thus forming a new generalization code of an attribute;

（4）Creating a third keyword index by Birthdate for PT. This would translate the Birthdate into Age internal, then generalizing the Age internal, and thus forming a generalization code representing internal.

IV. ANONYMOUS EVALUATION METRICS AND ANALYSIS

In E-commerce, the purpose of implementation of consumer anonymous is to disable an attacker to deduce a certain individual from external table achieved by getting external information or by the other way, and find the privacy of consumer purchase. However, the method based on k-anonymity recoded the individually identifying attributes and generalized the quasi-identifier attributes. A generalization level is higher, the anonymity is powerful, but the information loss is increasing. That is, for the recode value of individual identifying attributes, the same records is more, the analysis of consumer information became reducing. Therefore, a generalization has the anonymity as well as producing the information loss. We analyzed that the generalization impacted on the ratio of information loss and K (anonymity).

Definition 3. Ratio of Information loss

Given a table $T(A_1, A_2, ..., A_a)$, and a quasi-identifier |QT|=a. RT is the anonymous table of T, G_{Ai} represents the generalization level of attribute A, where *i=0,1,2，...,h, h* is the highest level number of generalization. Then, when RT suffices *k*-anonymity, the ratio of information loss is:

$$loss(RT) = 1 - \frac{\sum_1^a \sum_1^h \frac{G_{Ai}}{G_{A(i-1)}}}{h * a} \qquad (1)$$

If *T=RT*, then $G_{Ai} = G_{A(i-1)}$, $loss(RT) = 0$, that is said no information loss；if generalization level achieved the highest level h, and can be definite as no validate data members, namely $G_{Ah} = 0$, then $loss\ (RT) = 1$.

We abstracted randomly the records that a value of native-country field is United-States from the adult database of machine learning at United States Owen University. We generalized the field "fnlwgt" into five levels in the case of considering the ratio of loss on single attribute, and thus achieved the relation between generalizing level and the ratio of information loss. It illustrates in Figure 1.

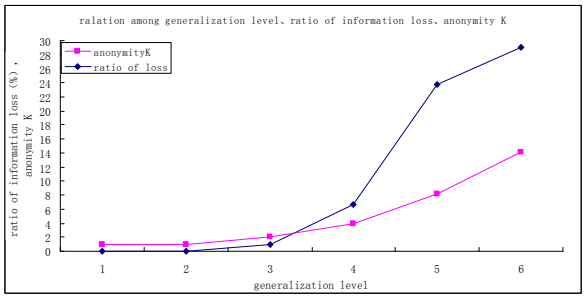

Figure 1. The generalization level influences on the ratio of information loss and K

According to the experience achieved by the data analysis process, we concluded that the ratio of information decreased with the number of record increasing, but K increased fast, thus the anonymity increasing. The data holder can selected a suitable k by two curves on the ratio of loss and K, next to generalize the data, finally release it. In the course of this, one needs to ensure the analyzability of data as well as achieve the purpose of anonymity.

V. CONCLUSION

Nowadays, online shopping has become a popular commercial operation mode. Commercial organizations mastered a lot of customers shopping information, including the user identity identification, Sex, Address, Zipcode, etc. When we analysis the user information, in order to protect the privacy of consumer against external attack, when commercial organizations release information, not only to protect the privacy of consumer, and analysis the data.

This paper presented a method of data processing before release data based on the concept of k-anonymity. The method

first recoded the information of identifying consumer to conceal, and adopt the generalization of information to protect the quasi-identifier. The paper analyzed the relation among the generalization level, the ratio of information loss, and K. The businesses selected the suitable result of a generalization through the steps of recoding, generalization, analysis on a specific database. This ensures to conceal the information of consumer, at the meantime providing the analyzability of data.

ACKNOWLEDGMENT

This work is partly supported by Shaanxi Provincial scientific research fund project (08JK481), and Xianyang Normal University research fund project（08XSYK337）. I thank Zhenglong XIE professor for his encouragement to write this paper. Finally, we wish to thank the reviewers for constructive and helpful comments.

REFERENCES

[1] Samarati P. Protecting respondents identities in microdata release[J]. IEEE Trans. on Knowledge and Data Engineering,2001,13(6):1010-1027.

[2] Sweeney L. Achieving K-anonymity privacy protection using generalization and suppression[J]. Int'l Journal on Uncertainty, Fuzziness and Knowledge Based Systems,2002,10(5):571-588.

[3] Sweeney L. K-Anonymity: A model for protecting privacy[J]. Int'l Journal on Uncertainty, Fuzziness and Knowledge Based Systems, 2002,10(5):557-570.

[4] Sweeney L. Achieving k-anonymity privacy protection using generalization and suppression[J]. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems,2002: 571-588.

[5] Wong RC,Li J,Fu AW,Wang K. (α,κ)-Anonyrnity:An enhanced K-anonymity model for privacy-preserving data publishing[J]. In : Eliassi-Rad T, Ungar LH, Craven M, Gunopulos D, eds. Proc. of the 12th Int' l Conf. on Knowledge Discovery and Data Mining. New York : ACM Press, 2006. 754-759.

[6] LeFewe KDeWm DJ, Ramakrishnan R. Incognito : Efficient full-domain K-anonymity[J]. In : 0zcan F, ed. Proc. ofthe Int' l Conf. on Management ofData. Maryland: ACM Press, 2005.49-60.