A Privacy Protection Model Based On K-Anonymity

Na Man[1, a *], Xin Li[1, b] and Kechao Wang[1, c]

[1]School of information Engineering, Harbin University, Harbin, China

[a]mn_0451@163.com, [b]114352231@qq.com, [c]erickcwang@126.com

Keywords: Privacy protection; K-anonymity; Sensitive attribute

Abstract. In this paper, we focuse that the existing k-anonymity does not fully consider the privacy protection degree issues of sensitive attribute, proposing a (p, α)-sensitive k-anonymity privacy protection model based on privacy protection degree grouping of sensitive attribute. The solution can not only effectively protect highly sensitive private information and reduce the risk of privacy leakage, but also reduce loss of information from the anonymous processing to improve the quality of the data.

Introduction

Networking technology, a large capacity storage technology and the rapid development of computing technology make large amounts of personal data automatically collected and published more and more conveniently. However, the problems of individual privacy leakage have become increasingly prominent in the process of data publishing, thus individual privacy protection issues are also paid more and more attention by people. In recent years, many anonymous privacy protection methods have been proposed, such as, sensitive attribute mining[1], anonymous model[2], anonymous publishing of high-dimensional data[3], and so on. The object of privacy protection in data publishing is primarily the correspondence between user' sensitive data and personal identity.

In order to solve problems of privacy attack and leakage in the data publishing, based on traditional methods of access control, Samarati Sweeney in 2002 proposed a new model for preventing leakage of private information, the model is called k-anonymity model [4], which is simple, practical and easy to implement. The model can make sure that the shared data authenticity and availability of information at the same time, and can also effectively protect privacy data from linking attack. The k-anonymity model have been widely used in recent years and become the main model of data publishing. At present, although an increasing number of researchers begin to pay attention to this field, the problems of privacy protection in data publishing still have a lot of contents that are worth studying. K-anonymity technology of protecting privacy information has important theoretical value and practical significance. The researches of k-anonymity technology will provide a strong support for protection of privacy information in the future, and promote study of practical application of anonymous privacy protection technology.

The problems of k-anonymity model

K-anonymity model is a typical data publishing model, and differs from privacy protection technology based on traditional access control and so on, in order to meet the anonymous need, it begins to preprocess the original data sets and then releases after processing data sets.[5] K-anonymity is proposed in specific application background. In various application occasions of needing data publishing such as electing, seeking job, medical care, and so on, it can not only conceal personal identify information, but also ensure that privacy information can't be inferred through the released information related to voters, job seekers, and patients, which are important significance of the privacy protection in data publishing.

K-anonymity can effectively prevent information disclosure from linking attack, but the published data of k-anonymity can still cause the disclosure of privacy, because it does not take into account diversity of sensitive attribute, and it is vulnerable to consistent attack. Although l-Diversity model[6] have solved the problem of consistent attack, due to the uncertainty of

background knowledge attack, how to set up l-Diversity Model's parameter have not good way. l-Diversity Model still has a privacy leakage when sensitive attribute distribution is uneven.

In most cases, the sensitivity degree level of sensitive attribute values are not the same, so they have the different level of protection, that is the higher sensitivity of attribute values, the higher the degree of their privacy protection. There are some diseases that the sensitivity degree is the highest, such as AIDS, but there are many common diseases, such as colds, stomach etc. Therefore, we may set different protection degree for different sensitive attribute values. K-anonymity is a good solution to the problem of data tables' linking attack, but it does not make any constraints for sensitive data.

(P，α)-sensitive k-anonymity model

On the basis of studying k-anonymity, this paper analyzes and researches the issue of the attribute disclosure of sensitive information [7]. In order to avoid attribute values of the high privacy protection degree at the same time occurring in the same equivalence class to cause the problem of privacy leak, we propose a (p，α)-sensitive k-anonymity model of privacy protection based on privacy protection degree groupings of sensitive attributes. The practical application of this model is more suitable for data sets of obviously strong or weak privacy protection degree of sensitive attributes, such as hospital case data sets. The basic idea of (p，α)-sensitive k-anonymity model is that divides sensitive attributes into groupings according to the privacy protection degree of sensitive attribute (that is, sensitivity). It requires that attribute values for the same sensitivity are in the same sensitive attribute grouping, so in this case, the model takes into account not only protection of a certain sensitive attribute value, but also protection of the whole grouping of sensitive attribute.

Firstly, this model analyzes sensitive attribute values according to privacy protection degree of sensitive attribute, and gives the corresponding grouping situation of sensitive attribute, which makes that sensitive attribute values in the same grouping are the same sensitivity, and the sensitivity of sensitive attribute values in the different grouping don't be same; then according to the high and low of privacy protection degree of sensitive attribute in sensitive attribute grouping, setting different constraint of Grouping Privacy Leakage rate, that is every privacy leakage rate of sensitive attribute grouping is not higher than the constraint value, if privacy protection degree are more higher, the constraint is more strict, which makes that attribute values of higher privacy protection degree seldom appear in the same equivalence class, so that to provide different intensity protection for the different sensitive attribute values, and ensure that each equivalence class consists of at least *p* different sensitive attribute groupings.

The correlative concepts in the model will be defined in the following section.

Definition 1 Sensitive attribute grouping Let S be a collection of sensitive attribute values in original data table T, and according to privacy protection degree of sensitive attribute in the collection S, S is divided into m groupings of ordered value (S_{g1}, S_{g2} ，...，S_{gm}) which are regarded as value domain of sensitive attribute collection S, which $S = \bigcup_{i=1}^{m} S_{gi}$, $S_{gi} \bigcap S_{gj} = F(i?\ j)$. When $S_{g1} £ S_{gk}$ $(1 \# l \quad k?\ m)$, S_{gl} is more sensitive than S_{gk}, and privacy protection degree of it is stronger.

Definition 2 Grouping privacy leakage rate (GPleakage) The percentage of leakage of private information. The privacy leakage rate of any sensitive groupings in the same equivalence class E is defined as the following formula:

$$GPleak(S_{gi}) = \frac{Count(E, S_{gi})}{Count(N_E)} \tag{1}$$

Among them, the function Count (E, S_{gi}) is the number of attribute values of privacy protection degree S_{gi} of sensitive attribute grouping in the equivalence class E, and Count(N_E) is the total number of tuples in the equivalence class E .

Definition 3 Privacy leakage rate constraint Given a data set D, quasi-identifier(QI) attribute and sensitive attribute grouping S is a collection of sensitive attribute values, where S={ S_{g1}, S_{g2}, …, S_{gm} } , S is divided into groups based on the sensitivity of sensitive attribute. (E, S_{gi}) expresses the level of privacy of sensitive attribute grouping S_{gi} in the equivalence class E. α is parameter that is given by experts, where $0 \# a \quad 1, a_{s_{g1}}, a_{s_{g2}}, \cdots, a_{s_{gm}}$ is respective privacy leakage rate constraint α of each sensitive groupings, where $a_{s_{g1}} < a_{s_{g2}} < \ldots < a_{s_{gm}}$. If the privacy leakage rate of privacy protection degree S_{gi} of sensitive attribute in all equivalence classes of the data set D is not higher than the sensitive attributes grouping's α values, namely $GPleak(S_{gi}) £ a_{s_{gi}}$, the data set D satisfies privacy leakage rate constraint α .

Setting constraint value is very important, which is generally given by experts. According to the grouping of privacy protection degree of sensitive attribute values, experts provide different constraint value α for different sensitive attribute grouping, the size of the value is determined by leaked severity of sensitive attribute grouping's attribute value in the equivalence class. In order to provide the higher level of protection for sensitive attribute grouping's values, the attribute grouping can be set as the smaller constraint value α , and the lower level of protection can be set as the larger constraint value α , which is even for 1, that is these sensitive attribute values may not be protected. Privacy leakage rate constraint α not only can intuitionistic reflect the level of data privacy protection, but also can avoid the higher sensitivity of attribute values to appear in the same equivalence class.

Definition 4 (p，α)-sensitive k-anonymity The original data table T is anonymized to data table AT which satisfies the following conditions: the property of k-anonymity; each equivalence class contain at least p different sensitive attribute groupings; any sensitive attribute groupings in the equivalence class satisfies privacy leakage rate constraint α , well then said anonymous table AT satisfies (p， α)-sensitive k-anonymity.

A specific example is given in the following, for original data Table 1 and the sensitive attribute grouping Table 2. According to the different requirement of protection level of sensitive attribute values, where set p=2, k=4, $a_{s_{g1}}$ =0.5, $a_{s_{g2}}$ =0.6, $a_{s_{g3}}$ =0.75, $a_{s_{g4}}$ =1. Table 3 shows that attribute values for different sensitive groupings in each equivalence class distribute evenly, and the probability that attribute value of the higher sensitivity is inferred is low, so as to effectively prevent leakage of attribute value of the higher sensitivity and enhance the security of data dissemination.

Table 1 The original data table

	Age	Race	Zip code	Disease
1	21	Black	14128	Cancer
2	39	White	23141	Hepatitis
3	48	Black	14150	Gastric pain
4	27	White	14156	Cancer
5	36	Black	23143	tuberculosis
6	42	Black	14156	Fever
7	22	Black	14152	AIDS
8	32	White	23114	Heart disease
9	49	Black	14156	Fever
10	24	Black	14120	AIDS
11	35	White	23141	High blood pressure
12	45	Black	14120	Fever

Table 2 Sensitive attribute grouping

Grouping ID	Sensitive property value collections	Grouping Privacy
1	AIDS, cancer	S_{g1}
2	Tuberculosis, hepatitis	S_{g2}
3	Heart disease, high blood pr essure	S_{g3}
4	Fever, stomach pain	S_{g4}

Table 3 (3，α)-sensitive 4-anonymous table

	Age	Race	Zip code	Disease	Grouping ID
1	[20,50)	Black	141**	Cancer	1
12	[20,50)	Black	141**	Fever	4
10	[20,50)	Black	141**	AIDS	1
6	[20,50)	Black	141**	Fever	4
2	[30,40)	*	231**	Hepatitis	2
11	[30,40)	*	231**	High blood pressure	3
8	[30,40)	*	231**	Heart disease	3
5	[30,40)	*	231**	Pulmonary tuberculosis	2
3	[20,50)	*	141**	Gastric pain	4
7	[20,50)	*	141**	AIDS	1
9	[20,50)	*	141**	Fever	4
4	[20,50)	*	141**	Cancer	1

Summary

This paper studied the existing privacy leakage problem of k-anonymity model. For the existing potential privacy leakage problems in k-anonymity model, taking into account privacy protection degree of sensitive attribute after anonymity, the paper propose a (p,α)-sensitive k-anonymity model based on privacy protection degree grouping of sensitive attribute, and the model effectively protects private information. The scheme of privacy protection model proposed in this paper still has some insufficient on the aspect of theories and technical problems. In the future data publishing for multiple sensitive attribute and anonymous model for supporting dynamic update to publish data table, and so on, will be improved and in-depth studied.

References

[1] Prakash M, Singaravel G. An approach for prevention of privacy breach and information leakage in sensitive data mining[J]. Computers & Electrical Engineering, 2015, 45: 134-140.

[2] Liu X, Xie Q, Wang L. Personalized extended (α, k)-anonymity model for privacy-preserving data publishing[J]. Concurrency and Computation Practice and Experience, 2016,29 (6).

[3] H Zakerzadeh , CC Aggarwal ,K Barker. Managing dimensionality in data privacy anonymization[J]. Knowledge and Information Systems,2015,45:1-33.

[4] L. Sweeney. k-anonymity:A model for protecting privacy. International Journal on Uncertainty, Fuzzi-ness and Knowledge-based Systems. 2002, 10(5):557-570.

[5] L. Sweeney. K-Anonymity:A Model for Protecting Privacy. International Journal on Uncertainty,Fuzzi-ness and Knowledge-based Systems. 2002,10(5):557-570.

[6] A. Machanavajjhala, J.Gehrke, and D. Kifer. l-diversity:privacy beyond k-anonymity. Proceedings of the 22nd International Conference on Data Engineering. Atlanta,GA,USA,

2006:24-36.
[7] D. Lambert. Measures of Disclosure Risk and Harm. Journal of Official Statistics. 1993,9:313-331.