Advanced Engineering Forum *Online: 2012-09-26*
ISSN: 2234-991X, Vols. 6-7, pp 64-69
doi:10.4028/www.scientific.net/AEF.6-7.64

Uncertain Data Privacy Protection Based on K-anonymity Via Anatomy

Ren Xiangmin [1,2, a] Yang Jing [1,b] Zhang Jianpei[1,c] Jia Zongfu[2,d]

[1]College of Computer Science and Technology, Harbin Engineering University, Harbin, China

[2]School of Software, Harbin University, Harbin, China

[a]min0070@sina.com, [b]yangjing@hrbeu.edu.cn, [c]zhangjianpei@hrbeu.edu.cn, [d]sjjzf@hrbu.edu.cn

Keywords: k-anonymity, uncertain data, privacy protection, UADK-anonymity

Abstract. In traditional database domain, *k*-anonymity is a hotspot in data publishing for privacy protection. In this paper, we study how to use *k*-anonymity in uncertain data set, use influence matrix of background knowledge to describe the influence degree of sensitive attribute produced by QI attributes and sensitive attribute itself, use BK(L,K)-clustering to present equivalent class with diversity, and a novel UDAK-anonymity model via anatomy is proposed for relational uncertain data. We will extend our ideas for handling how to solve privacy information leakage problem by using UDAK-anonymity algorithms in another paper.

Introduction

With the rising of data mining technology and the appearances of data stream and uncertain data technology etc, individual data, the enterprise data are possibly leaked at any moments, so the data security has become nowadays the main topic of information security. With the development of Sensor network, Web service and RFID in recent years, uncertain data has become ubiquitous in economy, military, logistics, finance, telecommunication areas and so on. Uncertain data management and privacy protection have become an important research direction and a hot area of research[1].

K-anonymity [2], a model put forward by Samarati P and Sweeney L in 1998 to avoid privacy leaks, requests existence of a certain amount of unrecognizable individuals in the publicized data which make the aggressor disable to distinguish the concrete individual of privacy, and prevent the leak of individual privacy. *K*-anonymity got the universal concern of the academic circles, and a lot of scholars research and develop the technology on different levels. But it was a *k*-anonymity privacy protection model of deterministic data, currently, research in uncertain data publishing based on *k*-anonymity is limited, it needs a new model to represent the *k*-anonymity privacy protection of uncertain data.

Charu C. Aggarwal [3] presents an uncertain version of the *k*-anonymity model, which has the additional feature of introducing greater uncertainty for the adversary over an equivalent deterministic model. He tests the effectiveness of the privacy transformation on the problems of query estimation and classification, and show that the technique retains greater accuracy than other *k*-anonymity models. Wu jiawei, et al. explore several new modeling methods[4]. A model space which consists of K_{attr}, K_{tuple}, $K_{upperlower}$ and K_{tree} model is built, the K_{attr} model uses the *attribute-ors* ways to describe the uncertainty in the quasi-identifier attribute(*QI*) values of the *k*-anonymity privacy protection model, the K_{tuple} model takes *QI* values as relations and use the *tuple-ors* ways to describe the relations. The completeness and closure about these models are discussed.

This paper explores a new *k*-anonymity privacy protection model for relational uncertainty data by anatomy.

Related concepts

k-anonymity

Definition 1:*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 $RT[QI_{RT}]$[5].

Table 1 is an example of *k*-anonymity, sensitive attribute(*SI*) is *Disease*, QI_T*={Race, Education, Age, Sex, ZIP}* and *k*=2. In particular, $t1[QI_T]= t2[QI_T]$, $t3[QI_T]= t4[QI_T]$, $t5[QI_T]= t6[QI_T] = t7[QI_T]$, $t8[QI_T]= t9[QI_T]$.

Table 1 Example of *k*-anonymity, where *k*=2 and *QI*={*Education*, *Age*, *Sex*, *ZIP*}

Num	Education	Age	Sex	ZIP	Disease
t1	Master	(15,40]	F	3115**	flu
t2	Master	(15,40]	F	3115**	lung cancer
t3	Master	(40,60]	M	3114**	lung cancer
t4	Master	(40,60]	M	3114**	lung cancer
t5	Bachelor	(40,60]	F	3114**	lung cancer
t6	Bachelor	(40,60]	F	3114**	short breath
t7	Bachelor	(40,60]	F	3114**	obesity
t8	Ph.D	(40,60]	Person	3114**	mammary cancer
t9	Ph.D	(40,60]	Person	3114**	mammary cancer

Generalization. Given an attribute *A*, a generalization for an attribute is a function on *A*. That is, each f: $A \to B$ is a generalization, it also says that: $A_0 \xrightarrow{f_0} A_1 \xrightarrow{f_1} \cdots \xrightarrow{f_{n-1}} A_n$ is a generalization sequence or a functional generalization sequence [6]. Fig.1 provides an example of generalization hierarchies.

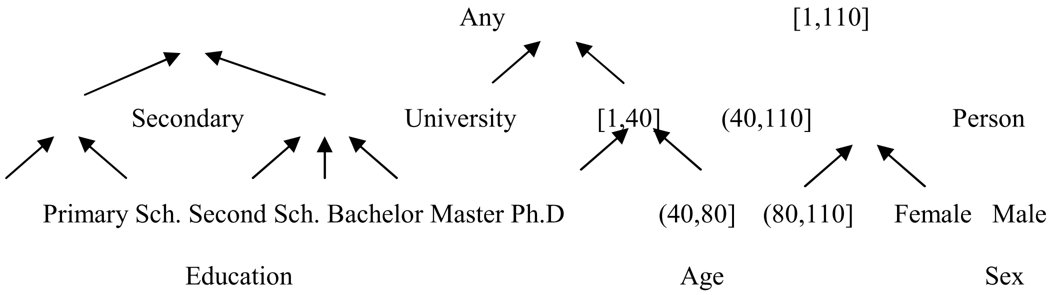

Fig.1 Gerneralization Hierarchies of {*Education*, *Age*, *Sex*}

Influence matrix based on background knowledge. Background knowledge describes the influence of a variety of *SI* produced by *QI*[7,8]. Background knowledge can be acquired from domain expert, and also can be acquired by analyzing basic data directly.

We use relation and sensitive degree matrix $M|S$ to describe the influence degree of *SI* produced by *QI* and *SI* itself, introducing notation as follows:

t_{ij} :the influence degree of NO. *j SI* produced by NO. *i QI*.

b_i :the weight of *SI* value of NO. *i*.

Influence matrix $M|S$ is with *m* rows and *n+1* columns, *m* is the number of *SI*, *n* is the number of *QI* attribute, then the matrix is as follows:

$$M|S = (t_{ij}|b_i)_{m\times(n+1)} = \left[\begin{array}{cccccc|c} QI_1 & QI_2 & QI_3 & QI_4 & \cdots & QI_n & S \\ t_{11} & t_{12} & t_{13} & t_{14} & \cdots & t_{1n} & b_1 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ t_{i1} & t_{i2} & t_{i3} & t_{i4} & \cdots & t_{in} & b_i \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ t_{m1} & t_{m2} & t_{m3} & t_{m4} & \cdots & t_{mn} & b_m \end{array}\right] \quad (1)$$

The weight value of t_{ij} and b_i is specified by expert or experience value, for example, we can divide weight of *QI* in Table 1 into 5 grades, 1,0.8,0.4,0.1,0, and divide weight of *S* in Table 1 into 5 grades, 0.10,0.30,0.50,0.80,0.90. The *flu* is common ailments, *disease* weight can use 0.11, because of the characteristic of local outbreaks of *flu*, *ZIP* weight use 0.8, *Sex* weight use 0.2 etc. The *disease* weight of *obesity* can use 0.12, the *disease* weight of *flu* and *obesity* are all 0.1, 0.01 and 0.02 denotes different ailment. The *disease* weight of *short breath* is 0.31, the major *diseases* weight of *lung cancer*, *mammary cancer* and *AIDS* use 0.91, 0.92 and 0.93, different *disease* must have different *disease* weight value. Then the relation and sensitive degree matrix based on Table 1 is as follows:

$$M|S = (t_{ij}|b_i)_{9\times(5+1)} = \begin{array}{c} \begin{array}{cccccc} \text{Race} & \text{Education} & \text{Age} & \text{Sex} & \text{ZIP} & \text{Disease} \end{array} \\ \left[\begin{array}{ccccc|c} 0 & 0 & 0 & 0.2 & 0.8 & 0.11 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0.4 & 1 & 0 & 0.92 \\ 0 & 0 & 0.4 & 1 & 0 & 0.92 \end{array}\right] \end{array} \quad (2)$$

Anatomy. Anatomy was proposed by Xiaokui Xiao et al, it means *QI* and *SI* published in different table, instead of publishing one single table with the generalized values, *QI* table included a unique identifier: equivalent class(QI-group) ID, *SI* table included equivalent class ID too, *SI* of each QI-group, and count. Anatomy overcomes the drawbacks of generalization. Extensive experiments confirm that anatomy permits researchers to derive from the published tables, highly accurate aggregate information about the unknown microdata, with an average error below 10% [9]. For example, table 3 satisfied 3-diversity anatomy table according to table 2 by anatomy. In paper [10], they proved that the resulting published tables NSS(*QI*) and SS(*SI*) satisfy *p*-sensitive *k*-anonymity property, that is to say, anatomy satisfy *p*-sensitive *k*-anonymity property.

Table 2 The original data table

Sex	**ZIP**	**Disease**
F	311578	flu
F	311579	chest pain
F	311579	hypertension
M	311581	obsity
M	311582	short breath
M	311582	hypertension
F	311588	obesity
F	311588	chest pain

Table 3 The 3-diversity data table by anatomy

QI attribute		
Sex	ZIP	ID
F	311578	1
F	311579	1
M	311579	1
M	311581	2
M	311582	2
F	311582	2
F	311588	3
F	311588	3
F	311589	3

Sensitive attrbute	
ID	Disease
1	flu
	chest pain
	hypertension
2	obsity
	short breath
	hypertension
3	obesity
	chest pain
	cancer

Constructs for Uncertainty. There are two different constructs for u-tuples(*uncertain tuples*)[11,12]:

attribute-ors: An *attribute-or* in a u-tuple specifies a set of alternative values for an attribute. For example, *t1* contains an *attribute-or* in its first field and represents one of two possible tuples: (Bachelor, insomnia) or (Master, insomnia).

Table 4 Uncertain data table with *attribute-ors* construct

Num	Education	Disease
t1	{Bachelor, Master}	insomnia
t2	Second School	flu
t3	Ph.D	mammary cancer
t4	Master	short breath
t5	Primary School	lung cancer
t6	Ph.D	mammary cancer

tuple-ors: A tuple-or in a u-tuple specifies a set of possible tuples. For example, the uncertainty in the previous example can also be represented by:

Table 5 Uncertain data with *tuple-ors* construct

(Bachelor, insomnia)‖(Master, insomnia)

K-anonymity privacy protection model for uncertain data via anatomy

First, we preprocess the uncertainty data table which makes the uncertainty data table become a deterministic data table, namely, the uncertain data has been generalized, for example, {*Bachelor*, *Master* }→*University*, then model the data of deterministic data table by *k*-clustering and anatomy. When we create a deterministic data table from an uncertainty data table, each uncertainty *QI*(*attribute-or*) is labeled with *QIID* or *TupleID* attribute in order to keep the uncertainty of *QI* of original data in uncertainty data table, *QIID* or *TupleID* is a appended attribute column, which value represents location in the deterministic data table. At the same time, we divide the uncertainty *SI* into two or more fields of *SI* in order to keep the uncertainty of *SI* of original data in uncertainty data table. That is to say, uncertain data is stored in a relational database, then we can use traditional *k*-anonymity model to represent the privacy protection of uncertain data.

UDAK-anonymity model. UDAK-anonymity model(uncertain data anatomy *k*-anonymity model) is built for *attribute-ors* construct, modeling process needs three steps: preprocessing, BK(L,K)-clustering [13], and anatomy.

Preprocessing

1. Create deterministic table by generalization and partition

Definition 2. Deterministic uncertian data table. $T(A_1, ..., A_n)$ is an uncertain data table, if T can be change into deterministic data table T' by generalization and partition, we say T' is a deterministic uncertian data table.

Table 6 The original data table including uncertain data with *attribute-ors* construct

Num	Education	Age	Sex	Disease
t1	{ Bachelor, Master }	25	M	insomnia
t2	Bachelor	21	M	{ obesity, flu }
t3	Ph.D	35	F	mammary cancer
t4	{Master, Ph.D}	{41,48}	M	{short breath, obesity}
t5	Master l	45	M	lung cancer
t6	Ph.D	36	F	mammary cancer

Generalize *QI*s which include uncertain data with *attribute-ors* construct and divide the uncertainty *SI* into two or more fields of *SI*. For example, Table 6 is the original data table including uncertain data with *attribute-ors* construct, Table 7(deterministic uncertian data table) is the deterministic data table by generalization and partition according to Table 6. {*Bachelor*, *Master*}→ *University* in t1, {*obesity*, *flu*}→t2[*disease1*]= *obesity*, *t2*[*disease2*]= *flu* in *t2*, {*Master*, *Ph.D*}→ *University*, {*41*,*48*} → max {*41*,*48*}=*48*, {*short breath*, *obesity*} → *t4*[*disease1*]= *short breath*, *t4*[*disease2*]=*obesity* in *t4*. In Table 7, *t1*[*QIID*]=*22* means the second field(*Education*) is the generalization value of uncertain data according to generalization hierarchies(Fig.1), and it has two uncertain data(two child nodes), *t2*[*QIID*]=*SI2* means the uncertainty *SI* attribute was divided into two fields of *SI* attribute, *t2*[*disease1*]=*obesity*, *t2*[*disease2*]=*flu*, so does *t4*. If t_i[*QIID*]=0, it represent that t_i is a deterministic data.

Table 7 The deterministic data table by generalization and partition according to Table 6

Num	QIID	Education	Age	Sex	Disease1	Disease2
t1	22	University	25	M	insomnia	
t2	SI2	Bachelor	21	M	obesity	flu
t3	0	Ph.D	35	F	mammary cancer	
t4	2232SI2	University	48	M	short breath	obesity
t5	0	Master	45	M	lung cancer	
t6	0	Ph.D	36	F	mammary cancer	

2. Create influence matrix based on background knowledge according to section 2.3

BK(L,K)-clustering

Definition 3. K-Clustering. *S* is a data set which has *n* tuples, *K* is anonymous parameter, K-Clustering is Clustering set: $\varepsilon = \{e_1, \ldots, e_m\}$ which satisfies the following conditions:

(1) $\forall i \neq j \in \{1, \ldots m\}, e_i \cap e_j = \varnothing$;(2) $\bigcup_{i=1,\ldots,m} e_i = S$;(3) $\forall e_i \in \varepsilon, |e_i| \geq K$

(4) $\frac{1}{m}\sum_{l=1,\ldots,m} |e_l| * MAX_{i,j=1,\ldots,|el|} \Delta(t(l,i), t(l,j))$ is minimum

Here $|e_l|$ is the amounts of tuples in clustering e_l , $t(l,i)$ is NO. *i* tuple in clustering e_l , $\Delta(t(l,i), t(l,j))$ is the distance between NO. *i* tuple and NO. *j* tuple, $MAX_{i,j=1,\ldots,|el|} \Delta(t(l,i), t(l,j))$ is maximum distance in clustering e_l [13].

Definition 4. BK(L,K)-clustering((L,K)-Clustering based on influence matrix of background knowledge). *T* ($A_1, \ldots, A_n$) is a table, if *T* satisfies K-Clustering, and satisfies the following conditions:

(1) $\forall b_i < c$ in clustering e_m , all tuples in e_m should be anatomized directly. Otherwise must satisfy condition (2) . Here, threshold *c*>0, b_i is *S* column vector in influence matrix $\mathrm{M}|\mathrm{S}$, $1 \leq i \leq |e_m|$, $|e_m|$ is the amounts of tuples in clustering e_m .

(2) $L = \sum_{j=1,\ldots,|em|} count(|b_i - b_j| > 0), 1 \leq i \leq |e_m|, b_i, b_j$ is *S* column vector in influence matrix $\mathrm{M}|\mathrm{S}$, *L* is the amounts of different sensitive attribute value, and *L* makes sensitive attribute diversity, otherwise further improve the generalization or suppression.

We say T satisfies BK(L,K)-clustering.

BK(L,K)-anonymity with anatomy

Definition 5. BK(L,K)-anonymity with anatomy. ((L,K)-anonymity with anatomy based on influence matrix of background knowledge). *T* ($A_1, \ldots, A_n$) is table, if *T* satisfies BK(L,K)-clustering, then we divided *T* into *QI* table(QIT) and *SI* table(ST). Specifically, the QIT includes all its exact *QI* values, together with its group membership in a new column Group-ID. However, QIT does not store any *SI* values, ST retains *SI* statistics of each QI-group, Group-ID and count.

Definition 6. UDAK-anonymity. *T* ($A_1, \ldots, A_n$) is an uncertain data table, *T'* is a deterministic uncertian data table from *T*, and satisfies BK(L,K)-anonymity with anatomy, we say *T'* satisies UDAK-anonymity.

For instance, Table 8 which were anatomized according to table 7 satisfied UDAK-anonymity.

Table 8 The anatomized tables according to table 7

QIT

Num	QIID	Education	Age	Sex	ID
t1	22	University	25	M	1
t2	0	Bachelor	21	M	1
t3	0	*	*	*	2
t6	0	*	*	*	2
t4	2232	University	48	M	3
t5	0	Master	45	M	3

ST

ID	Disease1	Disease2	
1	insomnia		1
1	obesity	flu	1
2	mammary cancer		2
2	mammary cancer		
3	short breath	obesity	1
3	lung cancer		1

Similarly, we can use UDAK-anonymity model to deal with *tuple-ors* construct of uncertainty, owing to the limitation of the scope, I won't discuss it in this post.

Conclusion

This paper proposed specific modeling method of k-anonymity privacy protection of uncertain data via anatomy, and presented new models of k-anonymity privacy protection, UDAK-anonymity. UDAK-anonymity model not only kept the characteristic of uncertian data, but also provided more useful information for the user, improved the utility of uncertain data. Owing to the limitation of the scope, we will extend our ideas for handling how to solve privacy information leakage problem by using UDAK-anonymity algorithms in another paper.

Acknowledgement

The authors want to thank the helpful comments and suggestions from the anonymous reviewers. This work was supported by the National Natural Science Foundation of China (Grant No. 61073043), the Natural Science Foundation of Heilongjiang Province of China (Grant No. F200901).

References

[1] ZHOU Aoying et al. A Survey on the M anagement of Uncertain Data, Chinese journal of Computers. 32(2009) 1-15.

[2] Samarati P, Sweeney L, Generalizing data to provide anonymity when disclosing information, Proceedings of the Seventeenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, PODS, (1998) 188.

[3] Aggarwal Charu.C, On Unifying Privacy and Uncertain Data Models, Proceedings of the 2008 IEEE 24th International Conference on Data Engineering. (2008) 386-395.

[4] Wu Jia-wei, Liu Guo-hua, Wang Mei,Modeling the Uncertain data in data in the K-anonymity Privacy Protection Mode., CPMPUTER EINGINEERING & SCIENCE. 33(2011) 7-13.

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

[6] L. Sweeney, Achieving k-anonymity privacy protection using generalization and suppression, International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10 (2002) 571-588.

[7] Xiangmin, Ren, et al., “Research on CBK(L,K)-anonymity algorithm”, International Journal of Advancements in Computing Technology.3(2011) 165-173.

[8] LI Tai-yong, et al., “k-Anonymity via Twice Clustering for Privacy Preservation”, Journal of Jilin University. 27(2009) 173-178.

[9] XiaoXiaokui, TaoYufei, Anatomy: Sample and effective protect preservation, Proceedings of the 32nd International Conference on VeryLarge DataBases. (2006)139-150.

[10] Xiaoxun Sun, Hua Wang, Jiuyong Li, David Ross, Achieving p-Sensitive k-Anonymity via Anatomy, Proceedings of the 2009 IEEE International Conference on e-Business Engineering. (2009) 199-205.

[11] Anish Das Sarma, Shubha U. Nabar, Jennifer Widom, Representing Uncertain Data: Uniqueness, Equivalence, Minimization, and Approximation.. Technical Report, Stanford Infolab (2005).

[12] Anish Das Sarma, Omar Benjelloun, Alon Halevy, Shubha Nabar, JenniferWidom, Representing Uncertain Data: Models, Properties, and Algorithms. The VLDB Joural. 18(2009) 989-1019.

[13] Ren Xiangmin, Yang Jing, Zhang Jianpei, Wang Kechao, " Research on CBK(L,K)-Anonymity Algorithm ", International Journal of Advancements in Computing Technology. 3(2011) 165 -173.