ISSN : 0974 - 7435 Volume 10 Issue 24

2014

BioTechnology

An Indian Journal

FULL PAPER

BTAIJ, 10(24), 2014 [14801-14808]

How to realize update in *K*-anonymity model

Jinling Song[1]*, Liming Huang[1], Gang Wang[1], Qianying Cai[1], Yan Gao[2]

[1]HeBei Normal University of Science & Technology, Qinhhuangdao, 066004, (CHINA)

[2]Liaoning Institute of Science and Technology, Benxi, 117004, (CHINA)

E-mail: songjinling99@126.com

ABSTRACT

K-anonymity is a typical privacy model which can guarantee the safety of publishing dataset, however, the k-anonymized dataset contains generalized value and it difficult to bring it into correspondence with the original dataset directly. We at first create the index table basing on the one-one mapping between original tuple and its generalized tuple, which can be used to update the generalized tuple. To locate the QI group where an original tuple is in or should be inserted in, the definition of tuple-QG semantic similarity degree is presented and the QI group is located basing on tuple-QG semantic similarity degree. To merge the QI group whose size is smaller than k, QG semantic similarity degree are presented and used to find the similar QI group. Finally, the update algorithms basing on Semantic for the k-anonymized dataset are presented.

KEYWORDS

k-anonymity; Update; Semantic; Similarity; Generalize.

INTRODUCTION

K-anonymity[1] is an important privacy model which protects privacy by making each tuple repeats at least *k* times on the quasi-identifier (*QI*) through generalizing attribute values. The generalizing process is called *k*-anonymization and the formed dataset is called *k*-anonymized dataset. Since *k*-anonymized dataset includes fuzzy (generalized) data, it is difficult to update naturally when the original dataset update. However, if we generate the *k*-anonymized dataset again after each update of original dataset, it will waste most resources of computer and may result in multiple versions of *k*-anonymized dataset and information leakage[2]. Because the updating dataset is small under normal circumstances, regenerate *k*-anonymized dataset is not fit for *k*-anonymity model. So, in this paper we will discuss how to update the *k*-anonymized dataset directly following it's original dataset, which is very important for *k*-anonymity.

Considering the fuzzy (generalized) data in the *k*-anonymized dataset, the connection between original tuple and generalized tuple ought to be built to update the generalized tuple directly. So we create an index table between original dataset and *k*-anonymized dataset at first, in which each original tuple mapping to one generalized tuple. Since the update operation contains insert, delete and modify, the update operation of *k*-anonymized dataset may be different. To insert or delete a generalized tuple, the *QI* group where the generalized tuple is in or should be inserted in will be located firstly. We present the definition of tuple-*QG* semantic similarity degree, and locate the *QI* group according to the least tuple-*QG* semantic similarity degree. To the modify operation, it can be decomposed into insert and delete. After one *QI* group is updated in the *k*-anonymized dataset, it maybe smaller than *k* and violate the *k*-anonymity constraint (the candidate of each *QI* group no less than *k*). So the *QI* group will be merged with other *QI* group to maintain the *k*-anonymity constraint. In order to find the similar *QI* group which can be merged, the *QG* semantic similarity degree is presented. Finally, the update algorithms basing on Semantic for the *k*-anonymized dataset are presented.

RELATED RESEARCH

Current researches on *k*-anonymity focused on anonymized methods or the improvement of *k*-anonymity model. Meyerson et al[3] and Agarwal[4] verified that achieving the highest precise *k*-anonymized table is a NP-hard problem. They presented O(*k*log*k*) and O(*k*) approximate algorithm respectively. Lefvre[5] gave a multi-dimension *k*-anonymity algorithm which can generalize multiple attributes simultaneously. The anonymized algorithm for high-dimension sensitive transactional data is proposed in[6]. A. Machanavajjhala et al[7] introduced a ℓ-diversity model which is better than *k*-anonymity model. Xiaokui Xiao[8] presented that the optimization of ℓ-diverse is NP-hard even there are 3 different sensitive value, then a (ℓ,d)-approximate algorithm is proposed. Basing on ℓ-diversity model, Junqiang Liu [9] presented ℓ^+-diversity model and an anonymized algorithm based on full sub-tree generalization. Ke Wang[10] pointed that the sensitive information in temporary data is slope and can't to satisfy ℓ-diversity, and proposed a tuple collocation strategy to construct ℓ-diversity. The quasi-sensitive attribute(QS) is presented in [11], in which QS ℓ-diversity and QS t-closeness model were proposed. Ren Xiangmin et al[12] proposed *CBK*(*L*,*K*)-anonymity algorithm which can make anonymous data effectively resist background knowledge attack and homogeneity attack by *K*-clustering based on influence matrix of background knowledge. Ren Xiangmin et al proposed *CBK*(*L*,*K*)-anonymity algorithm[13] to resist background knowledge attack and sample attack. Yinghua Liu et al[14] proposed a personalized privacy preserving parallel (alpha, *k*)-anonymity model based on *k*-anonymity to reduce high probability of the attributes in the equivalent group and reduce the probability of the likelihood of attack. An anonymized algorithm for multi-side cooperation under half-honesty model was proposed in [15].

Current researches on update algorithm of the *k*-anonymized dataset are as follows. Xiao X et al present "M-invariance" algorithm[16] to dynamic datasets, which assure the each QI group in different versions of generalized dataset has same sensitive attribute values when insert and delete operation is performed. K. LeFevre et al[17] update the *k*-anonymized dataset basing on Information loss metric. In this paper, we update *k*-anonymized dataset basing on the Semantic Similarity Degree, which will be an effective supplement for existing *k*-anonymity researches.

BASIC DEFINITION

In this paper, the dataset is a relational table as $R(A^{QI},A^S)$, where $A^{QI}=\{A_1^{QI},A_2^{QI},\ldots,A_n^{QI}\}$ is quasi-identifier, A^S is the sensitive attribute. For simplicity, we also use *R* denote dataset. For $A\subseteq A^{QI}\cup A^S$, $R[A]$ is the projection containing repetition values of table *R* on the attribute set *A*, $t[A]$ is the values of tuple *t* on the attribute set *A*.

Definition 1: ***k*-anonymity constraints** For dataset $R(A^{QI},A^S)$, if each tuple in $R[A^{QI}]$ counts at least $k(k\geq 2)$ times, then the dataset *R* satisfies *k*-anonymity constraints.

Example 1: When { *Age*, *Zip* } is quasi-identifier of dataset *R**(*Age*, *Zip*, *Problem*) (table 1(b)), for *R** [*Age*, *Zip*]={([21, 25], [11k, 20k]), ([21, 25], [11k, 20k]), ([41, 50], [21k, 30k]), ([41, 50], [21k, 30k]), ([51, 55], [51k, 60k]), ([51, 55], [51k, 60k]) }, so tuples ([21,25], [11k, 20k]), ([41, 50], [21k, 30k]) and ([51, 55], [51k, 60k]) are all count 2. Thus, *R** satisfies 2-anonymity constraints.

Definition 2: ***Generalization*** For a relation $R(A_1,A_2,\ldots,A_k)$, assume the domain of the attribute A_i is *D* and a partition of *D* is $\{u_1,u_2,\ldots,u_L\}$, where u_i $(1\leq i\leq L)$ is an integer interval. For any tuple $t\in R$, if there exists a function g: $t[A_i]\rightarrow u_i$ on the attribute A_i, where $t[A_i]\in u_i$, then we call *g* as *generalization function* of the attribute A_i, *g* $(t[A_i])$ is the generalization of $t[A_i]$.

Similarly, the generalization of R on the attribute set $\{A_1,A_2,\ldots, A_k\}$ denotes generalizing result of R on each attribute respectively, i.e. $g(t[A_1, A_2,\ldots, A_k])=(g(t[A_1]), g(t[A_2]),\ldots, g(t[A_k]))$.

Notes: Definition 2 fits generalized value too. The generalizing operation on the generalized value is a mapping process from a small range to a bigger range which includes the generalized value.

Table 1. Microdata R and its 2-anonymized dataset R^*, where $A^{QI}=\{Age, Zip\}$.

Tuple ID	*Age*	*Zip*	*Problem*
t1	21	12000	flu
t2	23	18000	gastritis
t3	48	28000	flu
t4	42	23000	gastritis
t5	49	25000	insomnia
t6	52	52000	flu
t7	53	59000	gastritis

(a) Microdata R

Tuple	*QG*	*Age*	*Zip*	*Proble*
t1*	1	[21, 25]	[11k, 20k]	flu
t2*	1	[21, 25]	[11k, 20k]	gastriti
t3*	2	[41, 50]	[21k, 30k]	flu
t4*	2	[41, 50]	[21k, 30k]	gastriti
t5*	2	[41, 50]	[21k, 30k]	insomn
t6*	3	[51, 55]	[51k, 60k]	flu
t7*	3	[51, 55]	[51k, 60k]	gastriti

(b) 2-anonymized dataset R^*

Example 2: Assuming the domain of attribute *Age* is [21-60], the first patition of the domain [21-60] is {[21-25],[26-30],…,[56-60]}, the second patition is {[21-30], [31-40],…,[51-60]}. Then the generalization on attribute *Age* in dataset R(table 1(a)) is: the values 21,23 were generalized to [21-25] respectively; the values 52,53 were generalize to [51-55] respectively; the values 48,42,49 were generalized by two steps, the first step result in [46-50], [41-45], [46-50] and the second step result in [41-50]. The generalization result were shown in table 1(b).

Definition 3: ***k*-anonymized dataset** For the dataset $R(A^{QI},A^S)$, if we generalize the values on A^{QI} and get the dataset R^* which satisfies k-anonymity constraints on A^{QI}, then the generalization process from R to R^* is called k-anonymization , dataset R^* is the k-anonymized dataset of R.

Example 3: R^*(table 1(b))is the generalization result of dataset R(table 1(a))on the quasi-identifier attributes $\{Age,Zip\}$. We know that R^* satisfies 2-anonymity constraints from example 1, so the generalization process from R to R^* is 2-anonymization of R and R* is a 2-anonymized dataset of R.

To distinguish tuples in R and R^*, we call tuples in R as 'tuple' and tuples in R^* as 'generalized tuple' below.

Definition 4: tuple-generalized tuple mapping Assuming the original dataset and the k-anonymized dataset are R and R^* respectively, for any tuple $t\in R$, if there exists $t^*\in R^*$ and where $t[A_i{}^{QI}] \in t^*[A_i{}^{QI}]$ $(1\le i \le n)$,$t[A^S]= t^*[A^S]$, then t^* is called the generalized tuple of t. Function gf: $t\rightarrow t^*$ is called the one-to-one mapping function from the tuple t to its generalized tuple t^*.

Example 4: For the original dataset R (table 1(a)) and its 2-anonymized dataset R^* (table 1(b)), $t1^*$ (in R^*) is the generalized tuple of $t1$ (in R), similarly, $t2^*$ is the generalized tuple of $t2$, …, $t7^*$ is the generalized tuple of $t7$.

We can build the index table between R and R^* basing on the one-to-one mapping between each tuple and its generalized tuple.

Definition 5: *QI* group For the k-anonymized dataset $R^*(A^{QI},A^S)$, the generalized tuples with the same value in $R^*[A^{QI}]$ are called a *QI* group, i.e. *QG*.

The *QI* groups in R^* are denoted as $QG(R^*)=\{QG_1, QG_2,\ldots, QG_m\}$, where $|QG_i|\ge k$(k-anonymity constraints), $QG_i \cap QG_j=\varnothing(1\le i, j\le m, i\ne j)$ and $|QG_1|+|QG_2|+\ldots+|QG_m|= |R^*|$.

Example 5: In the 2-anonymized dataset R^*(table 1(b)), for $t_1[Age, Zip]=([21, 25], [11k, 20k])$, $t_2[Age, Zip]=([21, 25], [11k, 20k])$, tupls t_1,t_2 are a *QI* group; in the same way, t_3,t_4,t_5 are a *QI* group, t_6,t_7 are a *QI* group, i.e. $QG(R^*)=\{QG_1=\{t_1,t_2 \}, QG_2=\{t_3,t_4, t_5\},QG_3=\{t_6,t_7\}\}$.

The update operations (insert, delete and modify) in the table R are expressed as follows:

INSERT (R, T) : Insert the tuple sets $T=\{t_1,t_2, \ldots,t_k\}$ to table R, where t_i $(1\le i \le k)$ is a tuple on the attribute set $\{ A^{QI},A^S \}$.

DELETE(R, φ_D): Delete the tuples satisfying condition φ_D in R.

MODIFY(R, φ_M, F_M): Modify the tuples satisfying condition φ_M in R with the modification expression F_M.

To be illuminated, φ_D and φ_M are boolean equation sets defined on the attribute set $\{ A^{QI},A^S \}$, whose normal form is $\varphi=\varphi_1\wedge\varphi_2\wedge\ldots\wedge\varphi_m$, where φ_i is an atom condition with models $(x\theta y+c)$or$(x\theta y)(x$ or y denote the attribute variable, c is a constant, $\theta\in\{=,<,\le,>,\ge\}$. F_M is an expression like $A=f(A_1,A_2,\ldots,A_k)$, where A, $A_1,A_2,\ldots,A_k$ are attributes in R, f is a computation function with the inputs $A_1,A_2,\ldots,A_k$. We use $\alpha(\varphi)$ denotes variables in φ in the following paper.

UPDATE OF K-ANONYMIZED DATASET BASING ON SEMANTIC

K-anonymity model arises k-anonymized dataset including fuzzy or generalized value in the k-anonymizaion process, which is strangling the natural update operations of k-anonymized dataset. However, the k-anonymized dataset need

to stay the same with the original dataset. So the *k*-anonymized dataset must be changed (insert, delete and modify) following the changes of original dataset. In this section, we consider how to update the generalized tuple directly according to the update operations of original dataset.

The update operation contains insert, delete and modify, in which the modify operation can also be decomposed into insert and delete. To insert or delete a generalized tuple, we need to locate the *QI* group where the generalized tuple is in or should be inserted in. We solve the problem by tuple-*QG* semantic similarity degree. After one *QI* group is updated in the *k*-anonymized dataset, it maybe smaller than *k* and violate the *k*-anonymity constraint(the candidate of each *QI* group no less than *k*). So the *QI* group will be merged with other *QI* group to maintain the *k*-anonymity constraint. In order to find the similar *QI* group which can be merged, the *QG* semantic similarity degree is presented.

The definitions of tuple-*QG* semantic similarity degree and *QG* semantic similarity degree are introduced in below. Since a tuple can be seen as a point and a *QI* group can be seen as a region composed by a set of points in the space, we measure the semantic between a original tuple and the *QI* group by the distance degree from the point to the region. Specifically, the tuple-*QG* semantic similarity degree is 1 when the point is inside of the region. For showing the semantic between two *QI* groups, we use the cosine similarity of the centers of the two regions.

Definition 6: Tuple-*QG* Semantic Similarity Degree To a tuple t in original dataset $R(A^{QI},A^{S})$ and a *QI* group QG_j in *k*-anonymized dataset $R^*(A^{QI},A^{S})$, the semantic similarity degree $T\text{-}QGSSD(t, QG_j)$ between t and QG_j is:

$$T\text{-}QGSSD(t,QG_j)=\begin{cases} 1 & t \in Range(\mathrm{QG_j}) \\ \dfrac{\sum_{l=1}^{n}|C_l - A_{jl}^{QI}|}{\sum_{l=1}^{n}|t[\mathrm{A_l^{QI}}] - A_{jl}^{QI}|} & t \notin Range(\mathrm{QG_j}) \end{cases}$$

Where $t[A_l^{QI}]$ is the value of tuple t on attribute A_l^{QI}, A_{jl}^{QI} are the centers of the range of QG_j on attribute A_l^{QI}, $\sum_{l=1}^{n}|t[\mathrm{A_l^{QI}}] - A_{jl}^{QI}|$ is the Manhattan distance of t to the center of QG_j, $\sum_{l=1}^{n}|C_l - A_{jl}^{QI}|$ is the Manhattan distance of the bound of QG_j to the center of QG_j, where the value of QG_j on the attribute A_l^{QI} is a interval $[b_l,c_l]$ and $A_{jl}=(b_{jl}+c_{jl})/2$. When $t \in Range(\mathrm{QG_j})$, it means that for each attribute A_l^{QI}, the value $t[A_l^{QI}]$of t must belong to the range value of QG_j $[A_l^{QI}]$.

Example 6: For the tuple t_1=(21,12000, flu) in R(table 1(a), the semantic similarity degree between t_1 and QG_1 in R^*(table 1(b))can be calculated as below: for 21∈[21, 25], 12000□[11k, 22k], so $T\text{-}QGSSD(t_1, QG_1)$=1. The semantic similarity degree between t_1 and QG_2 in R^* can be calculated as below: for 21∉[41, 50], 12000∉[11k, 22k], so $T\text{-}QGSSD(t_1, QG_2)=\dfrac{1}{|21-(50+41)/2|+|12-(30+21)/2|}=0.24$.

Definition 7 : *QG* Semantic Similarity Degree For two *QI* groups QG_i, QG_j in a *k*-anonymized dataset $R^*(A^{QI},A^{S})$, the semantic similarity degree between them is :

$$QGSSD(QG_i, QG_j)=\frac{\sum_{l=1}^{n} A_{il}^{QI} \bullet A_{jl}^{QI}}{\sqrt{\sum_{l=1}^{n}\left(A_{il}^{QI}\right)^2}\bullet\sqrt{\sum_{l=1}^{n}\left(A_{jl}^{QI}\right)^2}}.$$

Where, A_l^{QI} is the lth attribute on group of QG_i and QG_j, A_{il}^{QI} and A_{jl}^{QI} are the centers of the bound of QG_i and QG_j on attribute A_l^{QI}. Let the range on attribute A_l^{QI} of QG_i and QG_j are intervals $[b_{il},c_{il}]$ and $[b_{jl},c_{jl}]$, then $A_{il}=(b_{il}+c_{il})/2$, $A_{jl}=(b_{jl}+c_{jl})/2$.

Example 7: For the 2-anonymized dataset R^* (table1(b), semantic similarity degree between QG_1 and QG_2 is: $QGSSD(QG_1, QG_2)$=

$$\frac{23*45.5+15.5*25.5}{\sqrt{23^2+15.5^2}*\sqrt{45.5^2+25.5^2}}=\frac{1046.5+395.25}{\sqrt{529+240.25}*\sqrt{2070.25+625}}=$$

$$\frac{1441.75}{27.7*52.2}=0.997.$$

INSERT OPERATION

To an insert operation $INSERT(R, T)$, we need insert all the generalized tuples corresponding T to the *k*-anonymized dataset R^*. For each tuple $t \in T$, we firstly find the *QI* group whose semantic similarity degree is biggest with tuple t, then insert the generalized t^* of t to the *QI* group. If the biggest semantic similarity degree between t and QG_i is 1, then the generalized tuple t^* has the same value with QG_i on attributes A^{QI} and the same value with t on attributes A^{S}; else t^* has the generalized value of t and QG_i on attributes A^{QI} and the value of t on attributes A. In addition, when we insert many tuples into

QG_i, the size of QG_i may become very large (i.e. equal $2k$). In order to make the updated R^* satisfy k-anonymity constraints, we can divide the QG_i into two QI groups(each group size is k) according to information loss(IL) [15].

For any tuple $t \in T$, the procedure of the insert operation for the k-anonymized dataset is: 1.1 We calculate the semantic similarity between t and each QI group $QG_i(1 \le i \le m)$, if exit a QI group QG_i to make T-$QGSSD(t, QG_i)$ is 1, then insert t^* to QG_i directly, where $t^*[A^{QI}] = QG_i[A^{QI}]$, $t^*[A^S] = t[A^S]$. Otherwise, select a QI group QG_i where T-$QGSSD(t, QG_i)$ is the biggest, and let $t^*[A^{QI}] = g(t[A^{QI}], QG_i[A^{QI}])$, $t^*[A^S] = t[A^S]$. 1.2 If $|QG_i| = 2k$, then divide the QG_i into two QI groups.

Algorithm:

***INS* ($R(A^{QI}, A^S)$, T, $R^*(A^{QI}, A^S)$, *INDT*)**

Input: original dataset $R(A^{QI}, A^S)$, inserted tuples T, k-anonymized dataset $R^*(A^{QI}, A^S)$ corresponding to R, index table *INDT* between R and R^*

Output: k-anonymized dataset R^* corresponding R after insertion.

Initialization: $i = 0$; $ssd = 0$; $ssdmax = 0$;

1.for each $t \in T$

1.1 {for i =1 to m

{$ssd \leftarrow$ T-$QGSSD(t, QG_i)$;

if ssd =1 then

{$ssdmax$=1;

$QG \leftarrow QG_i$;

exit;

}

if $ssd > ssdmax$ then

{$ssdmax = ssd$;

$QG \leftarrow QG_i$;

}}

if $ssdmax$ =1 then /* $|A^{QI}|$= n */

{$R^* \leftarrow R^* \cup \{ t^* | t^*[A_j^{QI}] = QG[A_j^{QI}], t^*[A^S] = t[A^S] \}$;

update index table *INDT*;}

else

{$R^* \leftarrow R^* \cup \{ t^* | t^*[A_j^{QI}]$= the generalized value of $QG[A_j^{QI}]$ and $t[A_j^{QI}]$, $t^*[A^S] = t[A^S] \}$;

update other tuple $t^{*\prime}$in QG to same with t^* on each attribute A_j^{QI};

update index table *INDT*;}

1.2 if $|QG| = 2k$ then

{Divide QG into two QI groups QG', QG''(each group size is k) with least information loss;

update index table *INDT*;}}

2. return (R^*);

Example 8: When the insert operation *INSERT* (R, {(24,17000, insomnia), (55,62000, insomnia)})is performed in R(table 1(a)), the insert operation in R^*(table 1(b)) is: At first insert tuple t=(24,17000, insomnia), because T-$QGSSD(t, QG_1)$=2 is the biggest, so let t^*=([21, 25], [11k, 20k], insomnia) and insert into QG_1. The generalized tuple after insertion is $t3^*$ in table 2. For tuple t=(55,62000, insomnia), the semantic similarity degrees between t and each QI group are: T-$QGSSD(t, QG_1)$=0.095, T-$QGSSD(t, QG_2)$=0.2, T-$QGSSD(t, QG_3)$=0.765. Because the semantic similarity degree between t and QG_3 is the biggest, we insert t^* into QG_3. For $55 \in [51, 55]$, so $t^*[Age] = [51, 55]$; For $62000 \notin$[51k, 60k], we generalized them to [51k, 65k](i.e. $t^*[Zip]$= [51k, 65k]), $t^*[Problem]$= insomnia. $t9^*$ in table 2 is the generalized tuple after insertion. In addition, $t6^*[Zip]$ and $t7^*[Zip]$ in table 1(b) should be changed to [51k, 65k] too, which corresponding to $t7^*$, $t8^*$ in table 2.

Table 2. The increment update of insert to R^*.

Tuple	*QG*	*Age*	*Zip*	*Problem*
t1*	1	[21, 25]	[11k, 20k]	flu
t2*	1	[21, 25]	[11k, 20k]	gastritis
t3*	1	[21, 25]	[11k, 20k]	insomnia
t4*	2	[41, 50]	[21k, 30k]	flu
t5*	2	[41, 50]	[21k, 30k]	gastritis
t6*	2	[41, 50]	[21k, 30k]	insomnia
t7*	3	[51, 55]	[51k, 65k]	flu
t8*	3	[51, 55]	[51k, 65k]	gastritis
t9*	3	[51, 55]	[51k, 65k]	insomnia

DELETE OPERATION

To a delete operation (*DELETE*(*R*, φ_D)), the tuples satisfying delete condition φ_D will be deleted, so the corresponding tuples in *R** will be deleted too. The update operation of *k*-anonymized dataset *R** is: We first find the tuples satisfying the delete condition in *R*, then locate the generalized tuple in *R** of each deleted tuple and delete it. Delete operation on *k*-anonymized dataset *R**may make the size of some *QI* groups smaller than *k*, we need to check the size of each *QI* group and merge *QI* groups which size is less than *k* to maintain the *k*-anonymity constraints.

The delete operation of the *k*-anonymized dataset *R** is: we firstly find the tuple set *T* satisfying φ_D in *R*, for each tuple $t \in T$, search the generalized tuple *t** of *t* and delete *t** from *R**. Second, check each *QI* group size in *R**, if there is a QG_i less than *k*, then select another *QI* group QG_j which has the biggest semantic similarity degree with QG_i and merge with QG_i to one *QI* group.

Algorithm:

***DEL* ($R(A^{QI}, A^S)$, φ_D, $R^*(A^{QI}, A^S)$, *INDT*)**

Input : original table $R(A^{QI}, A^S)$, the delete condition φ_D of *R*, *k*-anonymized dataset $R^*(A^{QI}, A^S)$ corresponding to *R*, index table (*INDT)* between *R* and *R**.

Output: *k*-anonymized dataset *R** corresponding table R after delete operation.

Initialization: *ssd*=0; *ssdq*=0;

1. /*Locate the tuple need to be deleted, delete the corresponding tuple in *R** */

{$T \leftarrow$ {the tuples satisfying φ_D in *R*};
for *i* =1 to *m* /*delete the generalized tuple corresponding to *t* in *R** */
{ for each $t \in T$
ssd$\leftarrow$ *T-QGSSD*(*t*, QG_i);
if *ssd* =1 then /* | A^{QI} |= n */
{ *t**$\leftarrow$ { *t**|*t** $\in$ QG_i and *t**[A^S]= *t*[A^S]};
*R**$\leftarrow$ *R**- {*t**};
T$\leftarrow$*T*- *t*;
exit;
update the index table *INDT*;}}

2. /* Merge each *QI* group whose size is no less than *k* in *R** according to *QG* semantic similarity degree*/

for each $QG_i \in R^*$
if | QG_i |<*k* then
{for each $QG_j \in R^*$ and $QG_j \neq QG_i$
if *QGSSD*(QG_i, QG_j)> *ssd* then
{*ssd*$\leftarrow$*QGSSD*(QG_i, QG_j); *ssdq*=*j*;}
generalize QG_i,QG_{ssdq} and merge to one *QI* group;
update the index table *INDT*;}

3. return (*R**);

Example 9: When the delete operation in *R*(table 1(a)) is *DELETE*(*R*, (*Problem* = "insomnia")) , the deleted tuple *t*5* in *R** can be judge directly, the updated *R** was shown in table 3(a). When the delete operation in *R* is *DELETE*(*R*, (*Age*<25 □ *Zip*>15000)), the corresponding delete operation in *R** is: the tuple set satisfying φ_D=(*Age*<25 □ *Zip*>15000) is *T*={(23,18000, gastritis) }. For *T-QGSSD*((23,18000, gastritis), QG_1)=1, so the generalized tuple of (23,18000, gastritis) is *t*2*=([21, 25], [11k, 20k], gastritis), we delete *t*2* in *R**. Because |QG_1| is less than 2 after deleting *t*2*, so we need to merge QG_1 with another *QI* group. The semantic similarity degree between QG_1 and each other *QI* group QG_2 is : *QIGSSD*(QG_1, QG_2)= 0.997, *QIGSSD*(QG_1, QG_3)=0.97. Since *QIGSSD*(QG_1, QG_2) is bigger than *QIGSSD*(QG_1, QG_3), we merge QG_1 and QG_2 to one *QI* group. The updated *R** was shown is table 3(b).

Table 3. The increment update of delete to *R**

Tuple	*QG*	*Age*	*Zip*	*Problem*
t1*	1	[21, 25]	[11k, 20k]	flu
t2*	1	[21, 25]	[11k, 20k]	gastritis
t3*	2	[41, 50]	[21k, 30k]	flu
t4*	2	[41, 50]	[21k, 30k]	gastritis
t6*	3	[51, 55]	[51k, 60k]	flu
t7*	3	[51, 55]	[51k, 60k]	gastritis

(a)

Tuple	*QG*	*Age*	*Zip*	*Proble*
t1*	1	[21, 50]	[11k, 30k]	flu
t3*	1	[21, 50]	[11k, 30k]	flu
t4*	1	[21, 50]	[11k, 30k]	gastriti
t5*	1	[21, 50]	[11k, 30k]	insomn
t6*	2	[51, 55]	[51k, 60k]	flu
t7*	2	[51, 55]	[51k, 60k]	gastriti

(b)

MODIFY OPERATION

To a modify operation *MODIFY*(R, φ_M, F_M), the k-anonymized dataset R^* need to be modified correspondingly too. According to the modified value in R^* map is generalized value or precise values, the modify operation of R^* can be divided into the following two cases. First case, if φ_M and F_M only contain A^S, then we modify the tuple in R^* directly. Second, if φ_M and F_M contain attributes of quasi-identifier, the modify operation in R^* can be decomposed to delete and insert operations: For each tuple t satisfying φ_M in R, we delete it's generalizing tuple from R^* firstly, then insert the modified tuple t' to R^*. We can see that the modification procedure includes the delete operation, thus, we need to check the *QI* group after modification and merge the *QI* groups which size is less than k.

The modify operation for the k-anonymized dataset R^* includes two steps: 1. If φ_M and F_M only contain A^S, then we modify the tuple in R^* directly. Otherwise, 2. Search the tuple set T satisfying φ_M in R, for each tuple $t \in T$, we delete the generalized tuple of t from R^*, then, perform *INS* process to insert the modified tuple t' into R^*. 3. Merge *QI* groups which size is less than k.

Algorithm:

MOD ($R(A^{QI}, A^S)$, φ_M, F_M, $R^*(A^{QI}, A^S)$, *INDT*)

Input: original dataset $R(A^{QI}, A^S)$, k-anonymized dataset $R^*(A^{QI}, A^S)$ of $R(A^{QI}, A^S)$, the modification condition φ_M and the modification expression F_M for R, the index table *INDT* between R and R^*

Output: k-anonymized dataset R^* corresponding to R after modified

Initialization: ssd=0; ssdq=0;

1. if $\alpha(\varphi_M) = A^S$ and $\alpha(F_M) = A^S$ then

{$R^* \leftarrow$modify tuples in R^* satisfying φ_M basing the modification expression F_M;
Update the index table INDT;}

else

2./* Deleting the old tuple (before the modification) in $R^*$and inserting the new tuple (after the modification) */

$T \leftarrow$ {tuples satisfying φ_M on R};

/*modify each tuple $t \in T$ */

for i =1 to m

{for each $t \in T$

{$ssd \leftarrow$ T-$QGSSD(t, QG_i)$;

if ssd =1 then /* $|A^{QI}| = n$ */

{$R^* \leftarrow R^*$- {$t^* | t^* \in QG_i$ and $t^*[A^S] = t[A^S]$}; /*delete the old tuple in R^* */

update the index table *INDT*;}

/*insert the modified tuple to R^**/

$t' = F_M(t)$;

$R^* \leftarrow INS(R(A^{QI}, A^S), t', R^*(A^{QI}, A^S), INDT)$;

update the index table *INDT*;

$T \leftarrow T$- t;} }

3. /* merge *QI* group which size is less than k in R^* basing on QG semantic similarity degree */

for each $QG_i \in R^*$

if $|QG_i| < k$ then

{for each $QG_j \in R^*$ and $QG_j \neq QG_i$

if $QGSSD(QG_i, QG_j) > ssd$ then

{$ssd \leftarrow QGSSD(QG_i, QG_j)$; $ssdq = j$;}

merge QG_i and QG_{ssdq} to one *QI* group;

update the index table *INDT*;}

}

4. return (R^*);

Table 4. The increment update of modify to R^*.

Tuple	*QG*	*Zip*	*Proble*
t1*	1	[11k, 20k]	flu
t2*	1	[11k, 20k]	gastriti
t3*	2	[21k, 30k]	flu
t4*	2	[21k, 30k]	gastriti
t5*	2	[21k, 30k]	insomn
t6*	3	[51k, 65k]	flu
t7*	3	[51k, 65k]	gastriti

Example 10: When the modify operation *MODIFY* (*R*, *Age*<=25, *Age*= *Age*+10) was performed in *R*(table 1(a)), the corresponding modify operation in *R**(table 1(b)) is: Perform *Age*<=25 on *R* and got the tuple set *T*={(21,12000, flu),(23,18000, gastritis)}. For the tuple (21,12000, flu), the modified result is (31,12000, flu) according to the modification expression *Age*= *Age*+10. Because the result is different from the original tuple, we deleted it's generalized tuple ([21-25], [11k,20k], flu) from *R** and call *INS* process to insert (31,12000, flu) into QG_1. The result is shown in table 4. For the tuple(23,18000, gastritis), it's generalized tuple is ([31-35],[11k,20k], gastritis) which does not change before and after modify operation, so *R** is not changed too.

CONCLUSIONS

In this paper, we introduce a direct update method for *k*-anonymized dataset basing on semantic. The original dataset we considered is a relational table, but according to the information we have the update algorithms of *k*-anonymized dataset will be more complicated when the original dataset is a view derived from one or multiple relationship tables. So we will focus on the update method to the original dataset is a view in the following work.

ACKNOWLEDGMENT

Our work is supported by the National Natural Science Foundation of China (No.60773100, No. 61070032), Project of Science and Technology Office of Hebei Province (NO.13227427), Doctoral Fund of HeBei Normal University of Science & Technology (No.2013YB007) and Creative team Fund of HeBei Normal University of Science & Technology (No.CXTD2012-08).

REFERENCES

[1] Sweeney. L, "K-Anonymity: a model for protecting privacy", International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, **10(5)**, 557-570**(2002).**

[2] Xiao Xiaokui and Tao Yufei, "Dynamic Anonymization: Accurate Statistical Analysis with Privacy Preservation", in Proceedings of the ACM SIGMOD, 107-120**(2008)**.

[3] A.Meyerson and R.Williams, "On the complexity of optimal k-anonymity", in Proceedings of the ACM Symp, 223-228**(2004)**.

[4] G.Aggarwal, T.Feder, K.Kenthapadi, et al, "k-Anonymity:Algorithms and Hardness", Stanford University, California, USA, Tech Rep:2004-22**(2004)**.

[5] LeFevre Kristen, DeWitt. D J and Ramakrishnan Raghu, "Mondrian Multidimensional K-Anonymity", in Proceedings of ICDE, 25 **(2006)**.

[6] Ghinita G, Kalnis P and Tao Y F, "Anonymous Publication of Sensitive Transactional Data", IEEE Transactions on Knowledge and Data Engineering, **23(2)**, 161-174**(2011)**.

[7] Machanavajjhala. A, Gehrke. J and Kifer. D, "l-diversity: Privacy beyond k-anonymity", in Proceedings of ICDE, 1-12 **(2006)**.

[8] Xiao X K, Yi K and Tao Y F, "The Hardness and Approximation Algorithms for L-Diversity", in Proceedings of EDBT, 135-146**(2010)**.

[9] Liu J Q and Wang K, "On Optimal Anonymization for L+-Diversity", in Proceedings of ICDE, 213-224**(2010)**.

[10] Wang K, Xu Y B, Wong R C-W, et al., "Anonymizing Temporal Data", in Proceedings of ICDM, 1109-1114**(2010)**.

[11] Shi P, Xiong L and Fung B C M, "Anonymizing Data with Quasi-Sensitive Attribute Values", in Proceedings of CIKM, 1389-1392**(2010)**.

[12] Ren Xiangmin, Yang Jing, Zhang Jianpei and Wang Kechao, "Research on CBK(L,K)-Anonymity Algorithm", International Journal of Advancements in Computing Technology, **3(4)** , 165-173 **(2011)**.

[13] Ren Xiangmin, Yang Jing and Zhang Jianpei, "An Improved Strategy of Preventing Privacy Inference Attacks Based on K-Anonymity Data Set", International Journal of Digital Content Technology and its Applications, **4(10)**, 346-355 **(2012)**.

[14] Yinghua Liu, Bingru Yang and Guangyuan LI, "A Personalized Privacy Preserving Parallel (alpha, k)-anonymity Model", International Journal of Advancements in Computing Technology, **4(5)**, 265-271**(2012)**.

[15] Mohammed N, Fung B C M and Debbabi M, "Anonymity Meets Game Theory: Secure Data Integration with Malicious Participants", The VLDB Journal, 20(4), 567-588(2011).

[16] Xiao X and Tao Y, "M-invariance: towards privacy preserving re-publication of dynamic datasets", in Proceedings of the 2007 ACM SIGMOD international conference on Management of data, 689-700 **(2007)**.

[17] Traian M T and Alina C, "K-anonymization incremental maintenance and optimization techniques", in Proceedings of the 2007 ACM symposium on Applied computing, 80-387 **(2007)**.