Efficient and Tailor-made Anonymization for Relational and Transactional Medical Records

Tsubasa Takahashi, Koji Sabataka, Takuya Mori
Cloud System Research Laboratories, NEC Corporation, Japan
{t-takahashi@nk, k-sobataka@bx, moritaku@bx}.jp.nec.com

ABSTRACT

For health care revolution, patient records are expected to leverage in wide range of medical researches.In such leveraging, ensuring anonymity of patients by data anonymization is required to preserve privacy of them.Because most of patient records are complex data structure consisting of relational attributes (gender and birth year) and transactional attributes (diseases and drugs), existing k-anonymizations do not anonymize them efficiently.Further, in the actual medical analyses, researchers focus on their interests and objective cohorts. For the better data analyses, data anonymizations should follow their demands. In order to actualize an tailor-made anonymization with high-efficiency, this paper proposes 1) an efficient transactional recoding which concurrently discloses multiple items and 2) customizable pregeneralization to fit such demand. In our evaluation, we measure the efficiency of proposed method in several computational environments. As a result, we confirmed that proposed method can anonymize 100 million records in a reasonable time.

Categories and Subject Descriptors

H.2 [Information Systems]: Database Management

Keywords

data privacy, health care, anonymization

1. INTRODUCTION

Big data sciences are going to revolve digitized health care. Secondary use of health data can enhance health care experiences for individuals, expand knowledge about disease and appropriate treatments, strengthen understanding about effectiveness and efficiency of health care systems[12]. Thus, the revolution of public health care makes our life more smart and wealthy. However, in such data leveraging, patients' privacy should be taken care. Medical records such as patient records and medical receipts have some sensitive information. Therefore, ensuring anonymity of patients before leveraging medical records has been required.

Table 1: Complex Data

id	*birthyear*	*gender*	*diseases*	*drugs*
1	1970	Male	k, l, m	a, b, z
2	1971	Male	k, l, n	a, z
3	1989	Female	n, o	c, x, y
4	1982	Female	m, o	b, x, y

Table 2: 2-anonymous Complex Data

birthyear	*gender*	*diseases*	*drugs*
1970–1979	Male	k, l, M	a, z
1970–1979	Male	k, l, M	a, z
1980–1989	Female	M	x, y
1980–1989	Female	M	x, y

Patient records contains patient's characteristics and medical conditions (birthyear and gender, diseases and drugs are examples respectively in Table 1). Such records forms a complex data which has relational attributes and transactional attributes. By "complex data" we mean in which a tuple has the form $t_i = (a_{i1}, \ldots, a_{id})$, where t_i is the i-th tuple and a_{ij} be a value of i-th tuple at j-th attribute A_j in dataset T. In a relational attribute, every a_{ij} is the value whose size is 1. While, in a transactional attribute, every a_{ij} is the value whose size is more than 1 and forms $a_{ij} = \{item_1, \ldots, item_m\}$.

In order to publish patient records, k-anonymization is well known for a privacy protection technology. k-anonymity [1] has been proposed to reduce the risk of linking individuals for relational data (i.e. data having multidimensional single-valued attributes). k-anonymity assumes quasi-identifiers (QIs) which is possible to distinguish individuals from the dataset. k-anonymization makes duplicates for the combination of all QIs (Table 2). The goal of k-anonymization is to make released data with remaining utility as much as possible by recoding attribute values. However, existing k-anonymization researches mainly focus on either relational attributes or a transactional attribute. Furthermore, by the k-anonymization, the utility of patient records are degraded. Actually, health care researches focus on specific interests and objective cohorts. Thus, in order to actualize good public health care researches by utilizing patient records, anonymization should captures the demands of them.

KDD-DSSG, August 24, 2014, New York, NY, USA.
Copyright 2014 ACM X-XXXXX-XX-X/XX/XX ...$15.00.

This paper tackles a problem of tailor-made anonymization for large scale complex health care records. This paper proposes a complex data anonymization with an efficient transactional recoding and a customizable pre-order generalization. The contributions of this research are that 1) we propose a complex data anonymization having an efficient specialization for transactional attribute which can be transparently used with specializations for relational attributes, 2) we introduce pre-order generalization which generalizes attribute values following users' description and 3) we confirmed that the proposed method can anonymize medical receipts having over 100 million records in a half of day.

The rest of this paper is organized as follows. Section 2 describes related works. Section 3 describes a top-down anonymization framework for complex data. Section 4 introduces a pre-generalization. Section 5 proposes our recoding algorithm for transactional attributes in complex data. Section 6 presents experiments, their results and discussion about them. Finally, section 7 concludes this paper.

2. RELATED WORK

Data anonymization for health care data have been proposed, but most of them are anonymization only for relational attributes [10][11].

For anonymization of transactional data, several techniques have been proposed. The k^m-anonymity [5] requires that every subset of no more than m items is contained in at least k tuples. Let k^∞-anonymity denote k^m-anonymity with m being the longest itemset size in the dataset. The k^∞-anonymity is the same as the k-anonymity. The existing works in [5][6][8] obfuscate items using generalization hierarchies of items. Further, the existing works [5][6][7][8] assume that QIs are only a transactional attribute.

Data anonymization for complex data differs from the problem studied by most previous work, in which the implicit assumption is that each individual is associated with only a transactional attribute or relational attributes.

Therefore, each technique does not apply to complex data appropriately. For complex data, the existing work [9] proposed an anonymization by transparent top-down specialization for relational and transactional attributes. However, it is not efficient for large scale data.

3. K-ANONYMIZATION FRAMEWORK

This paper focuses on the anonymization for relational and transactional data. In the multi-dimensional case, top-down approach [2][4] is an efficient heuristic framework. Top-down approach starts with a general condition of all attributes, and specialize the condition so as to increase utility but reduce the anonymity.

Based on the existing top-down approach, we assemble a top-down k-anonymization for relational and transactional attributes. The top-down method for such multi-dimensional data is described in Algorithm 1. This algorithm abstracts multi-dimensional specialization for heterogeneous attributes. In this paper, the algorithm treats only both relational and transactional attribute in the common way. Especially, at the specialization phase, any specializations with/without generalization hierarchy are applicable. Moreover, at the choosing_dimension phase, in order to evenly decide the target attribute, the approach uses NCP[3], which is the normalized information loss measure for all attributes.

Algorithm 1 Top Down Anonymization(T)

```
$T^* \leftarrow$ initialize($T$)
while $|QI^*| \geq 1$ do
    $T^*_{tmp} \leftarrow T^*$
    $A_{div} \leftarrow choose\_dimension(QI^*)$
    $T^* \leftarrow specialize(A_{div})$
    if $T^*$ is not $k$-anonymous then
        $T^* \leftarrow T^*_{tmp}$
        $QI^* \leftarrow QI^* \backslash \{A_{div}\}$
    end if
end while
```

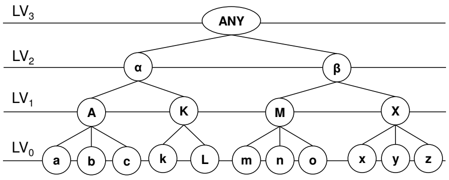


Figure 1: Generalization Hierarchy

Let equivalent class be a set of tuples which are duplicated over all attributes. If all equivalent classes are k-duplicated, all of them are k-anonymous. In this framework, all values are generalized into the top most generalized values and all tuples become members of the same equivalent class initially. At the specialization phase, equivalent classes are separated along with the specialization.

In the experiment part, we employ a specialization following a generalization hierarchy for a relational attribute. At a specialization phase, the specialization turns every generalized value of the selected attribute into specialized value which is a child of the generalized value by following its hierarchy.

4. TAILOR-MADE PRE-GENERALIZATION

In order to tune the anonymized data in to the demand of data usage, we introduce tailor-made pre-generalizations for relational attributes and transactional attributes. We propose a tailor-made anonymization which 1) first pre-generalizes attribute values into the generalized values designated by the user (pre-generailzation phase) and then 2) anonymizes attribute values to ensure k-anonymity by the top-down anonymization framework.

To input the demand of data usage, for all attributes the user is required to describes the generalization hierarchies and specify generalization levels for attributes. All pre-generalizations follows generalization hierarchies and generalization level.

The hierarchy has a root, nodes and leaves. The root, nodes and leaves are the highest generalized value, generalized value and original values of the attribute respectively. We introduce generalization level which represents degree of generalization. We define the generalization level of leaves are 0 and the parents of leaves are 1. While, for nodes, the generalization level is 1+ the maximum generalization level in their children.

Only for the attribute which is designated generalization level by the user, the values of the attribute are generalized into the designated values.

4.1 Pre-Generalization Description

In the most data analyses, values of objective attributes should be more carefully generalized because accuracy of the attribute values affect the accuracy of objective analysis. We introduce a precise and easy way of the description about the ways of pre-generalization.

Pre-generalization description PGD is a description about the ways of pre-generalization. For attribute A, let $PGD[A]$ be a pre-generalization description about overall generalization level of attribute A, $PGD[A]$ can be set a generalization level. Let $PGD[A, \ell]$ be a pre-generalization description attribute A at generalizaton level ℓ, $PGD[A, \ell]$ can be set a set of values in A. For any ℓ_1 and ℓ_2 ($\ell_1 \neq \ell_2$), $PGD[A, \ell_1] \cap PGD[A, \ell_2] = \emptyset$. $PGD[A, \ell]$ is always given priority over $PGD[A]$. If for any ℓ, $PGD[A, \ell]$ is not designated, all values are only generalized into the values at $PGD[A]$. If $PGD[A]$ is not designated, it means $PGD[A] = 0$. An example of the PGDs about attribute 'diseases' is as follows:

$$PGD[diseases] = 1$$
$$PGD[diseases, 0] = \{k, l\}$$
$$PGD[diseases, 2] = \{x, y, z\}$$

By the above PGD and following the hierarchy in Figure 1, k and l are not generalized, x, y and z are generalized into β and the others are generalized into the value at generalization level 1.

5. EFFICIENT ITEM SUPPRESSION

We propose efficient local item suppression for transactional attributes. This paper takes a stance of that generalizations of items are tasks for the pre-generalizations.

Based on this top-down anonymization, local item suppression for transactional attribute is that first all items are hidden, then each iteration items are disclosed if this disclosure does not break k-anonymity.

Existing method [9] requires a lot of round of specialization because it discloses an item at a round. In the ideal case, the minimum number of round is equal to the number of unique items in the attribute.

In order to actualize an efficient local item suppression, we propose a method which can concurrently disclose an item from every tuple at a round. The method first measures utility of all hidden items and ranks them in the each equivalent class. Second, from every tuple, it selects the highest ranked item for the candidate of disclose. Based on the candidate, it discloses the items if those items do not violate k-anonymity.

5.1 Ranking Items

To disclose items having high utility, the proposed method first evaluate all hidden items. The set of all items which can be disclosed in class c is described as $Cand(c)$. This evaluation, in each equivalent class, creates rankings based on the utility measure. In this paper, we utilize the frequency of item in the class for this utility measure. The frequency of item α in $Cand(c)$ is defined as follow:

$$f_c(\alpha) = |\{t \in c | \alpha \in t.A\}| \quad (1)$$

After the item evaluation, the ranking is created in descending order of the frequency in each class. In this ranking, the higher ranked items, the better utility the disclosure makes. Further, the item whose frequency is less than k is removed from the ranking. The rank of item α in class c is described as $rank_c(\alpha)$.

Table 3: Concurrent Specialization

(a) Orig. *drugs*	(b) Init. *drugs*	(c) 1st *drugs*	(d) 2nd *drugs*
a, b, z	*	a	a, z
a, z	*	a	a, z
c, x, y	*	x	x, y
b, x, y	*	x	x, y

5.2 Candidate Selection

Second, from every tuple t, it selects a pair $(t, \alpha^*(t))$, where $\alpha^*(t)$ is the highest ranked item. For items in all pairs in the class, we measure the frequency of them as follows:

$$f_c^*(\alpha^*) = |\{(t, \alpha^*(t)) | t \in c, \alpha^*(t) = \alpha^*\}| \quad (2)$$

Based on this frequency, we select $P(c) = \{\alpha^* | f_c^*(\alpha^*) \geq k\}$ which is the pivot of specialization.

5.3 Concurrent Specialization

At last, based on the $P(c)$, the proposed method specializes a class into multiple classes. In this specialization, all tuples having any item in $P(c)$ can disclose an item. Based on the equivalent of the disclosed items and nothing disclosed, the class is separated into multiple classes. Further, the specializations for all classes are concurrently executed (Table 3).

In this specialization, tuples which is nothing disclosed may be violated k-anonymity. In case of such violation, these tuples are suppressed if trash bin has a room. If trash bin has no room, the specialization of this class is canceled.

After such cancellation, it removes the item which led to the violation from $Cand(c)$ and backs to Ranking Items only classes which k-anonymity are violated. To ensure k-anonymous specializations, items whose $f_c^*() \geq |c| - k$ are removed.

In the ideal case, the minimum number of executions of this specialization is the maximum size of transaction in all tuples. In the most of cases, the maximum transaction size is smaller than number of unique items. Thus, this proposed method can efficiently anonymize transactional attributes than existing method. Further, by the pre-generalization, the size of transactions can be smaller and it makes anonymization faster.

6. EVALUATION

We evaluate our approach in terms of efficiency and utility. For evaluation of efficiency, we measure execution time of the proposed method and the existing method and compare the both. We call the proposed method CID (Concurrent Item Disclosure), and the existing method IR (Itemset Recoding). For utility, we measure the information remain which indicates how accurate anonymized data is. The value of the information remain is calculated as $(1 - NCP(A)) \times 100[\%]$. All k-anonymizations in this section are used k=10 for k-anonymity.

For experiments, we used a medical receipt data[1] and an artificial data. We use 3 relational attributes (birthyear, gender, treatdate) and 2 transactional attributes (diseases, drugs) from medical receipt data. The medical receipt data contains 4 million records. In this experimental evaluation, we used 1 million records which are randomly selected from the original one. The artificial data is randomly generated from medical receipt data for scalability analysis. The artificial data contains 150 million records with 6 relational attributes and 2 relational attributes.

Execution times are measured on a computer with 8 cores processors (2.3GHz) and 64GB memory running CentOS 6.5. We developed the proposed method by Java 1.7.0_51 and PostgreSQL 9.2.1. We also developed anonymization system on NEC's InfoFrame DWH Appliance (IDA)[2] which is a high performance data warehouse appliance.

6.1 Efficiency

This section measures execution times of anonymization methods for comparing efficiencies.

First, we compare the execution times of CID and IR for medical receipts having only a transactional attribute on PostgreSQL. For recoding relational attributes, both methods utilize the same way. Figure 2 shows the execution times. The proposed method CID shows around 3 times faster than IR for 1 million records. Because the proposed method shows faster than the existing method all the data, we can say that the proposed method can improve efficiency.

Next, we measure the execution times with employing IDA for big data anonymization. Figure 3 shows the execution times for 1, 10, 20, 50, 100 and 150 million records. The expression $RiTj$ in Figure 3 means that the target table contains i relational attributes and j transactional attributes.

For 1 million records, the proposed method can finish k-anonymization in several minutes. If the volume of database is increased, the execution time is linearly-increasing. Further, we confirmed that the proposed method can anonymize huge medical receipts having over 100 million records in a half of day. Consider medical receipts of all citizen in a county are generated each month, in the most case, out proposed method can anonymize them in a reasonable time.

6.2 Utility

Subsequently, we measure the utility of anonymized data. Figure 4 shows the values of information remains of each attribute in following settings; R3T1a and R3T1b have a different transactional attribute each other, and R3T2 has the both, R'3T1a, R'3T1b and R'3T2 are anonymized with designating generalization levels for all relational attributes. In this evaluation, we set these generalization levels are 1.

The non-designated anonymized data are extremely obfuscated and the utilities among attributes turns unbalanced. Because degree of specialization is differ in each attribute, balancing and optimizing utility of anonymized data is difficult. The balancing is one of future works. However, if the user have some demands of data usages, the demands can properly affected. Further, the total information remains are higher than non-designated ones. Thus, we can say that our proposed pre-generalization can be reflect user's demand.

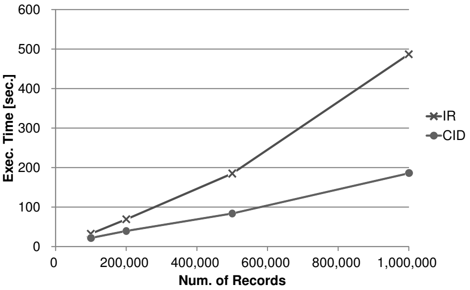


Figure 2: Execution Time (PostgreSQL)

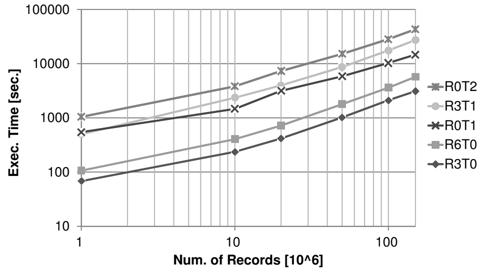
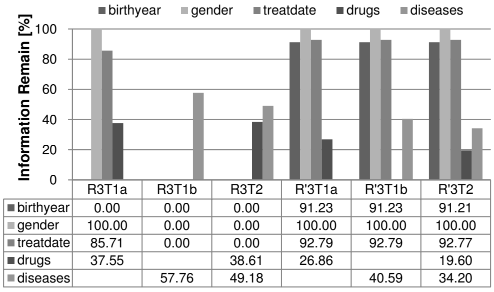


Figure 3: Execution Time (IDA)



	R3T1a	R3T1b	R3T2	R'3T1a	R'3T1b	R'3T2
■ birthyear	0.00	0.00	0.00	91.23	91.23	91.21
■ gender	100.00	0.00	0.00	100.00	100.00	100.00
■ treatdate	85.71	0.00	0.00	92.79	92.79	92.77
■ drugs	37.55		38.61	26.86		19.60
■ diseases		57.76	49.18		40.59	34.20

Figure 4: Utility of Anonymized Data

7. CONCLUSIONS

In this paper, we tackled the problem of anonymization for relational and transactional data. We introduced customizable pre-generalization based on the user's demand of data utilizations. We also proposed an efficient transactional recoding for large scale relational and transactional data. In the experimental part, we showed the data qualities and efficiency of the proposed method. The efficiency of the proposed method over came existing method. Furthermore, we confirmed that the proposed method can anonymize enormous medical receipts having over 100 million records in a half of day. Consider patient records of all citizen in a county are generated each month, in the most case, our proposed method can anonymize them in a reasonable time. In our future work, we refine the core algorithm of anonymization for complex data to improve utilities of anonymized data automatically.

ACKNOWLEDGMENTS

The authors thank Takahide Kohro for utilizing medical receipt data set.

[1] http://www.jmdc.co.jp/

[2] http://www.nec.com/en/global/techrep/journal/g12/n02/pdf/120203.pdf

8. REFERENCES

[1] Sweeney, L.: k-anonymity: a model for protecting privacy. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10(5), pp. 555–570 (2002).

[2] Fung, B.C.M., Wang, K. and Yu, P.S.: Top-down specialization for information and privacy preservation. Proc. $ICDE$2005, pp. 205–216 (2005).

[3] Xu, J., Wang, W., Pei, J., Wang, X., Shi, B. and Fu, A.: Utility-based anonymization using local recoding. Proc. $SIGKDD$2006, pp.785–790 (2006).

[4] LeFevre, K., DeWitt, D.J., and Ramakrishnan, R.: Mondrian multidimensional k-anonymity. Proc. $ICDE$2006 (2006).

[5] Terrovits, M. Mamoulis, N. and Kalnis, P.: Privacy Preserving Anonymization of Set-valued Data. Proc. $VLDB$2008 (2008).

[6] He, Y. and Naughton, F.: Anonymization of set-valued data via top-down, local generalization. Proc. $VLDB$2009 (2009).

[7] Xu, Y., Wang, K., Fu, A. and Yu, P. S.: Anonymizing Transaction Databases for Publication. Proc. KDD2008 (2008).

[8] Liu, J. and Wang, K.: Anonymizing Transaction Data by Integrating Suppression and Generalization. Proc. $PAKDD$2010 (2010).

[9] Takahashi, T., Sobataka, K., Takenouchi, T., Toyoda, Y. and Mori, T.: Top-down Itemset Recoding for Releasing Private Complex Data. Proc. PST2013 (2013).

[10] Mohammed, N., Fung, B., Hung, P. C., and Lee, C. K.: Anonymizing healthcare data: a case study on the blood transfusion service. Proc. $SIGKDD$2009, pp. 1285–1294 (2009).

[11] Mohammed, N., Fung, B., Hung, P. C., and Lee, C. K.: Centralized and distributed anonymization for high-dimensional healthcare data. ACM Transactions on Knowledge Discovery from Data (TKDD), 4(4), 18 (2010).

[12] Safran, C., Bloomrosen, M., Hammond, W. E., Labkoff, S., Markel-Fox, S., Tang, P. C., and Detmer, D. E. Toward a national framework for the secondary use of health data: an American Medical Informatics Association White Paper. Journal of the American Medical Informatics Association, 14(1), 1-9 (2007).