Clustering-based approach for achieving k-anonymization

Xiaoshuang Xu † Masayuki Numao ‡

† The University of Electro-Communications, 1-5-1 Chofugaoka, Chofu, Tokyo 182-8585

‡ The University of Electro-Communications, 1-5-1 Chofugaoka, Chofu, Tokyo 182-8585

E-mail: † xuxiaoshuang1988@gmail.com, ‡ numao@cs.uec.ac.jp

Abstract This paper proposed an efficient generalized clustering method which derives from the ***k***-means algorithm for achieving ***k***-anonymization with good data quality and minimum information loss. We defined the distance function for the three major attribute types: numerical type, categorical type, and structural type. Then we proceeded the method in two stages: preprocessing stage and postprocessing stage. The preprocessing stage is to partitions all records into $\left\lfloor \frac{n}{k} \right\rfloor$ groups, and then add the records that are naturally similar to each other into every group. The postprocessing stage is to add each remaining record into a cluster with respect to which the increment of the information loss is minimal. We experimentally compared our method with other two clustering-based ***k***-anonymization methods. The experiment showed that our method outperforms their method and also ensures the anonymization of data.

Keyword ***k***-anonymization, ***k***-means, distance function, information loss

1. Introduction

In recent years, the privacy leakage has now become one of the important major concerns. Many companies and organizations have collected and stored a large amount of information and operational data. The data normally contains a lot of personal details and sensitive information such as name, birthdate, address, e-mail and disease etc. In general, companies and organizations need to do some privacy-preserving techniques when publishing data. For example, a set of records with person's basic information may be released by an organization to facilitate useful data analysis or research. Records in Table I is an example of person's basic information records collected by a company. The company must ensure that no other organization can infer the personal information or even identify an individual since it will cause very serious impact on society. The process of protecting such a table is to remove all the explicit identifiers, such as name from the table. However, even though a table is free of explicit identifiers, some of the remaining attributes in combination could be specific enough to identify individuals. For example, the combination of {Gender, Region, Age} can be used to identify an individual and has been called a quasi-identifier in literatures. If an organization has the background knowledge about the person of ID A0002, that is: Gender is Male, Age is 38, and Region is Tokyo, he can accurately infer the person of ID A0002`s Disease, that is Bronchitis.

To prevent privacy leakage or against identifying individuals, a simple and practical privacy-preserving ***k*-anonymization** [1, 2, 3] was proposed. It requires that each record in a table is indistinguishable from at least $k-1$ other records [1]. Therefore, privacy related information cannot be revealed from k-anonymity protected table. For the past years k-anonymization has been extensively studied especially one of approaches called **generalization** [2, 5, 6, 10] which have successful achieved the privacy protection [9, 10, 11]. Table Ⅱ is a 3-anonymization version by using generalization approach. In table Ⅱ, the data values of Table Ⅰ in attributes Gender, Region and Age has been generalized as common values(for instance, the age of the person ID A0002 has been generalized to [30,40]) and the number of records in its two **equivalence classes** are both equal to three. As a result, given Table Ⅱ, even if an organization has the quasi-identifier values about the person of ID A0002, they cannot exactly identify the record of ID A0002 from the first equivalence classes. The purpose of data privacy preservation is then achieved.

TABLE I. PERSONAL INFORMATION IN A COMPANY

ID	*Gender*	*Region*	*Age*	*Disease*
A0001	Male	Kanagawa	40	Flu
A0002	Male	Tokyo	38	Bronchitis
A0003	Female	Chiba	37	Bronchitis
A0004	Male	Kyoto	28	Cancer
A0005	Female	Kobe	25	Flu
A0006	Female	Nara	23	Cancer

TABLE II. 3-ANONYMIZATION TABLE

G-ID	*Gender*	*Region*	*Age*	*Disease*
1	Male	Kantou	[30,40]	Flu
1	Male	Kantou	[30,40]	Bronchitis
1	Male	Kantou	[30,40]	Bronchitis
2	Female	Kansai	[20,30]	Cancer
2	Female	Kansai	[20,30]	Flu
2	Female	Kansai	[20,30]	Cancer

Although the generalization based k-anonymization have successfully achieved the privacy preservation, there is a serious issue on that it decreases data quality a lot after anonymized. To ensure data mining performance, utility should be taken into account. One of the direct measures of the utility of the generalized data is **information loss** [4]. Generally speaking, the less the information loss in the k-anonymity protected table makes, the larger the table usability is. Clustering [11] is a method commonly used to automatically partition a data set into many groups, which aims at grouping a set of records into clusters so that records in a cluster are similar to each other and are different from records in other clusters. In the clustering-based k-anonymity protected table, if the records that will be assembled as an equivalence class are more similar to each other with respect to an attribute set, it reduces the much more information loss for generalizing the equivalence class.

In this paper we proposed an efficient clustering based k-anonymization to ensure good data quality with minimum information loss and to realize the good performance. At the heart of every clustering problem are the distance functions that measure the dissimilarities among data points. The distance functions are usually determined by the type of data. As the data in the k-anonymity problem are always person-specific records that typically consists of the three major types: numerical type (such as age), categorical type (such as nationality), we need a distance function that can handle both types of data at the same time. This motivated us to define the distance function for the three major type respectively irrespective of the background for attributes. The distance function without regard to the background of data type is calculated by the numbers of each attribute to be generalized and the numbers of generalized attribute, which has a meaning of general purpose for any kind of data and makes the clustering calculate very easy and fast as well. Our clustering approach was based on the k-means clustering which will produce some clusters that do not satisfy the k size condition. Therefore, we proceed the partitions in two stages: preprocessing stage and postprocessing stage. The preprocessing stage is to partitions all records into $\lfloor \frac{n}{k} \rfloor$ groups, and then add the records in its nearest group. After the preprocessing stage, a set of clusters are constructed, but some of records might remain. The postprocessing stage is to adjust the remaining records, which adds each record into a cluster with respect to that the increment of the information loss is minimal.

The rest of this paper is organized as follows. Section 2 reviews the basic concepts of k–anonymization. In section 3, we talked about the exiting technique on this field. In section 4, we present the details of our generalization algorithm. In section 5, we experimentally evaluate the efficiency and effectiveness of our algorithm. Finally, the paper is concluded in Section 6.

2. Fundamental definitions

In this section, we give the definitions of basic concept about k-anonymization. Let T denote a microdata table like Table Ⅰ that contains the private information of a set of individuals and has privacy-related attributes.

2.1. k-anonymization table

Definition 1 (Quasi-identifier Attribute Set). A quasi-identifier attribute set QI = $\{A_1, A_2, \dots, A_d\} \subseteq \{A_1, A_2, \dots, A_n\}$ is a set of attributes in a table which can possibly be joined with other tables in order to reveal the personal identity of individual records. For example, attribute set {Gender, Region, Age} in Table I is a quasi-identifier set. If the table is joined with other tables, it may reveal more information of personal information.

Definition 2 (equivalence class). An equivalence class of a table with respect to a quasi-identifier attribute set is the set of all records in the table containing identical values for the quasi-identifier attribute set. For example, records 1 and 2 in Table Ⅱ form an equivalence class with respect to attribute set {Gender, Region, Age}. Their corresponding values are identical.

Definition 3 (k-anonymity Property). A table is k-anonymous with respect to a quasi-identifier set if the size of every equivalence class is k or more. k-anonymity requires that each record in a table with respect to a quasi-identifier set is indistinguishable from at least $k-1$ other records [3]. For example, Table Ⅱ satisfies 3-anonymity property since equivalence class {Male, Kantou, [30, 40]} and {Female, Kansai, [20, 30]} occur three times.

Definition 4 (k-anonymization). A table is said to be a k-anonymization of the table if the table after anonymized

satisfies the k-anonymity property with respect to the quasi-identifier set. For example, Table Ⅱ is a table of k-anonymization.

2.2. Information loss

k-anonymization by generalization or any other way usually causes information loss. The idea of information loss is used to measure the amount of information loss due to k-anonymization. There are various methods of devoting information loss. The measurement in this paper is based on the description given by Byun et al [4]. Please also refer to Byun et al. for more details.

Let T denote a set of records with m numeric quasi-identifiers $N_1, N_2, \ldots, N_m$ and s categorical quasi-identifiers $C_1, C_2, \ldots, C_s$. Let $P = \{P_1, P_2, \ldots, P_p\}$ be a partitioning of T, such that $\bigcup_{i=1}^{p} P_i = T$, and $P_i \cap P_j = \emptyset$ for any $i \neq j$. To generalize the values of each categorical attribute C_i (i = 1, 2,…,S), let T_{c_i} be the taxonomy tree defined for the domain of C_i.

Consider a cluster P in T which consists of some numerical and categorical attributes. Let $N_{i_{max}}$, $N_{i_{min}}$ devote the max and min values of the records in P and $T_{N i_{max}}$, $T_{N i_{min}}$ be the max and min values of the records in T with respect to numeric attribute N_i (i = 1, 2,..., m). Also let $\cup c_i$ devote the union set of values in P with respect to the categorical attribute C_i (i = 1, 2,…,S). Then the amount of information loss due to generalizing P, denoted by $IL_{(p)}$ is defined as:

$$\boldsymbol{IL_{(p)}} = |\mathbf{e}|.\left(\sum_{\mathbf{i=1}}^{\boldsymbol{m}} \frac{N_{i_{max}} - N_{i_{min}}}{T_{N i_{max}} - T_{N i_{min}}} + \sum_{\mathbf{i=1}}^{\boldsymbol{p}} \frac{H(\wedge(\cup c_i))}{H(T_{c_i})}\right) \quad \text{(1)}$$

where $|e|$ is the number of records in cluster P. $\wedge(\cup c_i)$ is the subtree rooted at the lowest common ancestor of every value in $\cup c_i$. And $H(T_{c_i})$ is the height of taxonomy tree T_{c_i}.

Consider that the total number of records in T is partitioned into P clusters, namely $P = \{P_1, P_2, \ldots, P_p\}$. The total information loss of T is the sum of the information loss of each P_1 (i = 1, 2,…,p). Therefore the total information loss will be:

$$\boldsymbol{IL_{(T)}} = \sum_{\mathbf{i=1}}^{\mathbf{p}} |\mathbf{e}|.\left(\sum_{\mathbf{i=1}}^{\boldsymbol{m}} \frac{N_{i_{max}} - N_{i_{min}}}{T_{N i_{max}} - T_{N i_{min}}} + \sum_{\mathbf{i=1}}^{\boldsymbol{p}} \frac{H(\wedge(\cup c_i))}{H(T_{c_i})}\right) \quad \text{(2)}$$

3. Existing Techniques

In the section of introduction, we talked about k-anonymization achieved by the generalization approach, which replaces real values with less specific but semantically consistent values. Typically, numeric values are generalized into intervals (e.g., [30−40]), and categorical values are generalized into a set of distinct values (e.g., {Japan, China}) or a single value that represents such a set (e.g., Asia).

3.1. Generalization based k-anonymization

Recently, many generalization approaches have been proposed. Mainly there are two ways to achieve k-anonymity, namely domain generalization and local generalization. The domain generalization happens at the domain level. If a lower level domain needs to be generalized to a higher level domain, all the values in the lower domain are generalized to the higher domain. This restriction could be a significant drawback in that it may lead to relatively high data distortion due to unnecessary generalization. A local generalization method generalizes attribute values at cell level. This method allows values from different domain levels to be combined to represent a generalization and hence may minimize the distortion of an anonymous table. Sweeney proposed MinGen [1,2] algorithm but it is impractical and another DataFly [2] is a global recoding algorithm. A lot of papers conclude that optimal k-anonymization is NP-hard. The objectives of ***k***-anonymization is to modify a table to satisfy the ***k***-anonymity property, and to minimize the distortion from its original table after anonymized.

3.2. Clustering based *k*-anonymization

Clustering is the problem of partitioning a set of records into groups such that records in the same group are more similar to each other than records in other groups with respect to some defined similarity criteria. In the k-anonymity protected table, if the records are more similar to each other with respect to an attribute set, it reduces the more information loss for generalizing the equivalence class. That is the reason why the k-anonymity model can be addressed from the viewpoint of clustering.

Byun et al. [4] proposed the greedy k-member clustering algorithm. This algorithm works by first randomly selecting a record r as the seed to start building a cluster, and subsequently selecting and adding more records to the cluster such that the added records incur the least information loss within the cluster. Once the number of records in this cluster reaches k, this algorithm selects a new record that is the furthest from r, and repeats the same process to build the next cluster. When there are fewer than k records not assigned to any cluster yet, this algorithm then individually assigns these records to their closest clusters. This algorithm has two drawbacks. First, it is slow. Second, it is sensitive to outliers. To build a new cluster, this algorithm chooses a new record that is the furthest from the first record selected for the previous cluster. If the

data contains outliers, it is likely that outliers have a great chance of being selected. If a cluster contains outliers, the information loss of this cluster increases. Their experimental results showed that the k-member algorithm causes significantly less information loss than another k-anonymization technique called "Mondrian" proposed by LeFevre et al.

To reduce the information loss and execution time recently Lin and Wei [8] proposed an efficient one-pass k-mean clustering problem. They showed that their algorithm performs better than the proposed algorithm of Byun et al. with respect to both execution time and information loss. Like Chiu and Tsai's algorithm, this algorithm forms all clusters at a time. According to their methods first sort all records by their quasi-identifiers, then determine approximate number of clusters, by $\left\lfloor \frac{n}{k} \right\rfloor$, where k is the cluster size. Then randomly select P records as seeds to build P clusters. For each record r the algorithm finds the cluster that is closest to$r$, assigns r to that cluster and subsequently updates the center point. Finally, if some clusters contain more than k records remove excess records from those clusters that are dissimilar to most of the records and then add these records to other similar clusters (whose size is less than k). Although this method has less execution time there is still a chance of being affected by extreme values. Again if this algorithm first selects p records that come from the same equivalent class then the total information loss will be higher.

4. Clustering based approach

In this section, we will present the details of our efficient generalized clustering method based k-anonymization. The objective of our algorithm is to modify a table which contains a variety of attribute types to satisfy the k-anonymity property, and to minimize the information loss from its original table after anonymized. Considering the data in the k-anonymity problem are always person-specific records and the attribute type in the data are nothing more than three major types: numerical type (such as age), categorical type (such as nationality, occupation). To apple the clustering approach, it is necessary to define the distance that measures the dissimilarities among data points. In this paper we defined the distance function for three major attribute irrespective of the background for attributes. The key of our distance function is to consider the numbers of each attribute to be generalized and the numbers of generalized attribute. And then give them the appropriate weight.

4.1. Distance function

Definition 5 (numeric values) About the values of numeric attribute type, we note that there are three kinds of values in the data. First one is continuous value like age (such as 35, 36, 37,…, etc.). Second one is 2-values type like gender. If the male is set as 1, the female can only be 0. The last one is hierarchical data like education. The value are of hierarchy such as grade of 1st-4th, 5th-6th, 7th-8th,9th,10th,11th,12th,...etc. We convert these hierarchical values into hierarchical numeric values: 0, 1,2,3,4,5,6,7,…,etc.

Let S_{N_i} be the standard deviation of the records in T with respect to numeric attribute N_i (i = 1, 2,…, m). Then the distance between numeric values N_1, N_2, devoted as:

$$\boldsymbol{NN_i(N_1, N_2)} = \frac{|R_{N_i}|-1}{|N_i|-1} \cdot \frac{N_1 - N_2}{S_{N_i}} \quad \textbf{(3)}$$

$|N_i|$is the numbers of numeric attribute to be generalized and $|R_{N_i}|$ is the numbers of generalized numeric attribute.

Definition 6 (categorical values). We specify the attribute such as nationality or occupation as the categorical type. Let C_i denotes S categorical quasi-identifiers$C_1, C_2, \dots, C_s$. The normalized distance between two categorical values v_i, $v_j \in N$ is defined as:

$$\boldsymbol{NC_i(v_1, v_2)} = \frac{|R_{C_i}|-1}{|C_i|-1} \quad \textbf{(4)}$$

$|C_i|$is the numbers of categorical attribute to be generalized and $|R_{C_i}|$ is the numbers of generalized categorical attribute.

Definition 7 (Distance between two records) Let T denote a set of records with m numeric quasi-identifiers $N_1, N_2, \dots, N_m$, s ccategorical quasi-identifiers $C_1, C_2, \dots, C_s$. We combine the accurate distance for clustering: Euclidean distance. Hence, the distance of two records $r_1, r_2 \in T$ is defined as:

$$\Delta(\boldsymbol{r_1, r_2}) = \sqrt{\sum_{i=1}^{m}(\boldsymbol{r_1[NN_i] - r_2[NN_i]})^2 + \sum_{j=1}^{s}(\boldsymbol{r_1[NC_i] - r_2[NC_i]})^2} \quad \textbf{(5)}$$

```
Input: a set T of n records and the value k
Output: a set of clusters P = {P_1, P_2, ..., P_p}
------------------The preprocessing stage----------
1. if( | n | ≤ k )
2.        return n;
3. end if;
4. Let P = ⌊n/k⌋;
5. result = ∅;
6. randomly select P distinct records r_1, r_2, ..., r_p;
7. for i= 1 to P
8.          Let P_i = {r_i} ;
9.          T = T - {r_i};
10. While (T ≠ ∅)
11.         r =randomly picked record from T ;
12.         Calculate the distance Δ (r, P_i) ;
13.         Add r to its closest P_i;
14.         update centroid of P_i;
15.         T = T - {r};
16. End While
----------------The postprocessing stage-----------
17. If there is Pwhich the size is < k then
18.         While (| P | > k)
19.          sort all clusters by the information
loss;
20.         find the cluster that information loss is
maximal
21.         remove r which is farthest from
centroid ofthe cluster;
22. End while
23. Add the record into the cluster with respect to
whichthe increment of the information loss is
minimal.
24. Else
25.      Return result;
26. End if
```

Fig. 1. *Clustering based approach*

4.2. Clustering based approach

Now we discuss our efficient generalized clustering algorithm based k-anonymization. This algorithm is based on the k-means clustering, but it will produce some clusters that do not satisfy the size condition. Therefore, we proceed the partitions in two stages. We give the details of our algorithm. Let $P = \left\lfloor \frac{n}{k} \right\rfloor$, where n is the number of records and k is the value for k -anonymization. During the preprocessing stage, the algorithm randomly picks P records as the initial values to build P clusters. Then, the algorithm finds the nearest cluster for each record , adds r to this cluster and subsequently updates the centroid of this cluster. The distance between a cluster and a record r is devoted as the number of records in the clusters times the distance between record r and the centroid of the cluster [8]. After the preprocessing stage, a set of clusters are constructed, but some of records might remain. The postprocessing stage is to adjust the remaining records, which adds each record into a cluster with respect to that the increment of the information loss is minimal. For those clusters with more than k records, calculate the inner cluster information loss. Then, sort all clusters by the information loss. Remove the record from the cluster that information loss is maximal. The records that are removed from the cluster are those most distant from the centroid of the cluster. Then add the record into a cluster with respect to which the increment of the information loss is minimal.

5. Experimental results

The main purpose of the experiments was to investigate the performance of our method in terms of Information Loss and Execution Time. To accurately evaluate our approach, we also compared our implementation with another two methods, greedy k member clustering proposed by Byun Ji-won [4] and one-pass k-means clustering proposed by Jun-Lin Lin [8].

5.1. Experimental Setup

All of the experiments were performed on a on a desktop PC with Intel Core2Duo 2.2 GHz CPU and 2GB of RAM under MS Window 7 operating system. And the implementation was built and run in Java. For our experiments, we used the Adult dataset from the UC Irvine Machine Learning Repository, which is considered a de facto benchmark for evaluating the performance of k -anonymity algorithms. For k -anonymization, we considered {age, work class, education, marital status, occupation, race, gender, and native country} as the quasi-identifier. Among these, age and education were treated as numeric attributes while the other six attributes were treated as categorical attributes.

5.2. Information time and Execution Time

The two metric used to measure the data quality are Information time and Execution Time. We experimentally implemented the two clustering-based k-anonymization method, greedy k member clustering proposed by Byun Ji-won [4] and one-pass k-means clustering proposed by Jun-Lin Lin [8].

Figure 2 shows the Execution Time of both three algorithms. We set k from 50 to 500 and wrote down the time respectively. According to Figure 2, our efficient generalized clustering algorithm took much less time than the two algorithms. Figure 3 recorded the Information Loss

caused by partitioning a set of clustering. It showed that our algorithm caused less information loss than the two clustering algorithm.

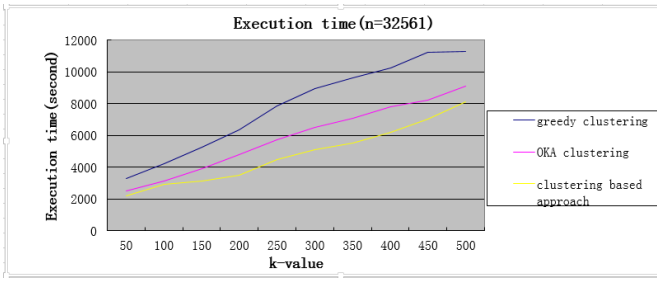

Fig. 2. *Execution time by different k*

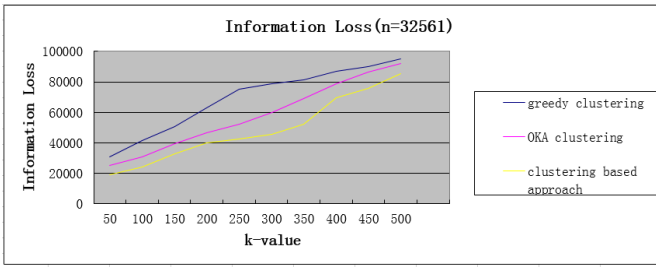

Fig. 3. *Information Loss by different k*

5.3. Histogram of the number of cluseters vs. Information loss

We also experimented on a dataset of 5000records include the attributes of age, gender and region. We let k=5 and k=10 to implement our clustering based approach. The final result was showed in the Fig4. Also we give the histogram of the number of Clusters VS. Information loss when k=5 and k=10

A0004	20代	22	女	愛知県	未婚	学生
A2501	20代	23	女	愛知県	未婚	パート・アルバイト
A3188	20代	23	女	愛知県	未婚	学生
A4422	20代	24	女	愛知県	未婚	学生
A4864	20代	23	女	愛知県	未婚	学生
A3198	40代	48	女	新潟県	既婚	専業主婦
A3647	40代	47	女	新潟県	既婚	無職
A4191	40代	49	女	新潟県	既婚	パート・アルバイト
A4607	40代	47	女	新潟県	既婚	無職
A4874	40代	48	女	新潟県	既婚	専業主婦

Fig. 4. *The final result by the dataset of 5000set*

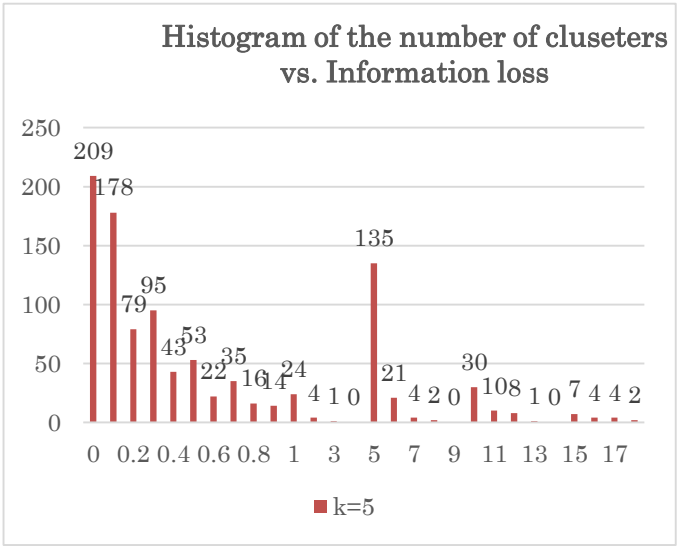

Fig. 5. *The histogram of the information loss when k=5*

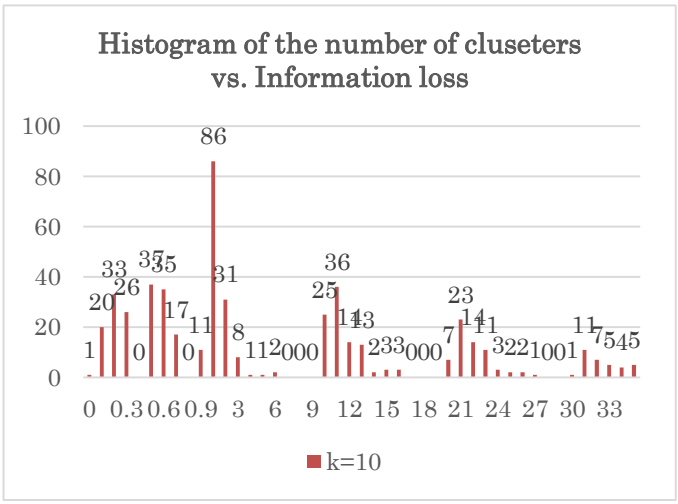

Fig. 6. *The histogram of the information loss when k=10*

6. Conclusions

In this paper, we proposed an efficient generalized clustering algorithm based k-anonymization.We defined the distance function for the numerical type (such as age), categorical type (such as nationality, occupation). We proceeded our algorithm in two stages: preprocessing stage and postprocessing stage. The preprocessing stage is to partitions all records into $\lfloor \frac{n}{k} \rfloor$ clusters, and the postprocessing stage is to adjust the remaining records in the preprocessing stage, which adds each record into a cluster with respect to that the increment of the information loss is minimal. We experimentally compared our method with two other clustering-based k-anonymization method. The experiment shows that our method outperforms their

method and also ensure the anonymization of data. Finally, many variations of the *k*-anonymization model have been proposed to further protect the data from identification, e.g., l-diversity [12], t-closeness [7].We should extend our algorithm to these models.

References

[1] Sweeney, L.: k-anonymity: a model for protecting privacy. Int. J. Uncertain. Fuzziness Knowl.-Based Syst. 10(5), 557–570 (2002)

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

[3] P. Samarati. Protecting respondent's privacy in microdata release. IEEE Transactions on Knowledge and Data Engineering, 13, 2001.

[4] J.-W. Byun, A. Kamra, E. Bertino, and N. Li.: Efficient k-anonymization using clustering techniques. In International Conference on Database Systems for Advanced Applications (DASFAA) (2007)

[5] B. C. M. Fung, K. Wang, and P. S.: Top-down specialization for information and privacy preservation. In ICDE '05: Proceedings of the 21st International Conference on Data Engineering (ICDE'05) (2005)

[6] V. Ciriani, S. D. C. di Vimercati, S. Foresti, and P. Samarati. K-anonymity. Security in Decentralized Data Management (to appear).

[7] N. Li and T. Li. t-closeness: Privacy beyond k-anonymity and l-diversity. In International Conference on Data Engineering (ICDE) (2007)

[8] Lin, J.L., Wei, M.C.: An efficient clustering method for k-anonymization. In Proceedings of the 2008 International Workshop on Privacy and Anonymity in Information Society (2008)

[9] Shyam Boriah Varun Chandola Vipin Kumar: Similarity Measures for Categorical Data: A Comparative Evaluation. Conference: Proceedings of the SIAM International Conference on Data Mining, SDM 2008, April 24-26, 2008, Atlanta, Georgia, USA

[10] G. Aggarwal, T. Feder, K. Kenthapadi, R. Motwani, R. Panigrahy, D. Thomas, and A. Zhu. Anonymizing tables. In International Conference on Database Theory, pages 246–258, 2005.

[11] MacQueen, J.B.: Some methods for classification and analysis of multivariate observations, Berkeley, pp. 281–297 (1967)

[12] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam. l-diversity: Privacy beyond k-anonymity. In Proceedings of the 22nd IEEE International Conference on Data Engineering (ICDE 2006), 2006.