Research on K-Anonymity Algorithm in Privacy Protection

Chen Wang, Lianzhong Liu, Lijie Gao
School of Computer Science and Engineering, Key Laboratory of Beijing Network Technology, Beihang University,
Beijing, 100191 China
abomb007@126.com, lz_liu@buaa.edu.cn, beautifulgaolijie@163.com

***Abstract*—Nowadays, people pay great attention to the privacy protection, therefore the technology of anonymization has been widely used. However, most of current methods strictly depend on the predefined ordering relation on the generalization layer or attribute domain, making the anonymous result is a high degree of information loss, thereby reducing the availability of data. In order to solve the problem, we propose a K-Members Clustering Algorithm to reduce the information loss, and improve the performance of k-anonymity in privacy protection.**

***Keywords*—privacy； privacy protection； k-anonymity； clustering**

I. Introduction

In recent years, the technology of data mining has been widely used in information systems [1], which made a great contribution to the development of information technology. However, at the same time, data privacy is threatened by this technology. Therefore, privacy protection technology [2] appears to protect data from being disclosed during the application process.

There are several ways to protect the privacy, such as access control [3] and anonymization. Anonymization is different from access control: the core thought of the access control is to protect data confidentiality, but anonymization does not guarantee the confidentiality of the data, and the data is completely open to the public. The core thought of anonymization is to protect the correspondence between the privacy and people.

One method to realize the anonymization is k-anonymity. Since the k-anonymity model [4][5] which first proposed by P. Samarati and L. Sweeney, researchers have been carried out extensive and in-depth research work for an effective way to seek to protect the privacy [6][7]. However, most k-anonymous methods are based on the technology of generalization and hiding, because of its strict dependence on predefined ordering relation on the generalization layer or attribute domain, making the anonymous result is a high degree of information loss, thereby reducing the availability of data [8]. This paper proposes a clustering-based anonymous method to protect privacy.

II. Related Concepts

A. Quasi-identifier

Quasi-identifier (QI) refers to the attribute or combination of attributes which link to other external data tables to identify the individual identity, such as gender, birthdate or zip code. The selection of the Quasi-identifier depends on the external data link table.

B. Linking Attack

Linking attack is a common method to obtain privacy data from the data table. Its basic idea is: attacker obtain the external data link operation through the published data and other ways, then infer privacy data and result privacy leaked. For example, in [9], attackers can uniquely ensure the medical diagnostic results of the patients through the medical information form and the voter registration form, however, the medical diagnostic result is the privacy of data.

C. K-anonymity

To resolve the loss of privacy issues caused by linking attack, L. Sweeney et al proposed a method: k-anonymity. K-anonymity release low precision data by generalization and hiding, so that in the table each record at least has the same Quasi-identifier attribute values with other k-1 records, thereby reduce the linking attack that led to the loss of privacy. Table 2 is a anonymized table of Table 1.

TABLE I. Integrity Table

Name	Race	Birth	Sex	Course	Score
Alice	Blank	1987-1-8	M	Math	77
Bob	Blank	1987-5-22	M	Math	65
David	Blank	1988-2-2	M	English	86
Helen	Blank	1988-10-14	M	English	94
Jane	White	1986-3-15	F	Chinese	72
Paul	White	1986-6-27	F	Chinese	81

TABLE II. Anonymized Table

Name	Race	Birth	Sex	Course	Score
*	Blank	1987	M	Math	77
*	Blank	1987	M	Math	65
*	Blank	1988	M	English	86
*	Blank	1988	M	English	94
*	White	1986	F	Chinese	72
*	White	1986	F	Chinese	81

III. The Clustering Algorithm of k-anonymity

A. Basic Idea

The basic idea of the algorithm is that look upon k-anonymity issues as a clustering problem, then divide data

Published by Atlantis Press, Paris, France.

objects into a number of classes or clusters, let the objects in the same cluster are highly similar while the objects in the different clusters are highly dissimilar.

B. *K members Clustering Issue*

The traditional clustering process requires specific cluster number, however, k-anonymity does not limit the number of clusters, but each cluster must contain k records at least. Therefore, k-anonymity issue can be seen as clustering issue, and usually be called as k members clustering issue.

Definition 1 (k members clustering issue): k members clustering issue is to divide a set which contain n records into a series of clusters, make each cluster containing at least k record, and require minimize the sum of cluster span inside. *S* is the set which contain n records, *k* is specific anonymous parameter, then the optimal solution of k members clustering issue is to produce a set of clusters *E*={ e_1, e_2 ,…, e_m } which is fit to the following requirements:

(1) $$\forall i \neq j \in \{1,2,...,m\}, e_i \cap e_j = \varnothing$$

(2) $$\bigcup_{i=1,2,...,m} e_i = S$$

(3) $$\forall e_i \in E, |e_i| \geq k$$

(4) $$\sum_{l=1,2,...,m} |e_l| \cdot \underset{i,j=1,2,...,|e_l|}{MAX} \triangle(p(l,i), p(l,j))$$

Wherein, |*e*| represents the size of the cluster of *e*, *p*(*l*, *i*) represent the *i* data point in e_l, △(*x*, *y*) represent the distance between the data points *x* and *y*.

C. *Distance and Cost Metrics*

The core of clustering is to define the distance function to measure the similarity between the data point, and define cost function to minimize the cost of clustering. Distance function is often decided by the data type of data point, and cost function is often defined by the object of clustering. Since the data involved by k-anonymity can contain either numeric attributes or classification attributes, thereby need to define a function which can handle different types of data. The following describe the cost function which is fit to k-anonymity.

Definition 2 (the distance between the numeric data): D is a finite value domain, any value v_i, v_j∈D, the standard distance between v_i and v_j can be defined as:

$$\delta_N(v_i, v_j) = \frac{|v_i - v_j|}{|D|}$$

Wherein, |*D*| represents the difference between the maximum and minimum values of the domain *D*.

Because most of classification domain does not have a complete order relation, therefore, the above definition does not apply to categorical data. A simple and intuitive solution is that assume the each value in domain is different from each other, if two attribute value are same, the distance is 0, otherwise the distance is 1. However, there are semantic relationships between the value of some domain, in these domains, these semantic relations defined distance function. Classification tree can usually reflect the semantic relationship, assuming a domain classification tree is a balanced tree, the leaf nodes represent domain classification. Thus, define the distance function among the categorical data.

Definition 3 (the distance between the classification data): D is classification domain, Td is the classification tree in D, for any classification value v_i, v_j∈D, the standard distance between v_i and v_j can be defined as:

$$\delta_C(v_i, v_j) = \frac{H(\Lambda(v_i, v_j))}{H(T_D)}$$

Wherein $\Lambda(v_i, v_j)$ represent the subtree with common ancestor root of v_i and v_j, $H(T)$ represent the height of classification tree T.

Combine the distance function in numerical domain and categorical domain, to define the distance between the two records are as follows:

Definition 4 (record distance): let Q_T = {N_1, N_2, … , N_m, C_1, C_2, … , C_n} is the Quasi-identifier of the data table *T*, wherein N_i (i = 1, 2, … , m) is the numerical attribute, C_j (j = 1, 2, … , n) is the categorical attribute, for arbitrary records r_1, r_2∈*T* the distance between r_1 and r_2 can be defined as follows:

$$\triangle(r_1, r_2) = \sum_{i=1,2,...,m} \delta_N(r_1[N_i], r_2[N_i]) + \sum_{j=1,2,...,n} \delta_C(r_1[C_j], r_2[C_j])$$

Wherein $r_i[A]$ represent the attributes value of record r_i.

Since the ultimate goal of k member clustering is to achieve the release of data k-anonymity, we should construct a cost function to cope with the extent of data distorted in generalization processing. Because the records in each cluster is generalized to the same quasi-identifier value, assuming that the numeric data can be generalized into interval [min, max], the categorical data can be generalized into a set of different attribute values.

Defined 5 (loss of information): let *e* = {r_1, r_2, … , r_k} is a cluster(equivalence classes), its quasi-identifier contains numerical attributes N_1, N_2, … , N_m, and categorical attributes C_1, C_2, … , C_n, T_{ci} is the classification tree of categorical attribute domain C_i, MIN_{Ni} and MAX_{Ni} are respectively the minimum value and maximum value of numerical attributes N_i in cluster *e*, $\cup_{ci}$ is the set of different attribute values in C_i, thereby the information loss generated by the generalization of cluster *e* can be defined as follows:

$$IL(e) = |e| \cdot (\sum_{i=1,2,...,m} \frac{(MAX_{N_i} - MIN_{N_i})}{|N_i|} + \sum_{j=1,2,...,n} \frac{(H(\Lambda(\bigcup_{C_j}))}{H(T_{C_j})})$$

Wherein |*e*| represent the number of records in the cluster *e*, |N_i| represent the difference between the maximum and minimum values of the numerical domain, $\Lambda(\bigcup_{Ci})$ represent the subtree with the root which has smallest common ancestor among all the values in $\bigcup_{Ci}$ of the classification tree, $H(T)$ represent the height of the classification tree.

Based on the above definitions, the total loss of information can be defined as follows:

Definition 6 (Total loss of information): Let E is the set of equivalence class of anonymous table AT, the total loss of information in AT can be defined as :

$$Total - IL(AT) = \sum_{e \in E} IL(e)$$

Because the cost function of k members clustering is the sum of distance in all clusters, wherein the distance in cluster is defined as the distance between the furthest data point within the cluster, so when generalize records in cluster, minimizing the loss of information is equal to minimizing the cost function at k member clustering process. Thereby the cost function which need to minimize when clustering is *Total- IL.*

D. The Clustering Algorithm of K-anonymity

The optimal solution of k-members clustering is exhaustive search, however, it has exponential complexity. To cope with the difficulty of the problem, we use a simple and efficient greedy algorithm. The basic idea is: for given records, first select a record randomly and look upon it as an individual cluster e, and then select records r_j make IL $(e_1 \cup r_j)$ minimum, repeat this operation until $| e |$ = k. When the size of the cluster e reaches k, then selected one record from the remaining records randomly and repeat the clustering process until the remaining number of records is less than k. Then insert the remaining records into the existing cluster, and make the minimum loss of information. It can be proved that the size of cluster which produced by greedy clustering algorithm of k members is between k and 2k-1, the time complexity is O(n). Algorithm process is as follows:

TABLE III K-MEMBERS CLUSTERING ALGORITHM

```
Algorithm：K-Members Clustering Algorithm
Input：Data set S and Anonymity parameter K
Output：Cluster set result, each contains at least k records
Process：
1.If the number of the records in data set is less than k, return;
2. Make cluster set result = ∅;
3. Randomly selected one record r from the data set S;
4. When the number of records in the data set S is not less than k, execute the loop:
  (1) Selected one record r randomly from the data set S;
  (2) S = S-{r};
  (3) c = {r};
  (4) When the number of records in cluster C is less than k, execute the loop:
    a. Select one record from the data set S, make the loss of information is minimum when add it to the cluster C;
    b. S = S-{r};
    c. c = c ∪ {r};
  (5) result = result ∪ {c};
5. when there are remaining records in the data set S, execute the loop:
  (1) Selected one record r randomly from the data set S;
  (2) S = S-{r};
  (3) selected cluster C from the cluster set result, make the loss of information is minimum when add the record to the cluster C;
  (4) c = c ∪ {r};
6. Return cluster set result
```

IV. CONCLUSION

By transform the k-anonymity into k-member clustering, we propose an effective k-anonymity algorithm, also proposed two important measurement criteria in clustering, the distance and cost metrics. The measurement criteria can characterize the data distorted in generalization process, can measure the quality metrics of k-anonymity data set, and can reduce the loss of information. Our future work is to improve the effective of the k-anonymity algorithm, and improve the performance of privacy protection.

ACKNOWLEDGMENT

This work has been supported by Co-Funding Project of Beijing Municipal Education Commission under Grant No. JD100060630.

REFERENCES

[1] J. Han and M. Kamber. Data Mining: Concepts and Techniques. 2nd edition, San Francisco: Morgan Kaufmann Publishers, 2006.

[2] V.S. Verykios, E. Bertino, I.N. Fovino, L.P. Provenza, Y. Saygin, and Y. Theodoridis. State-of-the-Art in Privacy Preserving Data Mining[J]. ACM SIGMOD Record, 2004, 3(1):50-57.

[3] R. Sandhu and P. Samarati, “Access Control: Principles and Practice,” IEEE Communications Magazine, Vol. 32, No. 9, September 1994, pp. 40-48.

[4] Samarati P,”Proteeting respondents’ identities in microdata release.” Proc of the TKDE’01: 1010-1027,2001.

[5] Samarati P, Sweeney L. Generalizing, “Data to provide anonymity when disclosing information”, Proc of the 17th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems.New York: ACM Press, 1998.

[6] Byun J W , Kamra A, Bertino E, et a1. “Eficient k-anonymization using clustering techniques”. LNCS 4443:Proceedings of DAS-FAA 2007. Berlin Heidelberg: Springer-Verlag, 2007: 188-200.

[7] Machanavajjhala A, Gehrke J, Kifer D. “l-diversity: privacy beyond k-anonymity[J]”. ACM Transactions on Kn owledge Discovery from Data. New York: ACM Press,2007,1(1):24-35.

[8] Wang Zhihui, Xu Jian, Wang wei. “Clustering data based on anonymous”. Journal of Software, 2010, 21(4):680-693.

[9] Sweeney L. K-anonymity. “A model for protecting privacy”. International Journal of Uncertainty, Fuzziness and Knowledge−based Systems, 2002, 10(5): 557−570.