Hybrid Recommendation Algorithm for User Satisfaction-oriented Privacy Model

Yinggang Sun[1], Hongguo Zhang[1], Luogang Zhang[1], Chao Ma[1*], Hai Huang[1], Dongyang Zhan[2,3] and Jiaxing Qu[4]

[1] Harbin University of Science and Technology, Harbin, 150040, China
[e-mail: syg15688708938@163.com, zhg07@163.com, 280411879@qq.com, machao8396@163.com, hust_hh@vip.163.com]
[2] School of Cyberspace Science, Harbin Institute of Technology, Harbin, 150001, China
[3] The Ohio State University, Columbus, 43202, USA
[e-mail: zhan.179@osu.edu]
[4] Heilongjiang Province Cyberspace Research Center, Harbin, 150001, China
[e-mail: smilingqu@126.com]
[*]Corresponding author: Chao Ma

Received May 4, 2022; revised August 15, 2022; accepted August 25, 2022; published October 31, 2022

Abstract

Anonymization technology is an important technology for privacy protection in the process of data release. Usually, before publishing data, the data publisher needs to use anonymization technology to anonymize the original data, and then publish the anonymized data. However, for data publishers who do not have or have less anonymized technical knowledge background, how to configure appropriate parameters for data with different characteristics has become a more difficult problem. In response to this problem, this paper adds a historical configuration scheme resource pool on the basis of the traditional anonymization process, and configuration parameters can be automatically recommended through the historical configuration scheme resource pool. On this basis, a privacy model hybrid recommendation algorithm for user satisfaction is formed. The algorithm includes a forward recommendation process and a reverse recommendation process, which can respectively perform data anonymization processing for users with different anonymization technical knowledge backgrounds. The privacy model hybrid recommendation algorithm for user satisfaction described in this paper is suitable for a wider population, providing a simpler, more efficient and automated solution for data anonymization, reducing data processing time and improving the quality of anonymized data, which enhances data protection capabilities.

Keywords: Anonymization, Historical Configuration Scheme Resource Pool, Privacy Protection, Positive Recommendation Process, Reverse Recommendation Process, Satisfaction

This work was supported by the National Natural Science Foundation of China (Grant No. 61976064), the National Natural Science Foundation of China (Grant No. 62172123), the Fundamental Research Foundation for of Heilongjiang Province, China (No. 2019KYYWF0214), the Postdoctoral Science Foundation of Heilongjiang Province, China (No. LBH-Z19067), the special projects for the central government to guide the development of local science and technology, China (No. ZY20B11), the Heilongjiang Provincial Natural Science Foundation of China, China (No. YQ2019F010).

1. Introduction

With the rapid development of new generation information technologies such as cloud computing, Internet of Things, artificial intelligence and mobile Internet, global data is showing an exponential growth trend. Through big data analysis and mining technology, we can discover the rules and trends of things, which can help managers make decisions. However, when analyzing and mining massive data, personal data will become easier to obtain and disseminate, and at the same time, violations of personal privacy rights are difficult to detect. Therefore, how to apply privacy protection technology to protect personal privacy rights from being violated has become an urgent problem to be solved. One of the solutions is to use anonymization technology.

As one of the most important privacy protection technologies, anonymization technology mainly implements privacy protection for the data release stage. The concept of anonymity was proposed by Samarati [1]. Anonymization technology refers to the anonymization of data through generalization or suppression technology in the data release stage, reducing the probability of attackers obtaining user identity and sensitive information, and protecting user privacy. The use of data anonymity technology effectively realizes the function of privacy protection of big data release [2]. At present, the anonymization process for the data release stage is usually: (1) importing the original data; (2) anonymizing the original data; (3) generating the anonymized data; (4) safely releasing the data.

Obviously, there are still many problems in the details of this process. For example, for data with different characteristics, what technology should the data publisher use to anonymize the original data, and how should the anonymized data be evaluated to determine whether it meets the requirements. In response to these problems, some researchers have supplemented some aspects of the process. After proposing the datafly anonymization algorithm, Sweeney [3] proposed a method based on the generalization level. It uses the Precision formula to compare the generalization level of each quasi-identifier in the data table before and after generalization to calculate the availability of data. Dankar [4] presents metrics on when and how to apply marketer risk measurement models to disclosure control risk.

These methods effectively complement the missing part of the anonymization process described above for evaluating post-anonymized data. Unfortunately, they have done excellent work in their respective fields, but have not integrated the entire anonymization process to improve, and no one has considered the difficulty of anonymizing data for non-professionals. For the data publisher, his ultimate goal is to make the published data have sufficient utility, that is, the information loss of the anonymized data is small, and at the same time, the published data is protected from the risk of re-identification as much as possible. For a data publisher with sufficient privacy protection knowledge, he can reconfigure the parameters according to the evaluation results and proceed to the next iteration. However, for a data publisher without sufficient privacy protection knowledge background, his configuration cannot meet the requirements with a great probability, and thus may not obtain satisfactory results. Moreover, even data publishers with sufficient knowledge of privacy protection cannot guarantee satisfactory results in a relatively short period of time. Therefore, the traditional anonymization process has the shortcomings of low efficiency, long time, and uncertain results. For these problems in reality, the above-mentioned researchers did not mention them.

In view of the above problems, this paper integrates various important aspects of anonymization technology, and designs a privacy model hybrid recommendation algorithm based on user satisfaction. Among them, for the first time, this paper proposes the concept of applying the resource pool of historical configuration schemes in the anonymization process,

and automatically recommends configuration parameters for users based on data characteristics and actual user needs to ensure that the anonymized data can meet user satisfaction requirements. The algorithm includes two processes: forward and reverse, aiming at users with two different knowledge backgrounds, and is dedicated to solving the problem of difficult parameter configuration.

The algorithm described in this paper reduces the difficulty for users to anonymize data, improves the quality of data after anonymization, reduces the time for data anonymization, enhances data protection capabilities, and provides new ideas for data security release.

The second part is to complete the paper The related work done. The third part of this paper gives the specific scenarios of the problem, and clarifies the problems that users may encounter in the process of anonymization, such as difficult parameter configuration, complicated data anonymization process, and data anonymization quality that cannot meet the requirements. The fourth part It is a specific definition of the problem solved in this paper, which quantifies the user satisfaction that the algorithm described in this paper needs to use, and sets constraints on the automatic recommendation scheme. The fifth part gives the specific solution to the problem, including the forward process and the specific steps of the reverse process and the pseudo code of the most important automatic recommendation algorithm. The sixth part is the experimental analysis of the solution described in this paper, which proves the feasibility and superiority of the solution described in this paper. The seventh part is the conclusion of this paper.

2. Related Works

With the digitalization process of various industries and the development of data collection technology, data privacy protection has become very important, and people have begun to study how to protect data privacy. In some hardware usage scenarios, [5, 6] propose some privacy-preserving solutions. In the blockchain scenario [7] proposes how to perform security authentication. At the same time, there are some solutions at the software level. For example, [8-11] describes how to ensure the privacy of user data in taxi-hailing software. [12, 13] devised the question of how to protect patient privacy during medical diagnosis. [14] proposed a scheme to protect data privacy during cloud media data sharing. However, this paper chooses to study the problem of data privacy protection from the K-anonymity method. Since the K-anonymity model proposed by Seweney in 1998, anonymization technology has been booming. In the new era and new background, people's demand for and dependence on privacy protection technology is becoming more and more profound. After proposing the K-anonymity model, Seweney proposed an improved version of the K-anonymity privacy protection model based on generalization and concealment technology on the basis of the K-anonymity model. To solve the property leakage problem of K-anonymous models, Machanavajjhala [15] proposed the L-diversity model. After that, in order to improve the flexibility of L-diversity and improve the personalized protection ability of anonymous data, LI ZD [16] proposed a (k,l)-anonymity model. In view of the lack of L-diversity model, LIN H [17] further proposed a T-approximation model. In order to adapt to the rapid development of social networks, graph-modified K-neighborhood [18], K-Degree [19], K-Isomorphism [20], K-Automorphism [21], clustering-based Partitioning [22], SANGREEA Anonymity protection models such as [23] and Generalization [24] have been proposed successively. Among the many models, K-anonymity, L-Diversity and T-Closeness are three classic privacy protection models, many of which are based on They are created to optimize and improve prototypes. For a long time, the anonymization technology has been continuously changed, the privacy model has gradually

increased, and the configuration of related parameters has become more and more complicated. For users without a background in privacy protection, using various anonymization systems or process configuration parameters becomes a big challenge. Therefore, based on the anonymization process constructed by the privacy model hybrid recommendation algorithm for user satisfaction described in this paper, 11 privacy protection models and their related parameters can be automatically recommended and configured, which is not only suitable for users with certain privacy protection knowledge, and is suitable for users with no or less knowledge of privacy protection.

In order to prove that the anonymized data still has availability and that the risk of the data meets the relevant standards, it is also indispensable for the data availability assessment and risk assessment. Many experts and scholars have devoted themselves to the research of these technologies, and have given different evaluation methods from different perspectives and different application scenarios. In terms of utility evaluation, Bayardo [25] et al. proposed an indistinguishable metric or DM, which assigns a penalty to each piece of data after generalization, and its value is the entire generalization The total number of data in the table that are indistinguishable from this piece of data. To deal with the problem of uneven distribution of data, El Emam [26] et al. proposed an entropy-based information loss metric. In terms of risk analysis of post-anonymized data, Motwani [27] proposed two natural methods for quantifying quasi-identifiers: sharpness ratio and separation ratio. In addition, there are a variety of risk assessment models, such as prosecutor model, reporter model, marketer model, etc., which can be used for risk assessment in different scenarios. The anonymization process based on the recommendation algorithm described in this paper applies the above evaluation method, which can clearly display the corresponding evaluation results to the user, so that it can make a basis for making a judgment on whether to publish the data.

3. Problem Scenario

The **Fig. 1** is a schematic diagram of the problem scenario, which respectively shows the problems encountered in the anonymization process by users with a certain privacy protection knowledge background and users with no or less privacy protection knowledge background.
For a user with a certain background in privacy protection, he knows how to configure parameters, and through these parameters, the anonymization process can be guaranteed to proceed normally. The user can also judge whether the anonymized data meets his own requirements through the corresponding evaluation model. Since users cannot predict in advance whether the output anonymized data meets their own requirements, they must enter the anonymization process. The anonymized data can only be evaluated after the anonymized data is obtained after the original data has been anonymized. Once the parameter configuration is deviated and the anonymized data fails to meet the user's requirements, the parameters need to be reconfigured, and then enter the anonymization process again. In most cases, users need to perform multiple iterations to approximate their desired results, which will consume more time. For larger amounts of data, the wasted time will be more.

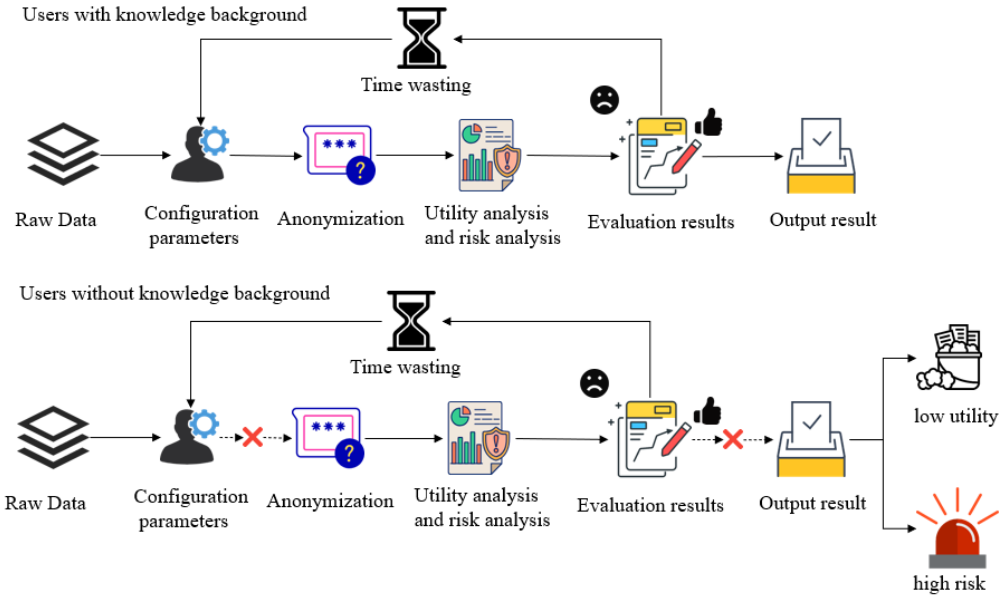

Fig. 1. Schematic diagram of the problem scenario

For a user with no or less knowledge of privacy protection, he does not know how to configure parameters. If the user is allowed to configure parameters freely, it is likely that the anonymization process cannot be performed normally. Moreover, the user cannot judge whether the anonymized data meets the requirements for publication based on the evaluation results of the evaluation model, so it is difficult to publish even if the available anonymized data is obtained. If anonymized data is released rashly, the utility of the anonymized data may be too low to be used normally, or the re-identification risk of the data may be too high, which may easily lead to privacy leakage. And the user spends more time in data anonymization.
In view of the above requirements, we need to find an automatic recommendation scheme that can satisfy two different types of users, which can automatically solve the optimal anonymization strategy in a limited number of iterations according to the user's requirements and expectations.

4. Problem Definition

The **Fig. 2** shows the anonymization process of the data used in this paper. A piece of raw data needs to be anonymized. Users first need to configure appropriate parameters for them considering their data characteristics, and then anonymize them according to the parameters. The processed anonymized data is subjected to utility analysis and risk analysis to evaluate the results. If the user evaluates the result to meet the requirements, the result is output directly. If the user is not satisfied with the result, the parameters need to be reconfigured for the next anonymization process. In this step, the data anonymization process creatively joins the historical configuration scheme resource pool to improve the efficiency of data anonymization.

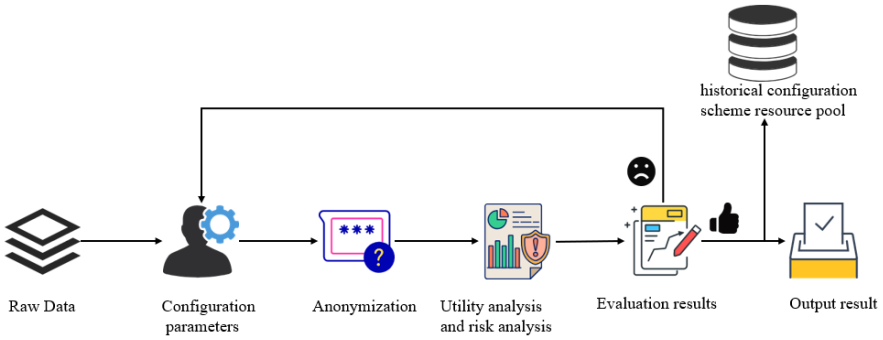

Fig. 2. Traditional data anonymization process

The historical configuration scheme resource pool is a configuration parameter database formed by parameters configured by a large number of users during the anonymization process. When a user obtains a satisfactory result, the historical configuration scheme resource pool will automatically record the data characteristics, configuration parameters and other information of the original data corresponding to the result. When the user needs to anonymize similar data, it will automatically the characteristic gives the corresponding configuration scheme. The core of the privacy protection model hybrid recommendation algorithm based on user satisfaction described in this paper is the automatic recommendation scheme of the historical configuration scheme resource pool.

$$\begin{bmatrix} u_1\ r_1\ s_1 \\ u_2\ r_2\ s_2 \\ u_3\ r_3\ s_3 \\ \vdots\ \ \vdots\ \ \vdots \\ u_i\ r_i\ s_i \end{bmatrix} = f\left(\begin{bmatrix} d_{11} & d_{12} & d_{13} & \dots & d_{1l} & p_{11} & p_{12} & p_{13} & \dots & p_{1n} \\ d_{21} & d_{22} & d_{23} & \dots & d_{2l} & p_{21} & p_{22} & p_{23} & \dots & p_{2n} \\ d_{31} & d_{32} & d_{33} & \dots & d_{3l} & p_{31} & p_{32} & p_{33} & \dots & p_{3n} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \dots & \vdots \\ d_{i1} & d_{i2} & d_{i3} & \dots & d_{il} & p_{i1} & p_{i2} & p_{i3} & \dots & p_{in} \end{bmatrix}\right) \quad (1)$$

(1) is a formalized representation of the historical configuration scheme resource pool, and a row in it corresponds to an instance in the istorical configuration scheme resource pool. The input values of this formula are the characteristics of the data and the configuration parameters. The output values are the utility of the anonymized data, the risk, and the satisfaction calculated from both. $f(x)$ for the anonymization process.

There are many data characteristics of a piece of raw data, including the semantic characteristics of data table fields, data table field type characteristics, attribute types, the number of corresponding attributes, etc. For example, In $D = \{\text{age, sex, zipcode, Integer, String, Integer}\}, \{\text{age, sex, zipcode}\}$ is the semantic feature of the data table field, $\{\text{Integer, String, Integer}\}$ is a data table field type feature. The configuration parameters include privacy model, parameters corresponding to privacy model, data suppression limit rate, attribute weight, etc. For example, In $P = \{\text{k-Anonymity, k} = 2, \text{suppression} - \text{limit} = 0.2, \text{weight} = 0.5\}$,k-Anonymity is privacy model, k=2 is privacy model parameters, suppression-limit=0.2 is data suppression limit rate, weight=0.5 is attribute weight.

Suppose the data characteristics of a piece of original data are $D_i = \{d_{i1}, d_{i2}, d_{i3}, \dots, d_{il}\}$, $i \in N$, $l \in N$. This feature corresponds to a configuration scheme $P_i = \{p_{i1}, p_{i2}, p_{i3}, \dots, p_{in}\}$, $i \in N$, $n \in N$. The result calculated using configuration scheme P_i is $R_i, i \in N$. R_i get utility u_i and risk r_i after utility analysis calculation and risk analysis calculation, $i \in N$.Satisfaction s_i can be calculated from u_i and $r_i, i \in N$. Then, s_i can be

described by the following formula,$i \in N$,

$$s_i = \frac{w}{1-u_i} + \frac{1-w}{r_i} \tag{2}$$

Among them, $w$is the weight of utility, indicating the importance of data utility in the process of anonymization. $u_i$is the result of the utility analysis obtained under the i -th scheme, r_i is the result of the risk analysis obtained under the i-th scenario.

U is the result of the utility analysis obtained under the i -th scenario , $0.5 < u_i < 1$, $0 < r_i < 0.5$.u_i and r_i are positively correlated, and when u_i decreases, r_i also decreases. When u_i increases, r_i also increases.

Assuming that the minimum utility threshold and the maximum risk threshold are given as u and r respectively, then the minimum satisfaction threshold can be obtained as

$$s_t = \frac{w}{1-u} + \frac{1-w}{r} \tag{3}$$

Then the set of schemes that should be automatically recommended is

$$P_s = \{P_i | s_i > s_t, i \in N\} \tag{4}$$

5. Solutions

5.1 Overview

According to the problem background and problem definition, we need to solve the problem of automatic parameter recommendation for users with different knowledge backgrounds. For the above problems, this paper proposes a hybrid recommendation algorithm of privacy model for user satisfaction as a solution. As shown in **Fig. 3**, the diagram of forward process and reverse process is designed as follows:

1. Positive recommendation process for users with certain knowledge background. This process applies KNN algorithm to automatically recommend configuration parameters based on user requirements and data features.
2. Reverse recommendation process for users without knowledge background. This process combines K-means semantic clustering with the KNN algorithm to automatically recommend configuration parameters based on the user's expectation of the utility and risk of the anonymized data.

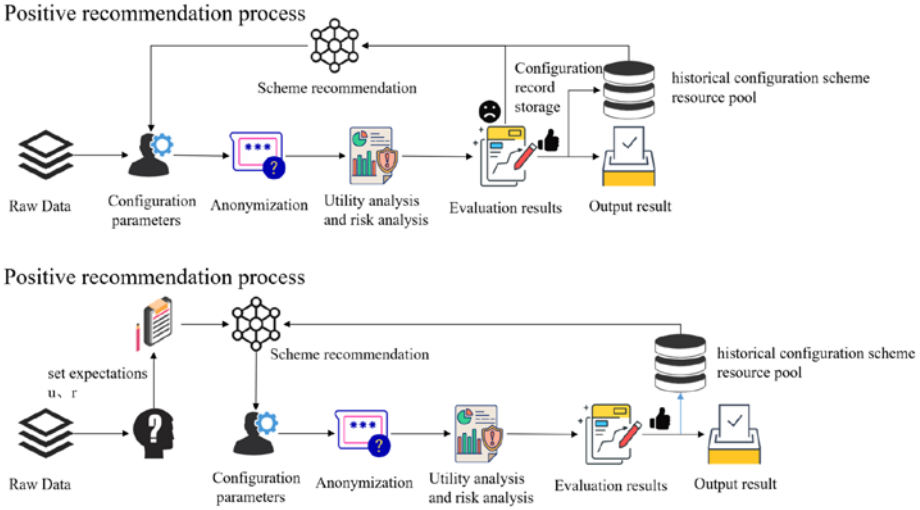

Fig. 3. Schematic diagram of the positive recommendation process and the reverse recommendation process

5.2 Scheme Design Positive Recommendation Process for Users with Certain Knowledge Background

The positive recommendation process is mainly for users with a certain knowledge background. These users know the basic anonymous technology and can configure relevant parameters by themselves. The positive process recommendation scheme for users with a certain knowledge background is described in detail as Algorithm 1 in **Table 1**. In the positive recommendation process, users can configure relevant parameters P_i according to the data features D in the original data set T_s combined with their own knowledge of anonymization processing (where i is the number of iterations, and the initialization $i = 0$). After configuring the relevant parameters, it will enter the data anonymization process. According to the parameters configured by the user, the corresponding anonymized dataset T_i will be obtained. The anonymized data will enter the utility analysis and risk analysis process. After the utility analysis and risk analysis, the corresponding utility assessment result u_i and risk assessment result r will be obtained, and the satisfaction degree s_i can be obtained through the satisfaction formula. According to u_i, r_i, and s_i, the user judges whether the satisfaction requirements are met. If the evaluation result meets the user's requirements, the corresponding T will be output directly, and the data feature D of the data and its corresponding configuration parameters P_i, utility u_i, risk r_i, and satisfaction s_i will all be stored in the historical configuration plan resource pool. If the evaluation results do not meet user requirements, use the historical configuration solution resource pool for automatic recommendation.

Table 1. Positive recommendation process algorithm

Algorithm 1 Satisfaction recommendation based on dataset and configuration parameter
Input: T_s: The original dataset with data feature D, P_0: The configuration parameters initialized by the user according to the dataset, n: Recommended result list length
Output: T_a: Complete anonymized datasets, S: satisfaction
1: $S_0, T_a \leftarrow satisfaction\_evaluation(T_s,\ P_0)$
2: **if** s_0 not satisfied **then**
3: $P_n \leftarrow \text{predict}(T_s,\ P_0)$
4: **for** $i = 0$ to n **do**
5: $S_i, T_a \leftarrow satisfaction\_evaluation(T_s,\ P_i)$
6: **if** S_i is satisfied **then**
7: $save(T_s,\ P_i)$
8: $S \leftarrow S_i$
9: **break**
10: **if** $i == n$ **then**
11: **return** satisfaction does not meet user requirements
12: **end for**
13: **else**
14: $save(T_s,\ P_0)$
15: $S \leftarrow S_0$
16: **end if**

17: **function** $satisfaction\_evaluation(T,\ P)$
18: $T_a \leftarrow data\_anonymization(T, P)$
19: $u, r \leftarrow analysis(T_a)$
20: $S \leftarrow w/(1-u) + (1-w)/r$
21: **return** S, T_a
22: **end function**
23: **return** T_a, S

The recommendation algorithm for the historical configuration scheme resource pool mentioned in it adopts the KNN algorithm. The basic idea of the KNN algorithm is to input the test data when the data and labels of the training set are known, compare the features of the test data with the corresponding features in the training set, and find the top K data that are most similar to it in the training set, then The category corresponding to the test data is the category with the most occurrences in the K data. Based on the above ideas, combined with our current problem scenario, the calculation steps of using the KNN algorithm for classification prediction are as follows: (1) Calculate the distance d between the test data T_s and each training data, (2) Sort according to the increasing relationship of the distance, (3) Select the K points with the smallest distance, (4) Determine the frequency of occurrence of the category where the first K points belong, (5) Return the category D with the highest frequency in the first K points as the predicted classification of the test data.

Among them, when calculating the distance between training data in step (1), we choose the commonly used euclidean distance calculation method, which is the most common distance representation between two points, such as $x = (x_1, x_2, \dots, x_n)$ and the euclidean distance of $y = (y_1, y_2, \dots, y_n)$ can be expressed as:

$$d(x,y) = \sqrt{(x_1-y_1)^2 + (x_2-y_2)^2 + \cdots + (x_n-y_n)^2} = \sqrt{\sum_{i=1}^{n}(x_i-y_i)^2} \quad (5)$$

$d(x,y) =$In step (3), it is more important to set the value of K. If the value of K is small, it means that the unclassified object is very close to its neighbors. A problem that arises in this way is that if the neighbor point is a noise point, the classification of unclassified objects will also produce errors, so that the KNN classification will produce overfitting. If the K value is relatively large, the points that are equivalent to too far away will also affect the classification of unknown objects. Although the advantage of this situation is strong robustness, the shortcomings are also obvious, which will cause under-fitting, that is, Unclassified objects are not actually classified. Therefore, the K value should be a result of practice, not something we set in advance. The idea of cross-validation is to use most of the samples in the sample set as the training set, and the remaining small part of the samples as the validation set for prediction to verify the accuracy of the classification model. We start by setting a small value of K, keep increasing the value of K, then calculate the variance of the validation set, and finally find a more appropriate value of K.

5.3 Reverse Recommendation Process for Users without Knowledge Background

The reverse process is mainly aimed at users without relevant background knowledge. These users do not understand the corresponding privacy protection knowledge and cannot configure appropriate parameters according to the data characteristics of the original data. The details of the reverse process recommendation scheme for users without knowledge background are described as Algorithm 2 in **Table 2**. In the reverse process, the user needs to upload the

original data T_s , and initialize the configuration parameter iteration number $i = 1$. Users need to set expectations for the utility and risk of anonymized data according to their actual needs, u_0 and r_0 respectively. The satisfaction threshold can be calculated by u_0 and r_0 as s_0. Then, the historical configuration scheme resource pool will automatically recommend the configuration parameter list $P_n = \{p_1, p_2, p_3, \dots, p_l\}$ according to the data feature D and the user-set u_0 and r_0. Starting from P_i, the configuration in the parameter list is added to the anonymous process in order, and the results are anonymous data is T_i. The utility analysis and risk analysis of anonymous data T_i are carried out, and the utility u_0 and risk r_0 are obtained respectively. According to the satisfaction formula, the corresponding satisfaction is s_i. Determine whether s_i is greater than or equal to s, and if s_i is greater than or equal to s, record in the high satisfaction list as HS. If $s_i$is less than s, it is recorded in the low satisfaction list as LS. After the recommended configuration parameter list P_n is processed, whether the high satisfaction list HS is empty is judged. If the high satisfaction list HS is not empty, T_a corresponding to the optimal satisfaction s_m in HS is selected as the optimal anonymity result. Users will assess whether the results meet their needs. Output the result if satisfied, and store the data feature D, configuration parameter P_i, utility u_i, risk r_i, satisfaction s_i of the original data into the historical configuration scheme resource pool. If the process is not terminated, the user needs to reset a reasonable expectation for iteration. If the list HS is empty, the T_i corresponding to the largest s_i in LS is selected as the optimal anonymous result. Users judge whether the results meet their needs. Output the result if the requirements are met and the process ends. The data feature D, configuration parameter P_i, utility u_i, risk r_i, satisfaction s_i of the original data are stored in the historical configuration scheme resource pool. If not satisfied, users re-set reasonable expectations for iteration.

Table 2. Positive recommendation process algorithm

Algorithm 2 Satisfaction recommendation based on utility and risk values of datasets and assessments

Input: T_s: The original dataset with data feature D,
P_s: The historical configuration scheme resource pool,
u_0: The expected utility value of the user,
r_0: The user's expected value at risk,
n: Recommended result list length

Output: T_a: Complete anonymized datasets,
S: satisfaction

$HS, LS, TL \leftarrow [\,]$
$T_k \leftarrow data\_kmeans(T_s, P_s)$
$P_n \leftarrow parameter\_recommend(u_0, r_0, T_k)$
$s_0 \leftarrow w/(1-u_0) + (1-w)/r_0$
for $i = 0$ to n **do**
  $T_a \leftarrow data\_anonymization(T_k,\ P_i)$
  $TL.\text{append}(T_a)$
  $u, r \leftarrow analysis(T_a)$
  $s_i \leftarrow w/(1-u) + (1-w)/r$
  if $s_i > s_0$ **then**
    HS.append(s_i)

12: LS.append(0)
13: **else**
14: $LS.\text{append}(s_i)$
15: HS.append(0)
16: **end if**
17: **end for**
18: **if** $len(HS) \neq 0$ **then**
19: $s_m, index \leftarrow max(HS)$
20: **if** s_m not satisfied **then**
21: **return** satisfaction does not meet user requirements
22: **else**
23: $T_a \leftarrow TL[index]$
24: **end if**
25: **else**
26: $s_m, index \leftarrow max(LS)$
27: **if** s_m not satisfied **then**
28: **return** satisfaction does not meet user requirements
29: **else**
30: $T_a \leftarrow TL[index]$
31: $P_s \leftarrow save(T_a,\ s_m)$
32: **return** T_a

In the above algorithm process, the recommendation process for the historical configuration scheme resource pool of the reverse process includes $data\_kmeans$ and $parameter\_recommend$. Among them, $data\_kmeans$ is the configuration plan information T_k of a set of similar attribute features found in the historical configuration plan resource pool P_s through the K-means algorithm according to the input feature attribute name. $parameter\_recommend$ is to use the KNN algorithm mentioned in the forward recommendation process again in T_k according to the utility and risk expectations input by the user, and filter out the configuration scheme through utility and risk, that is, the configuration scheme that the user is satisfied.

In the $data\_kmeans$ process, we use the K-means algorithm to cluster similar attribute features to realize the process of automatically finding similar configuration schemes. The K-means algorithm takes k as a parameter, and divides n objects into k clusters, so that the similarity within the cluster is high, and the similarity between the clusters is low. The process is as follows: (1) Randomly select k objects among the n objects as the initial clustering centers, (2) Group the remaining n-k objects into the nearest cluster, (3) Recalculate the center of each cluster, (4) Repeat (2) and (3) until the center of each cluster does not change.

Among them, when calculating the distance in step (2), we can use the cosine similarity calculation. It can be seen from the cosine theorem that the closer the Out is to 1, the smaller the angle between the two vectors, which means that the two texts are more similar. Calculated as follows:

$$Out = \frac{X^T Y}{\sqrt{X^T X}\sqrt{Y^T Y}} \tag{6}$$

The range of Out is $[0, 1]$, X represents the word vector of the first input feature, and$Y$ represents the word vector of the second input feature.

We need to convert attribute features into word vectors. We chose to use the word2vec [28] language model, which can train word vectors quickly and efficiently. There are two word2vec models, namely the CBOW model and the Skip-gram model, both of which use a current word to predict other words, thereby improving the accuracy of word-unit associations. Since the vocabulary generated during training is often more than tens of thousands, which greatly reduces the training speed, we choose the CBOW model and use negative sampling optimization to improve the training speed. This combination has the characteristics of fast operation. Any sampling algorithm should ensure that samples with higher frequencies are easier to be sampled. The essence of negative sampling is to update part of the weights of the neural network with one training sample at a time. The number of word vectors in the CBOW model is large, and the neural network has a huge number of weights. Unlike the original update of all the weights for each training sample, the negative sampling allows only a part of the weights to be updated for a training sample at a time, and all other weights are fixed, so that That is, the amount of calculation can be reduced, and at the same time, randomness can be increased to a certain extent, and the loss value can be reduced.

On the one hand, the forward process and the reverse process respectively anonymize the data for different users, and automatically recommend parameters through the resource pool of the historical configuration scheme to reduce the number of iterations of the anonymization process. On the other hand, in the process of continuous use, the resource pool of the historical configuration scheme will be continuously updated and upgraded, so that the recommended scheme will be more in line with user needs. Therefore, the anonymization process using the forward process and the reverse process will greatly improve the efficiency and quality of data anonymization.

6. Experimental Analysis

6.1 Experimental Environment

This paper uses IntelliJ IDEA as the compiler. The programming language is Java1.8, the experimental framework is Spring Boot and the computer hardware configuration is Intel(R) Core(TM) i5-10200H CPU @ 2.40GHz, 16GB RAM, 64 Bit Windows10 operating system.

6.2 Experimental Data

The datasets used in this experiment are six datasets from the real world:(1)US Census ,portion of data from the 1994 U.S. Census database, (2)Competition ,dataset in the 1998 KDD Data Mining Competition, (3)Accident Statistics, from the Fatality Analysis Reporting System, (4) Time Use Survey, data from the American Time Use Survey, (5)Health Interviews ， data from American Health Interviews, (6)Community survey, data from the U.S. Census Bureau, including people's emotions, social and economic characteristics, and more. These datasets have different characteristics, which are listed in **Table 3**.

Table 3. Data characteristics of different datasets

Dataset	Attributes Number	Attribute Complexity	Records Number	Records Complexity	Identifiability
US Census	9	Medium	30162	Low	Low
Competition	8	Medium	63441	Low	Medium
Accident Statistics	8	Medium	100937	Medium	Low
Time Use Survey	9	Medium	539253	Medium	Low
Health Interviews	9	Medium	1193504	High	Low
Community Survey	30	High	68752	Low	High

Table 3 shows the data characteristics of six different data sets, where Attributes Number represents the number of different field names in the data table. Attribute Complexity indicates the complexity based on the number of field names in the data table. The greater the number of attributes, the higher the attribute complexity. The Records Number represents the number of data contained in the data table, and the record complexity represents the complexity caused by the number of records in the data set. The greater the number of records, the higher the record complexity. Identifiability represents the number of quasi-identifiers in the data table. A combination of these quasi-identifiers can be used to identify a unique individual in the table. The higher the Identifiability, the higher the risk of privacy leakage in the data table.

6.3 Experimental Program

The main purpose of this experiment is to explore the difference in efficiency and user satisfaction between the anonymization process based on the recommendation algorithm described in this paper and the traditional anonymization process, as well as the user satisfaction of the anonymized data due to the size of the resource pool of the historical configuration scheme Impact.

In order to explore the difference in efficiency and user satisfaction between the anonymization process based on the recommendation algorithm described in this paper and the traditional anonymization process, the specific experimental scheme is shown in Experiment 1.

Experiment 1: We need to use the traditional anonymization process and the anonymization process based on the recommendation algorithm described in this paper to anonymize the above data. The traditional anonymization process is an anonymization process that does not use the historical configuration scheme resource pool to recommend configuration solutions. In this process, the user is the only parameter configure. If the user is satisfied with the evaluation result, the parameter should be added to the historical configuration scheme resource pool as shown in **Fig. 2**.

Before the anonymization process based on the algorithm described in this paper, the above-mentioned various data have been processed many times using the traditional anonymization process. The configuration scheme with the highest satisfaction obtained every time and the corresponding data characteristics are stored in the historical configuration scheme resource pool. On this basis, there are already 5,000 configuration schemes in the historical configuration scheme resource pool, which ensures that the amount of data in the historical configuration scheme resource pool is sufficient to support the normal progress of the positive recommendation process and the reverse recommendation process.

On the basis of the existing 5000 pieces of data in the historical configuration scheme resource pool, the positive recommendation process described in this paper is used to anonymize the above data. Through the historical configuration scheme resource pool, n configuration schemes can be recommended for users to choose.

On the basis of the existing 5000 pieces of data in the historical configuration scheme resource pool, the reverse process is used to anonymize the data. This process requires the user to set initial expectations for utility and risk, and the historical configuration scheme resource pool can provide n configuration schemes for the user to choose from.

We will record the results of each iteration, including the value u_i of the utility analysis, the result r_i of the risk analysis, and the value s_i of the satisfaction. The number of iterations is all set to n times.

In specific experiments, we set n to 6. For the parameters in the reverse process, we set the user's expected value of utility u = 0.7, the expected value of risk r = 0.25, and the satisfaction threshold calculated by formula (2) is 3.67.

Fig. 4 shows the increasing trend of satisfaction with the results obtained by three different anonymization processes as the number of iterations increases when there are 5000 samples. As can be seen from **Fig. 4**, compared with the traditional anonymization process, both the forward recommendation process and the reverse recommendation process can achieve a higher level of satisfaction within a short number of iterations. Especially in the reverse recommendation process, after the data in the historical configuration plan resource pool reaches 5,000, it can rely on its recommended parameters to keep the satisfaction of the anonymized results at a high level.

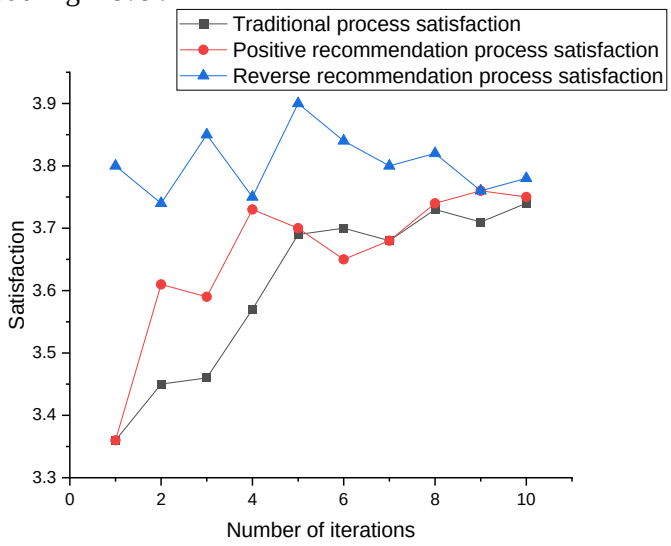

Fig. 4. Three Process Satisfaction Line Charts

Fig. 5 shows a histogram of the time spent by three different anonymization processes as the number of iterations increases when there are 5000 samples. It can be seen from **Fig. 5** that at the beginning of the iteration, the execution time of the traditional anonymization process is shorter when the number of iterations is small, but as the number of iterations increases, the time spent increases exponentially each time. It takes 2.4 times more time for the traditional process to iterate ten times than using the forward recommendation process, and 3 times more time than using the reverse recommendation process.

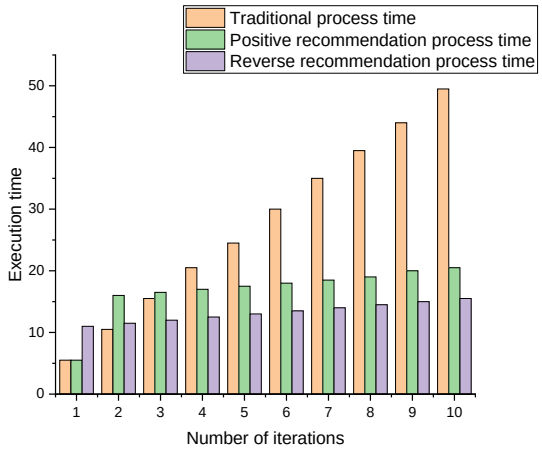

Fig. 5. Execution time at different iterations

In order to explore the influence of the size of the historical configuration scheme resource pool on the user satisfaction of anonymized data, the specific design scheme is shown in Experiment 2.

Experiment 2: Change the size of the historical configuration scheme resource pool, and compare the highest satisfaction achieved by the configuration parameter solutions recommended in the historical configuration scheme resource pools of different sizes.

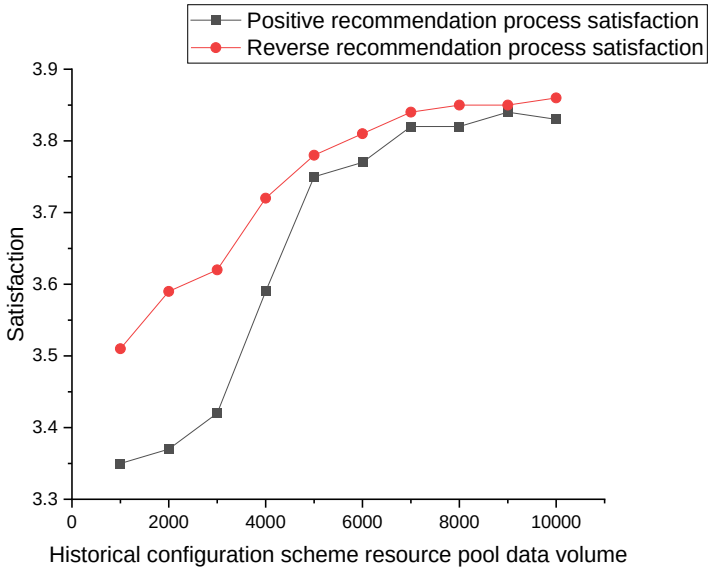

Fig. 6. The impact of historical configuration scheme resource pool on satisfaction

Fig. 6 shows the satisfaction level of the anonymized results obtained through the forward recommendation process and the reverse recommendation process as the amount of sample data in the historical configuration scheme resource pool increases when the number of iterations is 10. It can be seen that, whether it is a forward recommendation process or a reverse recommendation process, as the amount of sample data in the historical configuration scheme resource pool gradually increases, the satisfaction of the obtained anonymized results is higher. However, when the amount of sample data in the resource pool of historical configuration

solutions exceeds a certain threshold, the increase in satisfaction will tend to be flat.

From the results of Experiment 1, it can be seen that the anonymization process with the automatic recommendation algorithm described in this paper is more in line with the needs of the public than the traditional anonymization process, and can obtain better results in a given number iterations. Improve the efficiency of anonymization and reduce unnecessary time.

From the results of experiment 2, it can be seen that the size of the resource pool of the historical configuration scheme is not as large as possible. When the amount of data in the resource pool of the historical configuration scheme not reaches a certain threshold, the amount of data in the resource pool of the historical configuration scheme can be continuously increased. However, when the amount of data in the historical configuration scheme resource pool reaches a certain threshold, it is necessary to appropriately adjust the strategy for adding configuration schemes to the historical configuration scheme resource pool, reduce the number of configuration schemes stored in the library, and optimize the automatic recommendation mechanism for configuration schemes. Reduce matching time and maximize the benefits of automatic recommendation.

7. Summary

For a long time, the research focus of researchers on anonymization technology is how to improve the privacy model or improve the measurement method of data to provide better anonymization effect. This clearly promotes the development of anonymization technology. But at the same time, how to deal with the needs of users with different knowledge backgrounds for anonymization technology in the big data environment is still unresolved. The privacy model hybrid recommendation algorithm described in this paper focuses on combining various anonymization techniques into a realistic system, and provides data anonymization solutions for a wider range of users. Compared with the previous anonymization process, this scheme improves the efficiency of anonymized data and increases the quality of anonymized data. And the scheme has sufficient scalability. When a new anonymization technology appears, it is easy to add this technology to achieve the purpose of rapid update and iteration.

References

[1] P. Samarati, L. Sweeney, “Generalizing data to provide anonymity when disclosing information,” *PODS*, vol.98, no.188, pp.10-1145, 1998. Article(CrossRef Link)

[2] T. Li, N. Li, J. Zhang and I. Molloy, "Slicing: A New Approach for Privacy Preserving Data Publishing," *IEEE Transactions on Knowledge and Data Engineering*, vol. 24, no. 3, pp. 561-574, 2012. Article(CrossRef Link)

[3] L. Sweeney, “Datafly: A system for providing anonymity in medical data,” *Database Security XI*, Springer, Boston, pp. 356-381, 1998. Article(CrossRef Link)

[4] F. K. Dankar, K. El. Emam, “A method for evaluating marketer re-identification risk,” in *Proc. of the 2010 EDBT/ICDT Workshops*, pp.1-10, 2010. Article(CrossRef Link)

[5] H. Lu, C. Jin, X. Helu, C. Zhu, N. Guizani, Z. Tian, “AutoD: Intelligent Blockchain Application Unpacking Based on JNI Layer Deception Call,” *IEEE Network*, vol.35, no.2, pp.215-221, 2021. Article(CrossRef Link)

[6] H. Lu, C. J. Jin, X. H. Helu, M. Zhang, Y. B. Sun, Y. Han, Z. H. Tian, “Research on Intelligent Detection of Command Level Stack Pollution for Binary Program Analysis,” *Mobile Networks and Applications*, vol. 26, pp. 1723-1732, 2021. Article(CrossRef Link)

[7] J. Zhang, G. Xu, X. Chen, H. Ahmad, X. Liu, “Towards privacy-preserving cloud storage: a blockchain approach,” *Computers, Materials & Continua*, vol. 69, no.3, pp. 2903–2916, 2021. Article(CrossRef Link)
[8] H. Yu, X. Jia, H. Zhang, X. Yu, J. Shu, “PSRide: Privacy-Preserving Shared Ride Matching for Online Ride Hailing Systems,” *IEEE Transactions on Dependable and Secure Computing*, vol.18, no.3, pp.1425-1440, 2021. Article(CrossRef Link)
[9] H. Yu, X. Jia, H. Zhang, J. Shu, “Efficient and Privacy-Preserving Ride Matching Using Exact Road Distance in Online Ride Hailing Services,” *IEEE Transactions on Services Computing*, vol. 15, no. 4, pp. 1841-1857, 2022. Article(CrossRef Link)
[10] H. Yu, J. Shu, X. Jia, H. Zhang, X. Yu, “lpRide: Lightweight and Privacy-Preserving Ride Matching Over Road Networks in Online Ride Hailing Systems,” *IEEE Transactions on Vehicular Technology*, vol.68, no.11, pp.10418-10428, 2019. Article(CrossRef Link)
[11] H. Yu, H. Zhang, X. Yu, X. Du, M. Guizani, “PGRide: Privacy-Preserving Group Ridesharing Matching in Online Ride Hailing Services,” *IEEE Internet of Things Journal*, vol.8, no.7, pp.5722-5735, 2021. Article(CrossRef Link)
[12] X. Zhang, X. Sun, X. Sun, W. Sun, S. K. Jha, “Robust reversible audio watermarking scheme for telemedicine and privacy protection,” *Computers, Materials & Continua*, vol.71, no.2, pp.3035–3050, 2022. Article(CrossRef Link)
[13] A. Bhardwaj, A. A. Mohamed, M. Kumar, M. Alshehri, A. Abugabah, “Real-time privacy preserving framework for covid-19 contact tracing,” *Computers, Materials & Continua*, vol.70, no.1, pp.1017–1032, 2022. Article(CrossRef Link)
[14] T. Zheng, Y. Luo, T. Zhou, Z. Cai, “Towards differential access control and privacy-preserving for secure media data sharing in the cloud,” *Computers & Security*, 113, 2022. Article(CrossRef Link)
[15] A. Machanavajjhala, D. Kifer and J. Gehrke, “l-diversity: privacy beyond K-anonymity,” *ACM Transactions on Knowledge Discovery from Data*, vol. 1, no. 1, pp. 3-es, 2007. Article(CrossRef Link)
[16] Z. Li, G. Zhan and X. Ye, “Towards an Anti-inference (K, ℓ)-Anonymity Model with Value association rules," in *Proc. of International Conference on Database and Expert Systems Applications*, Springer, Berlin, Heidelberg, pp. 883-893, 2006. Article(CrossRef Link)
[17] N. Li, T. Li and S. Venkatasubramanian, “t-Closeness: Privacy Beyond k-Anonymity and l-Diversity,” in *Proc. of 2007 IEEE 23rd International Conference on Data Engineering*, pp. 106-115, 2007. Article(CrossRef Link)
[18] B. Zhou and J. Pei, “Preserving privacy in social networks against neighborhood attacks,” in *Proc. of 2008 IEEE 24th International Conference on Data Engineering*, pp. 506-515, 2008. Article(CrossRef Link)
[19] K. Liu and Terzi E, “Towards identity anonymization on graphs,” in *Proc. of the 2008 ACM SIGMOD international conference on Management of data*, pp. 93-106, 2008. Article(CrossRef Link)
[20] J. Cheng, A W. Fu and J. Liu, “K-isomorphism: privacy preserving network publication against structural attacks,” in *Proc. of the 2010 ACM SIGMOD International Conference on Management of data*, pp. 459-470, 2010. Article(CrossRef Link)
[21] L. Zhou, L. Chen and Özsu M T, “K-automorphism: A general framework for privacy preserving network publication,” *Proceedings of the VLDB Endowment*, vol. 2, no. 1, pp. 946-957, 2009. Article(CrossRef Link)
[22] S. Bhagat, G. Cormode and B. Krishnamurthy, “Class-based graph anonymization for social network data,” *Proceedings of the VLDB Endowment*, vol. 2, no. 1. pp. 766-777, 2009. Article(CrossRef Link)
[23] A. Campan, M. Traian and Truta, “A clustering approach for data and structural anonymity in social networks,” in *Proc. of 2nd ACM SIGKDD International Workshop on Privacy, Security, and Trust in KDD(Pin KDD)*, 2008. Article(CrossRef Link)
[24] M. Hay, G. Miklau and D. Jensen, “Resisting structural re-identification in anonymized social networks,” *Proceedings of the VLDB Endowment*, vol. 1, no. 1, pp. 102-114, 2008. Article(CrossRef Link)

[25] R. J. Bayardo and Rakesh Agrawal, "Data privacy through optimal k-anonymization," in *Proc. of 21st International Conference on Data Engineering (ICDE'05)*, pp. 217-228, 2005. Article(CrossRef Link)
[26] K. El Emam, A. Brown and P. AbdelMalik, "Evaluating predictors of geographic area population size cut-offs to manage re-identification risk," *Journal of the American Medical Informatics Association*, vol. 16, no. 2, pp. 256-266, 2009. Article(CrossRef Link)
[27] R. Motwani and Y. Xu, "Efficient algorithms for masking and finding quasi-identifiers," in *Proc. of the Conference on Very Large Data Bases*, pp. 83-93, 2008. Article(CrossRef Link)
[28] Q. Le, T. Mikolov, "Distributed Representations of Sentences and Documents," in *Proc. of the 31st International Conference on Machine Learning*, pp.1188-1196, 2014. Article(CrossRef Link)

Yinggang Sun received his B.Eng. degree in Software Engineering from Harbin University of Science and Technology, China (2020). He currently studying for M.E. degree in Computer Science and Technology at Harbin University of Science and Technology. His research areas include Federated Learning, Privacy Computing, and Service Recommendation.

Hongguo Zhang received his B.Eng. degree in Computer and Application from Wuhan Technical University of Surveying and Mapping, China (1991). He received his M.E. degree in Computer Application Technology from Harbin University of Science and Technology, China (1997). His Ph.D. degree in Computer Application Technology was obtained from Harbin Institute of Technology (2009). His research areas include Enterprise Intelligent Computing, Cloud Manufacturing and Scheduling Optimization, Service Engineering and Service Computing, Software Engineering and Application.

Luogang Zhang received his B.Eng. degree in Measurement and Control Technology and Instrumentation Program Control from Wenzheng College of Soochow University, China (2020). He currently studying for M.E. degree in Software Engineering at Harbin University of Science and Technology. His research areas include Privacy Computing, and Service Recommendation.

Chao Ma received his B.Eng. degree in Computer Science and Technology from Northeast Agriculture University, China (2006). He obtained his M.E. degree in Computer Application Technology from Harbin Institute of Technology, China (2008). His Ph.D. degree in Computer Application Technology was also obtained from Harbin Institute of Technology (2013). His research areas include Service Computing and Service Recommendation, Privacy Computing and Federated Learning, Intelligent Manufacturing and Big Data Analysis.

Hai Huang received his B.Eng. degree in Automation from Harbin University of Science and Technology, China (2004). His M.E. degree in Pattern Recognition and Intelligent Systems was also obtained from Harbin University of Science and Technology (2007). He received his Ph.D. degree in Microelectronics and Solid State Electronics from Harbin Institute of Technology, China (2013). His research areas include Information Security, Reconfigurable Computing, and Integrated Circuit Design.

Dongyang Zhan received his B.Eng. degree in Computer Science and Technology from Harbin Institute of Technology, China (2014). His Ph.D. degree in Computer Science and Technology was also obtained from Harbin Institute of Technology (2019). His research areas include Cloud Computing Security, and System Security.

Jiaxing Qu received his B.Eng. degree in Computer Science and Technology from Heilongjiang University, China (2002). He received his M.E. degree in Software Engineering from Northeastern University, China (2007). His Ph.D. degree in Computer Application Technology was obtained from Harbin Engineering University, China (2017). His research areas include Network Security, Informatization, National Defense Science and Technology Research and Application.