Privacy-Preserving Data Mining Metrics for Social Networking Services

Ayahiko Niimi
Faculty of Systems Information Science
Future University Hakodate
2-116 Kamedanakano, Hakodate
Hokkaido 041-8655, Japan

Takahiro Arakawa
School of Systems Information Science
Future University Hakodate
2-116 Kamedanakano, Hakodate
Hokkaido 041-8655, Japan

Abstract

This study aims to determine whether privacy-preserving data mining method can be effectively applied in data mining for a social networking service (SNS). By implementing privacy-preserving data mining on personal information collected by a SNS, it becomes possible to provide secure personalized services to SNS users. In this study, we consider using privacy-preserving data mining by the anonymization approach. By this approach, all input information is anonymized while performing data mining. We examine whether the anonymization approach can be applied to data that can be partially anonymized, such as the SNS data, and how many users can be identified by the anonymization approach. In the experiments conducted in this study, we anonymized some dataset attributes and then counted whether we can identify the data of the anonymized attributes from the data of the non-anonymized attributes. The ratio of anonymization to all data is defined as the security level of an anonymization approach. In this study, it became clear that the anonymization approach can be applied to data, which can be partially anonymized, such as a SNS data. In addition, it became clear that it is necessary to consider removing a small number of attributes as identifiers because they can be narrowed down even if they are anonymized as quasi-identifiers. We also propose clutering using randomization as a privacy-preserving data mining algorithm for SNS.

Keywords: Social Networking Service (SNS), Privacy, Data Mining, Partially Anonymized Data, Clustering Using Randomization

1. Introduction

In recent years, social networking services (SNSs) utilizing personal information such as address and birthdays have been widely used. By performing data mining on the personal information stored in a SNS, it becomes possible to provide services. Data mining is a technique used to obtain relevant knowledge from a large amount of accumulated data. However, such a techniques poses risk of the leakage of personal information during data processing. As a result, research on privacy-preserving data mining is being conducted [1]–[3]. Privacy preserving data mining is a technology that obtains relevant knowledge from a large amount of data while protecting personal and confidential information.

In recent years, privacy-preserving data mining has gained considerable attention. However, studies considering the possibility of identifying an individual when secret information is combined with publicly available information that cannot be hidden are scarce. By performing privacy-preserving data mining on personal information used on a SNS, the provision of services those are more personalized and safe has become possible. This study aims to evaluate a privacy-preserving data mining method that can be used for SNS data mining. To archive this goal, we consider the application of privacy preserving data mining using an anonymization approach. In data mining, all input information is anonymized and mined. However, the data published on the Internet that is the browsing history information of a SNS cannot be hidden. Anonymization of all the input information combined with publicly available data may be sufficient to identify an individual. Thus, an individual’s ingenuity is required. Based on this characteristic, we consider the points to be noted when applying privacy-preserving data mining to a SNS’s history information using the anonymization approach. In this study, we investigate the anonymization approach and verify its effects on its application to some data. We also propose cluttering using randomization as a privacy-preserving data mining algorithm for SNS.

The remainder of this study is organized as follows. Section 2 describes privacy-preserving data mining methods. Section 3 discusses the method proposed in this study. Section 4 presents the outline, results, and consideration of the experiment with the proposed method. In section 5, the randomization decision tree is introduced, and randomization clustering is proposed. Section 6 concludes the study.

2. Privacy-Preserving Data Mining

This section describes the types of data handled, main methods used in privacy-preserving data

| DOI: 10.20533/ijds.2040.2570.2021.0208

mining, and relationship between this study and privacy-preserving data mining.

2.1. Data Types

Attribute data that can directly identify a specific individual, such as an individual number introduced in Japan or a social security number introduced in the United States, is called identifier data. Attribute data that can indirectly identify a specific individual, such as gender, birthday, and address, in combination with other attribute data is called quasi-identifier data.

2.2. Method Assuming a Third Party (Ideal Model)

In this method, a trusted third party (TTP) that, does not leak any information aggregates data and performs data mining. This method is considered to be the safest. However, TTP installation is often unrealistic as it must be performed by the government or a reliable institution.

2.3. Anonymization

In the anonymization approach, data is processed, and data mining is performed to avoid the identification of a particular individual. Specifically, processes such as the deletion of the identifier, integration of multiple variable values of the quasi-identifier into one category and converting a variable value to an ID are performed. Thus, even if the identifier is deleted, there is a possibility that the individual can be indirectly identified with the combination of other data. To achieve anonymity, it must be designed to achieve anonymity definitions such as k-anonymity and l-diversity[4], [5]. k-Anonymity is the property in which there are at least k amount of data with the same number of attribute values. l-Diversity is the property in which there are at least l variations in the attribute values of confidential data in the k-anonymity data.

2.4. Randomize

In randomization, random noise is added to personal information, and data mining is performed. Specifically, processes such as adding random noise to variable values, random exchange with the data of other individuals, and replacing variable values with random values are performed.

Randomization is a lossy operation in which the restoration of original data is difficult; thus, privacy is protected. Computational cost for this method is low; however, accuracy and safety are statistical. Moreover, the higher is the degree of randomization, the higher is the safety but the less accurate are the results.

2.5. Encrypt

In encryption, data is encrypted, and data mining is performed. Secret calculation, which is one of the encryption approaches, is a technology that performs calculations such as statistical analysis and machine learning, while maintaining the confidentiality of personal information. In addition, it only outputs the results[6]. With data encryption, privacy is protected as the data is randomized and encrypted during the secret calculation. However, the computational cost of this method is high but accuracy and security are more stringent than the aforementioned methods due to encryption.

2.6. Relationship with This Study

This study is related to anonymization. In the existing method, all data is anonymized and data mining is performed. In this study, we consider the effect of anonymization of only a part of the data and performing data mining to reproduce the data published on the Internet, such as personal information of SNS.

3. Privacy-Preserving Data Mining Metrics for Social Networking Services

Privacy-preserving data mining using the anonymization approach anonymizes all input information and performs data mining. As a characteristic of a SNS information, data published on the Internet cannot be hidden. Anonymization of all input information combined with publicly available data may lead to the identification of an individual. Based on this characteristic, we developed a privacy-preserving data mining method that could be used for SNS data mining.

The application of privacy-preserving data mining to personal information used in a SNS allows the provision of safe and personalized services. In this study, we consider the application of privacy-preserving data mining using the anonymization approach. With this method, all input information is anonymized and mined. We also determine whether the anonymization approach can be applied to data that can be partially anonymized, such as SNS data, and whether this allows the identification of an individual.

In this study, we propose a method to evaluate the degree of security of anonymization when only a part of the input data is anonymized. The proposed method, first, some attributes of the dataset are anonymized. Second, whether the data of the anonymized attributes can be identified from the data of the non-anonymized attributes of the dataset is determined. This leads to creation of a dataset in which only some data are anonymized.

For the created dataset, the amount of data that applies to the attributes' data that has not been anonymized from the dataset is counted. The ratio of anonymization to all data is defined as the degree of security of anonymization. This makes it possible to evaluate the degree of anonymization of the data in which only some data is anonymized, using the defined degree of anonymization. The procedures of the proposed method are as follows.

i. Anonymize datasets using tools to anonymize some attributes
ii. Search and count anonymous attributes with non-anonymous attributes to determine safety

The degree of security of anonymization may change not only with the anonymization method but also with the data. In this study, we investigated the distribution of all the attributes of the dataset and judged the attributes with a particularly large distribution bias.

i. The identification of non-anonymous attributes is quite difficult, even if they are published for attributes with a small distribution bias.

   Example: If the gender ratio is similar, we can narrow down by half when searching for data for a particular woman.

ii. It is thought that attributes with a large distribution bias can identify non-anonymous attributes to some extent if they are published.

   Example: If we have a small number of old people, we can narrow down to a small number when looking for data for a particular old person.

In the next section, based on such distribution bias, the degree of safety of anonymization for data, in which only some data is anonymized will be experimentally examined.

4. Experiment

In this section, we experimentally examine the degree of safety of anonymization for data, in which only some data is anonymized against the bias of data distribution.

4.1. Experiment Summary

Our experiment aimed to investigate whether the anonymization approach can be applied to partially anonymized data, such as SNS data. Moreover, we specifically examine whether an individual can be identified by using an anonymization approach on a dataset with a defined degree of security.

In the experiment, we first anonymized some dataset attributes. Second, we examined whether the data of the anonymized attribute can be identified from the data of the non-anonymized attribute of the dataset. This led to the creation of a dataset, in which only some data is anonymized. For the created dataset, the number of data that applies to the attribute data that has not been anonymized from the dataset is counted. In addition, the ratio of anonymization to all data is defined as the degree of security of anonymization. This allows the evaluation of the degree of anonymization for data in which only some data is anonymized using the defined degree of anonymization.

4.2. Experimental Procedure

The procedures of the experiment are as follows:

i. Anonymize datasets using tools to anonymize some attributes

ii. Search and count anonymous attributes with non-anonymous attributes to determine safety

In this study, we investigated the distribution of all attributes of the dataset and judged the attributes with a particularly large distribution bias. It is difficult to identify non-anonymous attributes even if they are published for attributes with a small distribution bias. For example, if the gender ratio is about the same, you can narrow down by about half when searching for data on a specific woman. It is thought that attributes with a large distribution bias can identify non-anonymous attributes to some extent if they are published. For example, if there are few elderly people, you can narrow down to a small number when searching for data on a specific elderly person.

In this experiment, the UT Dallas Anonymization ToolBox [7] was used. With this tool, parameters such as k-anonymity and l-diversity can be set according to the anonymization method. In this experiment, Census Income attached to the UT Dallas Anonymization ToolBox was used as the dataset. Such dataset is also known as “Census Income” dataset or “Adult” dataset [8]. Census data collected by the United States Census Bureau in 1994 and 1995. The purpose was to determine whether the annual income is **$ 50,000/year** or more or less than **$ 50,000/year** from the census data. It has 42 attributes, such as age, occupation, education, race, gender, assets, place of residence, and place of birth, and 95130 data. Although this data is not considered as SNS data, it is a dataset that contains information similar to public information and non-public information common in SNS. This made us think, that it was a dataset suitable for research purposes, and thus, we used it in the experiment.

The settings of the anonymization tool are presented. The anonymization method was Datafly, and the privacy definition was k-anonymity ($k = 400$). The attributes **class_of_worker** and **detailed_industry_recode** were used as identifiers and thus deleted. The attribute **sex** was considered as a quasi-identifier and was thus mapped to $[0:1]$. The attribute **age** was also considered as a quasi-identifier and was mapped to $[0:100)$. The attribute salary was confidential information.

4.3. Experimental Results

The Figures 1, 2, and 3 present the attributes that are biased in the distribution of each attribute of the data before anonymization.

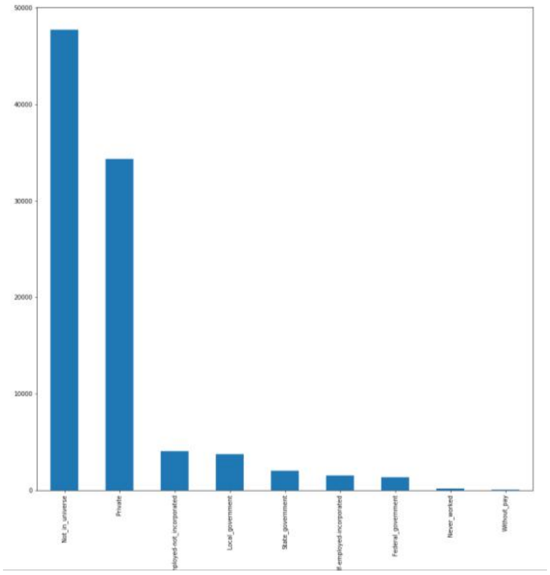

Figure 1. class_of_worker

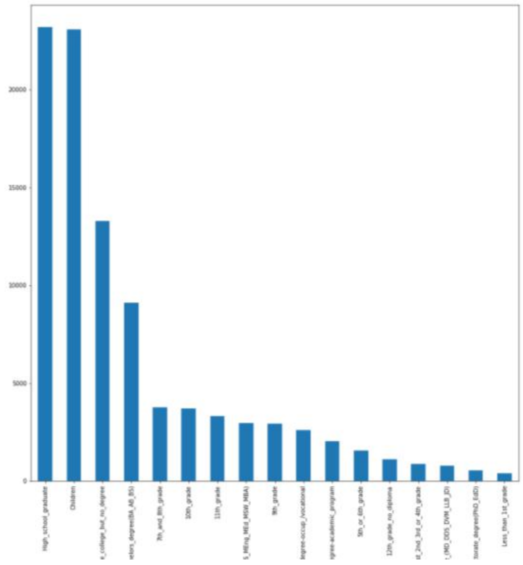

Figure 2. education

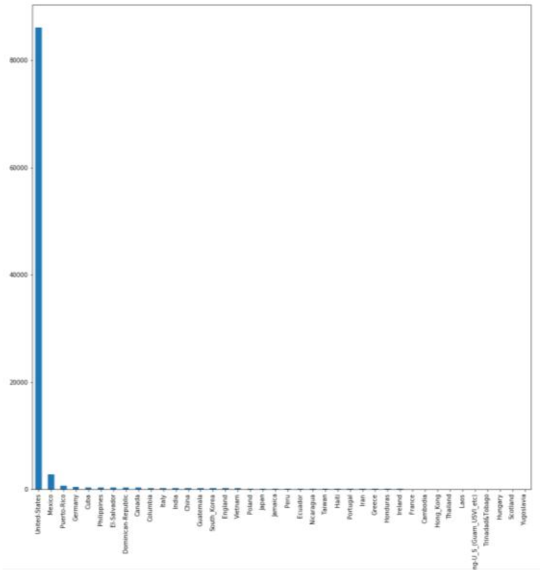

Figure 3. country_of_birth_self

When the attribute **country_of_birth_self** (non-anonymous) was set to "Yugoslavia," it was narrowed down to 28 out of 95130 cases, and when it was set to "Scotland," it was narrowed down to 33 cases. Eight nodes with attribute **age** (anonymous) existed, as presented below:

- 11.25419304676219:50.24556028129236
- 11.51152827104121:51.229290967271915
- 36.622210690192006:47.8680000254938
- 36.638232161874335:47.395535106915915
- 60.6195601310248:51.80496134138072
- 61.10766495456341:53.88034820224511
- 80.3943580886586:20.037342952946876
- 81.12478093235191:21.60588082302219

The attribute **country_of_birth_self** (non-anonymous) was narrowed down to "Yugoslavia," the number of age (anonymous) nodes was 7, as presented below. Compared with the case in which narrowing down was not performed, the number of nodes with attribute **age** (anonymous) was reduced by one:

- 11.25419304676219:50.24556028129236
- 36.622210690192006:47.8680000254938
- 36.638232161874335:47.395535106915915
- 60.6195601310248:51.80496134138072
- 61.10766495456341:53.88034820224511
- 80.3943580886586:20.037342952946876
- 81.12478093235191:21.60588082302219

When the attribute **country_of_birth_self** (non-anonymous) was narrowed down to "Scotland," the number of **age** (anonymous) nodes was 7, as presented below. Compared with the case in which narrowing down was not performed, the number of nodes with attribute **age** (anonymous) was reduced by one:

- 11.51152827104121:51.229290967271915
- 36.622210690192006:47.8680000254938
- 36.638232161874335:47.395535106915915
- 60.6195601310248:51.80496134138072
- 61.10766495456341:53.88034820224511
- 80.3943580886586:20.037342952946876
- 81.12478093235191:21.60588082302219

4.4. Consideration

From the experimental results, it can be concluded that the non-anonymized attributes include those render the narrowing down of data easy and those that make it otherwise. This indicates that there are attributes that cannot be anonymized to increase the degree of security. Since attributes that make it easy to narrow down data are those that contain data with a small number of distributions, it is safer to delete such attributes as identifiers when using the anonymization approach.

In this study, we consider an appropriate anonymization method when performing data mining including non-public information for a dataset whose information has already been published, such as SNS. The SNS data is a dataset that easily reflects individual tastes and behaviors. It is expected that various findings can be obtained by data mining. However, data analysis cannot be performed solely by the organization operating SNS. Moreover, there may be cases in which data analysis is outsourced to an external organization.

The data mining is not only data performed by data collectors; it can also be performed by sharing information with related items. In addition, cloud-based data mining environments have become easier to use and have made data mining conceivable. In these two cases, handling of all the data in a controlled environment is impossible. Thus, it is necessary to consider data anonymization. Through this experiment, it has been confirmed that some attributes need to be anonymized and that some do not contribute much to safety even if anonymized.

5. Randomization Decision Tree and Randomization Clustering

A study on privacy preserving data mining, privacy preserving decision tree learning using randomization has been proposed [1]. Here, we first introduce privacy preserving decision tree learning using randomization. Next, we propose privacy preserving clustering using the proposed randomization.

5.1. Privacy Preserving Decision Tree Learning

An example of research on privacy protection data mining, Agrawal and Srikant proposed privacy protection decision tree learning using randomization [1]. The general decision tree learning algorithm is as follows. The decision tree is a model of the tree structure for classification. The decision tree generation algorithm is shown in the Algorithm 1.

Algorithm 1: decision tree generation algorithm

```
Procedure Partition (Data $S$)
    (1) If all the data contained in $S$ belong to the same class, end
    (2) Evaluate how to divide by each attribute
    (3) Divide $S$ into $S_1$ and $S_2$ according to the best way to divide $S$
    (4) Execute Partition ($S_1$)
    (5) Execute Partition ($S_2$)
```

The goodness of division is evaluated by the gini index.

On the other hand, decision tree learning using randomization is being studied. This randomizes the input data. Randomization includes methods such as adding random noise to the data, exchanging the data randomly, and replacing the data with random values. After performing decision tree learning, the results of decision tree learning are restored by statistical inference. There are three algorithms to remove the influence of randomization when restoring the result: Global, ByClass, and Local. Global reconstructs each attribute at once at the start. ByClass reconstructs each attribute and each class at the start. Local reconstructs each attribute and each class at the time of classification evaluation of each node.

5.2. Privacy Preserving Clustering

We propose clustering using randomization. This is a method of adding random noise to the data during clustering. Global, ByClass, and Local are proposed as methods for adding random noise, with reference to previous research. Global reconstructs each attribute at once at the start. ByClass reconstructs each attribute and each class at the start. Local reconstructs each attribute and each class at the time in clustering steps. Local has the highest degree of privacy protection, but the problem is the increase in the amount of calculation due to randomization each time. Glocal does not have a high degree of privacy protection, but it requires less calculation and is easy to implement.

In the future, we are planning an experiment to verify the effectiveness of the proposed algorithm.

6. Conclusion and Future Works

In this study, we have taken up the issue on how to consider anonymization methods for datasets that contain not only publicly available information, such as SNS data, but also confidential information. With regard to this problem, Section 2 has described the types of data handled, the main methods used in privacy-preserving data mining, and the relationship of this research with privacy-preserving data mining.

In Section 3, we have proposed a method for evaluating the degree of anonymization when only a part of the input data was anonymized. Moreover, we considered the data distribution.

In Section 4, the degree of safety of anonymization for data in which only some data was anonymized was examined experimentally against the bias of data distribution.

Based on the experimental results, it has been confirmed that the non-anonymized attributes include those that render the narrowing down of data easy and those that make it otherwise. This indicates that there are attributes that cannot be anonymized to increase the degree of security. Since attributes that make it easy to narrow down data are those that contain data with a small number of distributions, it is safer to delete such attributes as identifiers when using the anonymization approach.

In Section 5, we introduced privacy protection decision tree learning using randomization. Next, we proposed privacy protection clustering using the proposed randomization.

7. References

[1] R. Agrawal and R. Srikant, “Privacy-preserving data mining,” SIGMOD Rec., vol. 29, no. 2, pp. 439–450, May 2000. http://doi.acm.org/10.1145/335191.335438 (Access Date: 15 January, 2021).

[2] C. C. Aggarwal and P. S. Yu, A General Survey of Privacy-Preserving Data Mining Models and Algorithms. Boston, MA: Springer US, 2008, pp. 11–52. https://doi.org /10.1007/978-0-387-70992-5 2 (Access Date: 15 January, 2021).

[3] J. Sakuma and S. Kobayashi, “Privacy-preserving data mining,” The Japanese Society for Artificial Intelligence, vol. 24, no. 2, pp. 283–294, mar 2009.

[4] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam., (2006). L-diversity: privacy beyond k-anonymity, in 22nd International Conference on Data Engineering (ICDE’06), April, pp. 24–24.

[5] L. Sweeney., (2002). k-anonymity: A model for protecting privacy, International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, vol. 10, no. 05, pp. 557–570. http://www.worldscientific.com/doi/abs/10.1142/S0218488502001648 (Access Date: 15 January, 2021).

[6] R. Cramer, I. Damg˚ard, and J. B. Nielsen, (2001). Multiparty computation from threshold homomorphic encryption, in Advances in Cryptology - EUROCRYPT 2001, B. Pfitzmann, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, , pp. 280–300.

[7] Utd anonymization toolbox. http://www.cs.utdallas.edu/ dspl/cgi-bin/toolbox/index.php (Access Date: 15 January, 2021).

[8] M. Lichman, “UCI machine learning repository, (2013). http://archive.ics.uci.edu/ml(Access Date: 15 January, 2021).

8. Acknowledgements

This research work is supported by JSPS KAKENHI Grant Number JP17K00310.