Electronic Preprint for Journal of Information Processing Vol.29

Regular Paper

Privacy Risk of Document Data and a Countermeasure Framework

Tomoaki Mimoto[1,a)] Masayuki Hashimoto[1,b)] Shinsaku Kiyomoto[2]
Koji Kitamura[3] Atsuko Miyaji[4,5]

Received: March 8, 2021, Accepted: September 9, 2021

Abstract: A huge number of documents such as news articles, public reports, and personal essays have been released on websites and social media. Once documents containing privacy-sensitive information are published, the risk of privacy breaches increases, thus requiring very careful review of documents prior to publication. In many cases, human experts redact or sanitize documents before publishing them; however, this approach can be inefficient with regard to cost and accuracy. Furthermore, such measures do not guarantee that critical privacy risks are eliminated from the documents. In this paper, we present a generalized adversary model and apply it to document data. This work devises an attack algorithm for documents using a web search engine, and then proposes a privacy-preserving framework against the attacks. We evaluate the privacy risks for actual accident reports from schools and court documents. In experiments using these reports, we show that human-sanitized documents still contain privacy risks and that our proposed approach can contribute to risk reduction.

Keywords: privacy, document data, sanitization

1. Introduction

1.1 Background

Personal data are essential for building an efficient and sustainable society, but they must be carefully handled according to the sensitivity of the data. In contrast to security, a key challenge of preserving privacy in personal data is that an attacker could be mistaken for a particular "authorized user." Therefore, it is important to maintain a balance between privacy and utility in each use case, and anonymization techniques for achieving an optimized balance have been studied extensively. Several anonymization methods have been optimized for specific types of personal data such as medical records. Governments, public offices, and enterprises exchange or publish huge numbers of documents containing personal data. Citizens have a right to request governmental information. Moreover, in several healthcare organizations, medical data are utilized for epidemiological research and disease prevention. These responsible organizations require appropriate protection measures to be defined due to the confidentiality and sensitivity of these information. These protective measures are generally performed by humans, and no systematic rules have been developed for automatic analysis. Furthermore, the impact on privacy through the release of such documents is not uncommon, and some risk to privacy still persists even after the documents have been sanitized. To solve these problems, analysis methods of privacy risks and sanitizing algorithms for documents need to be incorporated in the publication of documents that contain privacy-sensitive information. Researchers have proposed several privacy models in order to analyze privacy risks, and they have also developed anonymization algorithms to reduce potential privacy risks [2], [3], [4], [5], [6], [7], [8]. However, almost all of them are intended for structured data, such as datasets in relational databases, and few studies have focused on unstructured data (e.g., document data).

There are many privacy-preserving techniques for structured data. K-anonymization [2], [9] is a well-known technique for input privacy. K-anonymity is ensured by the fact that in each published record, every combination of values of quasi-identifiers is matched to at least k respondents. Furthermore, k-anonymization has been used as a technique for transforming a dataset into its anonymized dataset satisfying k-anonymity [10], [11]. Many types of extended privacy metrics have also been proposed [4], [5]. Differential privacy [3], [12] is another state-of-the-art privacy model, and it is based on the statistical distance between two datasets differing by at most one record. Differential privacy does not need to define the attacker's knowledge, and the privacy risk is mathematically proven in a certain attack model that sends queries to the dataset. However, document data are unstructured and contain risk words, where the risk depends on the contexts of the document data. Consequently, a new approach to preserving privacy is needed to reduce the risk posed by publishing document data.

1 Advanced Telecommunications Research Institute International, Sorakugun, Kyoto 619–0288, Japan
2 KDDI Research, Fujimi, Saitama 356–8502, Japan
3 National Institute of Advanced Industrial Science and Technology, Koto, Tokyo 135–0064, Japan
4 Osaka University, Suita, Osaka 565–0871, Japan
5 Japan Advanced Institute of Science and Technology, Nomi, Ishikawa 923–1292, Japan
a) to-mimoto@atr.jp
b) masayuki.hashimoto@atr.jp

© 2021 Information Processing Society of Japan

1.2 Contribution

In this paper, we focus on privacy risk analysis for document data. An attacker has a target record (or a dataset), a processed dataset, an attack algorithm, and a risk evaluation subroutine. This paper is the modified version of Ref. [1], and we modified the algorithm and performed additional evaluations. Specifically, we introduce a preprocessing part and adjust the algorithm to improve the processing speed. The aim of an attacker in this paper is to re-identify a person relevant to documents. If a person is re-identified, an attacker can link the person to additional information in the document and it causes a serious breach of privacy.

Our proposal provides the following contributions:

- We define an actual adversary model for document data and propose a realistic web-based attack algorithm. The proposed model is an extension of the general adversary model described in Section 3. In comparison with other models [13], [14], the aim of an attacker in this model is to find words that are missed by sanitizers and to re-identify a person related to a document. Our proposal does not need to define or designate sensitive words as opposed to previous works, and it also does not need to evaluate relationships among sensitive words and words in documents. Hence, we can handle any type of document without considering the length of articles.
- The proposed privacy-preserving algorithm operates as a web-based attack algorithm. The attack algorithm searches for words with a high volume of information and searches articles related to the document on the internet by using those words as keywords. If the articles contain additional information, the privacy-preserving algorithm sanitizes the keywords. This algorithm is able to find words that pose the risk of re-identification by combining other words, thus reducing the privacy risk.
- We apply the attack algorithm to two actual Japanese document datasets, which were previously sanitized manually, and we verify that the algorithm works effectively in both cases. The results of the experiments show that manual sanitization is not sufficient, and the proposed algorithm reduces re-identification risks by sanitizing risk words that have risks of re-identification.

1.3 Organization

The rest of the paper is organized as follows. First, we introduce related works in Section 2 and clarify issues of document sanitization. Then, Section 3 presents a generalized adversary model. In Section 4, we structure document data to apply the adversary model; then, we present a document sanitization algorithm called `DocumentSanitization`. Section 5 describes the document data and explains the idea of an actual privacy evaluation. Section 6 presents empirical results obtained by the algorithm for actual documents. The paper is concluded in the last section.

2. Related Work

There are many types of data that contain privacy information, such as location data, medical data, and purchase history data. Such document data as judicial records and medical records also contain privacy information. Previous works [13], [14], [15] attempted to evaluate risk or sanitize document data. In one study [14], documents were defined as being composed of entities and terms, and each entity was set based on related terms. In the case of a compendium of diseases, for instance, each disease is an entity, and the items pertaining to its context, including its symptoms and the drugs used to treat the disease are terms. In this proposal, the entities to be protected are determined beforehand. Under this assumption, the authors propose the idea of K-safety, which is similar to k-anonymity, as well as an algorithm to achieve K-safety. The anonymization algorithm sanitizes the terms so that more than K entities are inferred from the terms. In another work [15], a sanitization tool was proposed, and this tool had certain functions: understanding the contents of a sentence by leveraging linguistic content analysis and understanding the sensitivity of the content in general by using inference detection algorithms. Through their interviews and feedback, the authors claimed that sanitization is an alternative approach in which a document is revised to hide sensitive content while retaining as much cohesion and utility as possible. In a further work [13], sensitive words, such as AIDS, were defined beforehand, and words that have high relevance to the sensitive words were sanitized. In contrast to other approaches [16], [17], the authors proposed a method to automatize the detection of terms that may disclose sensitive data to secure their protection. These previous works checked pre-defined sensitive words in a document as well as words that have high relevance to the sensitive words, and their experimental results show that the algorithms try to emulate manual sanitization, although their accuracy is less than that by manual sanitization.

Moreover, combinations of general words, which seem to have no relationship at first glance, sometimes cause privacy leakage. Therefore, manual sanitization may miss privacy risk on a document. For example, the two general terms “football lesson” and “flood” may lead to re-identification of a victim in an accident. An accident during a football lesson or an accident by flood can sometimes occur in isolation. However, an accident by flood during a football lesson is very rare, and when sanitization is conducted manually, the words “football lesson” and “flood” can be missed because these words are so common. Therefore, algorithms that imitate manual sanitization may be insufficient from the viewpoint of de-identification. In our experiments, we actually confirmed that a person in such a situation could be re-identified using those methods. Furthermore, in cases where the sanitizer can only access partially sanitized documents, the previous algorithms may not work effectively, because it is impossible to evaluate pointwise mutual information (PMI).

3. Generalized Adversary Model

In this paper, we first generalize an adversary model for re-identification. Let an original dataset be D, an anonymized dataset, namely, the target dataset for evaluation, be $\mathcal{A}(D) = d = \{d_1, \ldots, d_n\}$, and $d_i = \{A_i, g(R_i)\}$. Here, R_i is the set of information containing re-identification risk r_{ij}, which are abused by an attaker and should be protected, and $g(\cdot)$ is an anonymization

© 2021 Information Processing Society of Japan

function. A_i is the other information. We consider an attack simulator $\mathcal{A}^*$ and an ideal dataset D^*. Here, D^* may be the original dataset $\mathcal{A}^{-1}(d) = D$ or other datasets including the information of D, namely, $D^* \supseteq D$. An attacker who has d_i wants to know additional information $R_i^* \supseteq R_i$.

Definition 1 *Generalized Adversary*: Let a target dataset for risk evaluation be $d = \{d_1, \ldots, d_n\}$, an attack simulator be $\mathcal{A}^*$, and an ideal dataset be D^*. An attacker has d, $\mathcal{A}^*$, and D^* is called a "generalized adversary."

The goal of the generalized adversary is to identify a participant of d_i or to learn additional information of the person from d_i by comparing d with D^*. Here, the identifiers of record i such as the user name and user ID, are included in R_i. In other words, the goal of the generalized adversary is to compare an output of $\mathcal{A}^*(d_i, D^*)$ with R_i and obtain additional information of d_i.

Definition 2 *Generalized Privacy Metrics*: Let a target dataset for risk evaluation be d, an attack simulator be $\mathcal{A}^*$, and an ideal dataset be D^*. We define the privacy risk of a record $d_i \in d$ against a generalized adversary as

$$E(\mathcal{A}^*(d_i, D^*), R_i), \tag{1}$$

where $E(\cdot)$ is an evaluation function.

We can easily expand this definition for the risk of target dataset d. For instance, the privacy risk of a dataset d against a generalized adversary can be expressed as

$$\max_i(E(\mathcal{A}^*(d_i, D^*), R_i)). \tag{2}$$

The generalized privacy metrics include major privacy metrics. In the case of k-anonymity, for example, $D^* = D$, $R_i = QI_i$, which is the set of quasi-identifier, and $d = \mathcal{A}(D)$, where $\mathcal{A}(\cdot)$ is a k-anonymization algorithm. $\mathcal{A}^*(d_i, D^*)$ links d_i to $d_j^* \in D^*$ and outputs a suspicious record set, while $E(\cdot)$ outputs the number of given datasets. When $\mathcal{A}(\cdot)$ works well, the privacy of each record is more than k, namely $\max_i(E(\mathcal{A}^*(d_i, D^*), R_i)) \geq k$, and it is the same estimated value as k-anonymity. Many privacy metrics have been proposed, and some of them use a target record (or dataset), another dataset, an anonymization algorithm, and an evaluation function; accordingly, they can be expressed as the generalized privacy model.

4. Framework for Document Sanitization

4.1 Overview of Document Sanitization

In this section, we provide a framework for document sanitization. The framework is assumed to be employed when a user, who is the reporter or who has the sanitization authority, checks the document privacy (hereinafter, the users are called "sanitizers"). There are some anonymization techniques such as generalization and data deletion, and we call the anonymization techniques "sanitization" collectively. This framework consists of three parts: preprocessing, simulated attack, and sanitization. We assume a real attack model for document data based on the generalized adversary model and propose a privacy preserving algorithm against this attack. We apply a real attack situation to the generalized adversary model and also define the preconditions.

First, the preprocessing part provides data structuring. Documents are not structured, and thus it is difficult to handle them. Therefore, we divide documents d into word sets w. Furthermore, this part calculates the amount of information of each word for the next part. Preprocessing also classifies input dataset subsidiarily. A simulated attacker accesses the internet and searches for documents related to a target document. Hence, the risk for the attack depends on the degree of interest of a target document. For example, big disasters are featured in the news, and documents related to catastrophes can be easily found on the internet compared with minor accidents. Preprocessing tags the target dataset, and the attack part utilizes this information. If a sanitizer is a data holder, namely, the sanitizer is not a sanitization proxy, and has an additional dataset D^{**}, the dataset is available as optional input. The attack part provides a simulated attack on a target document d_i and makes a list of searched documents related to the target. A simulated attacker searches the internet using words with a large volume of information and outputs a document list $List_i$, which can be found on the web. The sanitization part first evaluates the documents in $List_i$. If there are documents including the name of a person, we consider the keywords used for the web search have a risk of re-identification. In this paper, we refer to the words that have risks of re-identification including identifiers as "risk words". Finally, sanitization sanitizes the keywords and outputs a sanitized document d_i'.

Algorithm 1 `DocumentSanitization`$(d, (D^{**}))$: Document sanitization framework.

Input: A target document dataset $d = \mathcal{A}(D)$ (and a document dataset $D^{**} \supseteq D$.)
Output: A sanitized document dataset d'

1: $(w, I(w), Label) \leftarrow$ `Preprocessing`$(d, (D^{**}))$
2: $List \leftarrow$ `Attack`$(w, I(w), Label)$
3: $d' \leftarrow$ `Sanitization`$(d, List)$
4: **return** d'

4.2 Adversary Model for Document Data

We consider a real attack model for document data based on the generalized adversary model and propose a privacy preserving algorithm against the attack. This framework is assumed to be used when a sanitizer, who is the reporter or who has sanitization authority, checks the document privacy. A sanitizer is assumed to have target documents $\mathcal{A}(D) = d = \{d_1, \ldots, d_n\}$. d is a set of documents d_i, and $d_i = \{A_i, g(R_i)\}$. Here, $g(\cdot)$ is a generalization function and R_i is a set of information containing re-identification risk that may include such information as the name of the person associated with document d_i. This assumption is that the sanitizer is not a data owner, namely, that the sanitizer is commissioned to sanitize a dataset, and thus the sanitizer does not have D. In this paper, for simplicity, we regard a sanitizer as having only d, but the sanitizer may use another dataset $D^{**} \supseteq D$ to calculate the amount of information on words and to tag documents. Furthermore, if the sanitizer is the data owner, the sanitizer can evaluate and sanitize D. On the other hand, an attacker is assumed to have a sanitized dataset d and access to D^*. D^* is an ideal dataset and includes a part of R_i. In this paper, D^* is a document dataset on the web, and a simulated attack $\mathcal{A}^*$ is used for linking d_i to $d^* \in D^*$ and obtaining R_i from d^*. This attack is valid when a

© 2021 Information Processing Society of Japan

target document d_i is related to events that can be known publicly, and we focus on documents related to accidents that occur at schools and court documents. It is easy for human beings to judge whether a word in d^* is r_i, so, the evaluation function can be very simple. If the attack succeeds, namely, an article about d_i exists on the web and includes additional information such as the name of the victim, $E(\mathcal{A}^*(d_i, D^*), R_i) = 1$, and 0 otherwise. However, the evaluation should be mechanical and exclude human factors. When a sanitizer has D, a simulated attacker knows that $R_i \in D_i$ and $E(\mathcal{A}^*(d_i, D^*), R_i)$ can be evaluated. However, a sanitizer may not have D because some data owners commission an outside agency to evaluate the risk of sanitized datasets. Therefore, we need to consider a more flexible evaluation function with caution. The evaluation function is discussed in Section 5.

4.3 Preprocessing Part

Documents are not structured, and some processing is thus needed to handle them. In the preprocessing part, a morphological analysis is applied in the first step, and we define documents as a set of words following the practice of previous research [13], [14]. More precisely, we define a document i including m words as $d_i = w^i = \left\{w^i_1, \ldots, w^i_m\right\}$ (denoted as $w^i_j = w_j$ for simplicity). Some words have risk of re-identification, and without loss of generality, we denote $A_i = \{w_1, \ldots, w_l\}$, $R_i = \{w_{l+1}, \ldots, w_m\}$. The preprocessing algorithm, then, runs *CalculationI* to calculate the amount of information of each word $I(w) = \left\{I(w^1_1), \ldots, I(w^n_m)\right\}$. *CalculationI* requires word set $w^i = \{w_1, \ldots, w_m\}$ and d and calculates the volume of information of each word $w_j \in d_i$. The volume of information of w_j is defined as

$$I(w_j) = -\log P(w_j) + \epsilon(w_j). \tag{3}$$

Algorithm 2 `Preprocessing`$(d, (D^{**}))$: Preprocessing for document sanitization.

Input: A target document dataset $d = \mathcal{A}(D)$ (and a document dataset $D^{**} \supseteq D$.)
Output: A word set w, information content $I(w)$, and *Label* for each document.

1: **for** $i < n$ **do**
2: $w^i \leftarrow MorphologicalAnalysis(d_i)$
3: $I(w^i) \leftarrow CalculationI(w^i, d)$
4: $Label_i \leftarrow Labeling(d_i, d)$
5: **end for**
6: **return** $w, I(w), Label$

Here, $P(w_j)$ is the appearance probability of w_j, and $\epsilon(w_j)$ is a moderator variable. In general, $P(w_j)$ is calculated as $P(w_j) = \frac{\#d_i(w_j)}{\#D^{**}}$, where $\#D^{**}$ is the number of words in D^{**} and $\#d_i(w_j)$ is the number of $w_j \in d_i$. However, when w_i does not appear in D^{**}, $P(w_j) = -\log 0 = \infty$ and the moderator variable $\epsilon(w_j)$ does not work even if $\epsilon(w_j)$ is very large. Therefore, we define $P(w_j)$ as

$$P(w_j) = \begin{cases} \frac{\#d_i(w_j)}{\#D^{**}} & (w_j \in D^{**}) \\ \frac{\#d_i(w_j)}{\#D^{**} + \#d_i} & (w_j \notin D^{**}), \end{cases} \tag{4}$$

where $\#d_i$ is the number of words in d_i and $\epsilon(w_j)$ is the parameter for selecting appropriate words. Tuning is generally required based on data in natural language processing (NLP), and it is expressed as $\epsilon(w_j)$ in this paper. More precisely, we give weight to parts of speech. Since prepositions, conjunctions, and adjectives are rarely used in a web search, we give weight to nouns and verbs. Furthermore, we give a negative weight to nouns and verbs with a similar meaning to words. For example, we consider a document and find that $-\log P(w_j)$ of w_j = "first" and w_j = "second" are high. They may be regarded as candidates of risk words when $I(w_j) = -\log P(w_j)$, but they have similar meanings. Hence, $I(w_j)$ of one of them decreases by adjusting $\epsilon(w_j)$ and is excluded from the candidates. For the other words, we adjust $\epsilon(w_j)$ so as to find words that have different vectors. Furthermore, the preprocessing part applies a classification protocol and tags to each document d_i. The classification is optional but has benefits in terms of computation cost and accuracy. A simulated attacker in the framework accesses documents on the website and compares them with a target document, so the degree of interest of the target document is relative to the privacy risk. This attack is made repeatedly, and this is the bottleneck. The preprocessing part classifies documents from the standpoint of interest, and documents with a high interest label are attacked intensively for efficiency. These tasks require d, and the accuracy is expected to improve by using $D^{**} \supseteq D$. In this paper, we focus on reports of accidents at a school and court documents. When a target document is sensational, such as a murder, a greater weight assigned to the document.

4.4 Attack Part

Algorithm 3 `Attack`$(w, I(w), Label)$: Web-search attack algorithm.

Input: A word set w with the volume of information $I(w)$, and Label
Output: A searched document set *List*

1: **for** $i < n$ **do**
2: $KW_i \leftarrow SetKeywords(w^i, I(w^i), Label_i)$
3: $List_i \leftarrow WebSearchAttack(KW_i)$
4: $List = List \cup List_i$
5: **end for**
6: **return** $List$

The attack part follows the preprocessing part. The attack algorithm in this part inputs a word set $w = \left\{w^1, \ldots, w^n\right\}$ and additional information $I(w) = \{I(w^1), \ldots, I(w^n)\}$. w^i is a word set of document d_i, and $I(w^i) = \{I(w_1), \ldots, I(w_l), I(g(w_{l+1})), \ldots, I(g(w_m))\}$ represents the amount of information of words. The attack algorithm, `Attack`, inputs a word set $w = \left\{w^1, \ldots, w^n\right\}$ with the volume of information $I(w)$ and *Label*. Here, $Label = \{Label_1, \ldots, Label_n\}$ is the tag set of documents, and when the preprocessing algorithm does not run the classification protocol, $Label_i = \phi$. Moreover, some parameters are actually required, but we omit them here for simplicity. The attack algorithm calls *SetKeywords* and *WebSearchAttack*. *SetKeywords* requires w^i, $I(w^i)$, and $Label_i$ and outputs a set of words that have a large volume of information KW_i. Subsequently, the attack algorithm calls *WebSearchAttack*. *WebSearchAttack* searches for documents related to d_i using $kw \in KW_i$ and returns $List_i = \bigcup(List_i(kw), kw)$. $List_i(kw)$ is the set of documents found using kw as keywords. Note that we have to limit the number of documents to search for and the number of kw sets. Regarding the document search, the

© 2021 Information Processing Society of Japan

number of documents has little effect on the run time and the evaluation result. We fix $\#List(\cdot) = 10$, which represents the number of web pages that are displayed in a web browser at one time. Moreover, kw generates the $2^{|KW_i|} - 1$ combinations, and the run time of the algorithm strongly depends on $|KW_i|$. This must be a parameter of the algorithm, and in this paper, we set $|KW_i| = 3$. After that, we only need to check whether $r_{ij} \in R_i$ is included in the $List_i$ to confirm that the attack has succeeded. If $r_{ij} \in R_i$ is included in $List_i$, document i is at risk and an attacker may obtain additional information about d_i.

In this paper, we focus on documents about accidents that occurred at a school. One document includes the exact date and time of the accident, the place, gender, grade, and the name of the student victim, his/her medical history, the compensation value of the accident, etc., in addition to a report on the accident situation. The details of the dataset are stated in Section 5. The name of the victim in each accident was deleted from the documents we received. Accidents at a school are often included in news documents, and the name of the student is often included in a news document. Therefore, we assume that D^* represents news articles on the web, and we define the attack as searching for documents containing the name of the victim on the internet. The metadata, which is information incidental to a document such as medical history, has utility value, but if the name of a student relevant to the document is revealed, the metadata and the student are linked, and consequently, the impact of privacy violation is serious.

Algorithm 4 `Sanitization`$(d, List)$: Privacy-preserving algorithm for documents using an attack simulator.

Input: A target document dataset d and a document set $List$.
Output: A sanitized document dataset d'.

1: **for** $i < n$ **do**
2: **if** $RiskEvaluation(d_i, List_i) == 1$ **then**
3: $RiskWords_i = RiskWords_i \cup kw$
4: **end if**
5: $List \leftarrow List - List(kw)$
6: **end for**
7: $d'_i \leftarrow Reconstruct(d_i, RiskWords_i)$
8: **return** $d' = \bigcup d'_i$

4.5 Privacy-preserving Algorithm Based on an Attack Algorithm

Finally, we construct an algorithm to sanitize risk words by using the `Attack` algorithm. The sanitization algorithm requires a target document dataset d and $List$. $List_i \in List$ includes keywords kw and documents found on the web. In the sanitization algorithm, *RiskEvaluation* and *Reconstruct* are executed. *RiskEvaluation* compares d_i with the documents included in $List_i$ and then extracts risk words. *Reconstruct* removes or generalizes the words in d_i and the sanitized document d'_i is output. Subsequently, the privacy risk of each document is calculated by *RiskEvaluation*, the function of which is to evaluate the privacy risk of kw. When a document $d_{search} \in List_i(kw)$ includes words $r_{ij} \in R_i$, kw violates the privacy of d_i and *RiskEvaluation* returns 1. The kw is input in $RiskWords_i$, and finally this algorithm runs *Reconstruct* and outputs a sanitized document d'_i. An overview of our model is shown in **Fig. 1**.

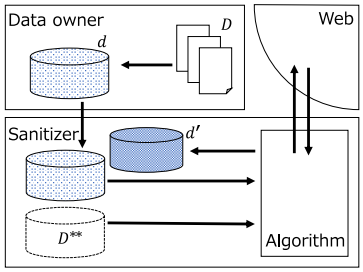


Fig. 1 Overview of the model.

5. Datasets for Experiments

5.1 Accident Documents

The target documents considered first in this paper are accident documents from a school. We perform morpheme analysis of the documents using Mecab [*1] and define all words that appear in the results as A_i. These documents are owned by the Japan Sport Council (JSC), and we can use (nearly) original document data D and sanitized data $\mathcal{A}(D) = d$. The original data D have the exact date and time of the accident, the location, the diagnosis (or cause of death), the gender and grade of the student victim, his/her medical history, and the compensation value of the accident, etc., in addition to a report on the accident situation. The name of the victim is clearly included in the original data, but the dataset we received does not include such information. On the other hand, d has almost the same information as D, but the report on the accident situation is sanitized. Note that some sanitized data are open to the public by JSC [*2], and the original data are processed to obscure the sensitive information. Not only are the document data sanitized, but metadata are also generalized and deleted, such as the date and time of the accident, medical history, and the compensation value of the accident. The published data do not include sensitive information, such as medical history and the compensation value of the accident, so even if the name of a victim is revealed by the attack (with this information being publicized by a news report), the attacker cannot obtain additional information from the open data. However, the lack of information leads to a lack of value of the data. For instance, some research institutes review these accident data, analyze the scale of the accident, and use the results to prevent a future accident. In this case, the medical history information and the compensation value of the accident could be useful. Therefore, the linkability between the data and the name of the victim should be reduced to protect the victim's privacy, and at the same time, the utility of the data should be maintained. We have approximately 700 original fatal accident documents (OADs) and more than 4,000 sanitized accident documents (SADs). All sanitized versions of the OADs are included in the SADs. We reviewed the documents on a contractual basis for research purposes, and they are not published. On the other hand, the sanitized documents were manually sanitized by staff members of the JSC and are disclosed online.

[*1] https://github.com/neologd/mecab-ipadic-neologd
[*2] https://www.jpnsport.go.jp/anzen/anzen_school/anzen_school/tabid/822/Default.aspx

© 2021 Information Processing Society of Japan

5.2 Court Documents

The other targets in this paper are court documents in Japan. In particular, murder cases were used in our experiments because these incidents tend to be broadcasted/published as news in a public domain such as the internet. Accordingly, our attack algorithm can easily capture the information related to the documents. As in the experiments on the accident documents, we performed a morpheme analysis of the documents using Mecab and defined all words in the overview section as A_i. We collected these documents from a website [*3]. In contrast to the accident documents, suspects, victims, and other information such as their age are anonymized, but the degree of anonymization of these documents is lower than that of the school data. We denote all court documents as sanitized court documents (SCDs). Contrary to accident documents, court documents do not have metadata, so we define the top 20 words with a large volume of information as metadata. For experiments, we downloaded 30 documents concerning murder and 1,000 documents randomly to calculate the volume of information.

5.3 Risk Words of Documents

A definition of risk words is required in order to estimate the privacy risks of documents. In this paper, the objective of an attacker of documents is to reveal the names of the people vulnerable to privacy breaches. R_i needs to be a set of words w_j which are linked to public news about accidents and incidents on websites. A sanitizer has an original dataset with the name of a victim, and R_i is the name of the victim when the sanitizer uses our algorithm; however, we need to reproduce the attacker's behavior for an attack algorithm. A simulated attacker is assumed to have a sanitized dataset d and access to D^*. The risk words of d are $g(R)$ and may not include the name of the victim. We first manually checked each document and searched for the names of the people relevant to our experiments by using *SetKeywords* and extracting the candidates keywords. Then, we searched the internet and set the name as R_i when we found the target name in the searched documents. It was also manually judged whether the name is correct. In the first experiment, *RiskEvaluation* returns 1 when a document $d \in List(kw)$ contains words $r_{ij} \in R_i$, implying that it also contains the tagged name of a relevant person, which is defined as r_{ij}. On the other hand, if a sanitizer does not have the name of a relevant person, this process takes huge computational time as well as a manual search of the risk documents. Consequently, we need some indices for automation. We focus on the number of words that are included in the metadata of each document. We assume if the number of words appearing in both a searched document and the metadata increases, the possibility that the two documents indicate the same accident will also increase. In the case of court documents, no document has metadata, but there are many words including a large volume of information, and thus these words are handled as metadata. After calculating the volume of information of each word and setting kw, we set words with a high volume of information other than kw as R_i. The validity of this assumption is considered in the following section.

[*3] https://www.courts.go.jp/app/hanrei_jp/search1

6. Experiment

6.1 Experiments on Accident Documents

We use both the original documents and the sanitized documents in the experiments for accident documents. We first analyze the risk of the OADs; then, we apply our algorithm to SADs and confirm that some privacy risk of identifying the victims remains in the SADs. We set 700 OADs as D, the corresponding 700 SADs as d, the other 1,600 SADs as D^{**}, and the victim's actual name as R_i. `Preprocessing` applies the morpheme analysis to d and evaluates $I(w_j)$ for each word. In this experiment, we obtained $\#d \fallingdotseq 150{,}000$ as the result of the morpheme analysis. The words with the highest value are $\#d_i(w_j) = 1$ and $I(w_j) = 23.80$, followed in order by 22.79 and 22.21. *SetKeywords* outputs KW_i, which is the set of words $w_j \in d_i$, where it has a large volume of information. We focus on the words having the top three volumes of information, namely the words w_j s.t. $I(w_j) \geq 22.21$. In the experiments, the words having the top three volumes of information are named potential risk words (PRWs) to compare OADs and SADs. We classify the words w_j into three classes: (1) words that appear only in $D_i \in D$; (2) words that appear in both D_i and d_i, i.e., that are not sanitized manually; and (3) words that have a risk and appear in d_i, that is, words that are keywords leading to the acquisition of the name of the victim by web search (**Fig. 2**). There are $93 + 40 + 36 = 169$ PRWs in the OADs (1), and we find 35+20+21 = 76 PRWs remaining in the corresponding SADs (2). This result shows that $169 - 76 = 93$ PRWs are sanitized in SADs, but nevertheless many PRWs remain in the SADs. Then, we run *SetKeywords* and launch *WebSearchAttack* using the PRWs in the `Attack` algorithm. KW_i is input to *WebSearchAttack*, and it outputs a set of searched documents $List_i$. The set of searched documents is input to *RiskEvaluation*, and it outputs 1 and the keyword *RiskWords*, where the set includes the risk words (e.g., the name of the victim). In the experiment, the *RiskEvaluation* algorithm found 12 out of 76 PRWs that are linked to the name of the victim. $RiskWords_i$ is a word set classified into Class (3).

The experimental results focusing on PRWs are listed in **Table 1**. Here, we can see 93 (= Class (1)\Class (2)) PRWs are sanitized manually. However, the other 76 (= Class (2)) words are not sanitized and remain. In the OADs, we found many proprietary nouns such as facility names are regarded as PRWs and they are all sanitized in the SADs. However, some events such

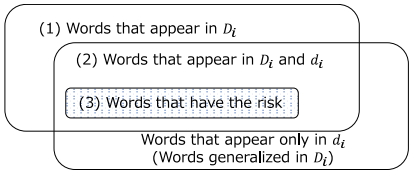


Fig. 2 Relationship of each word class.

Table 1 Relationship between $I(w_j)$ and the class of words.

$I(w_j)$	Class (1)	Class (2)	Class (3)
23.80	93	35	9
22.79	40	20	3
22.21	36	21	0
#PRWs	169	76	12

© 2021 Information Processing Society of Japan

as "plane accidents" that are PRWs are not sanitized even in the SADs and are published. The attack simulator in the algorithm abuses the PRWs and 12 (= Class (3)) of them link to actual articles that include the names of the victims. As a result, it is clarified that attackers can re-identify persons relevant to documents and link other information such as medical histories to the re-idenfied ones. In conclusion, the risk of identification still exists in manual sanitization, but our algorithm efficiently detects risk words missed by manual sanitization. Finally, *Reconstruct* removes or generalizes the 12 words from d and outputs d'.

Our algorithm performs a simulation attack and deletes only high-information words that cause the re-identification of a person relevant to a document. In this experiment, there were 700 documents that were sanitized manually and we found 12 words that cause re-identification. As mentioned before, we focus on the risk of re-identification, which is a critical issue for privacy, and we can minimize the decrease of utility due to deletion or generalization. The words we sanitized can recover the linkability between a person and a document and it must be a critical issue. Thus, even if an analyst demands to maintain utility, the words must be sanitized at least. Therefore, we can say our algorithm preserves privacy while maintaining utility. Furthermore, we found almost all of the words classed in Eq. (3) in the experiment have top three volume of $I(w_j)$, namely $I(w) \leq 22.21$. Therefore, the large number of KW_i does not mean a strong privacy protection, and even if $|KW_i|$ is not large, we can prevent a re-identification attack enough.

6.2 Experiments on Court Documents

In the case of court documents, we collected sanitized documents on the web, but due to contractual issues, we could not obtain the original documents D. This is not a special case, so a sanitizer needs to define *RiskEvaluation* carefully. We perform an experiment to confirm whether our algorithm successfully attacks SCDs. As mentioned above, we set 30 SCDs concerning murder as d, the other 1,000 SCDs as D^{**}, and the name of the persons relevant to d_i as R_i. We set $KW_i = 3$, so that the algorithm would select three words from each document as Class (2) words. *WebSearchAttack* searches the articles in $List_i$ using these words. We compared the words included in $List_i$ with R_i, and the words that link to the relevant persons are assigned to Class (3). The results are shown in **Table 2**. The words with a large volume of information are at high risk of being identified, and it is also confirmed that there are some risk words in the documents, even if the documents are manually sanitized, as in the case of the accident documents.

Note that the parameter $\epsilon(w_j)$ was changed from the previous experiment due to the difference between accident documents and court documents. We place high priority on the person's name in OADs, but we can reduce the value of the person's name in SCDs because court documents include the judge's name. Furthermore, we increase the importance of the age of the person in SCDs. The victims in accident documents are all students, and even if the age of a victim were included, the information would not be worth much in terms of re-identification. On the other hand, the age of suspects and victims in articles might be effective identifiers of the relevant person. In this way, some tuning is needed depending on the document type, and a sanitizer can easily optimize the policy of the parameter setting d.

Table 2 Relationship between $I(w_j)$ and the class of words.

$I(w_j)$	Class (1)	Class (2)	Class (3)
31.50	–	41	14
31.21	–	28	8
30.83	–	15	0
#PRWs	–	84	22

6.3 Labeling Option

`Preprocessing` provides a labeling option. The attack accuracy depends on the interest of the target document, and so we assign an interest label to target documents. In this section, we confirm the effectiveness of the labeling option. We use FastText [18], which is published as open source by Facebook AI Research, to classify documents. FastText handles words as a vector, as does Word2Vec [19], and classifies documents at high speed. Of our 2,300 SADs, we set half of them as training data. We assigned two labels, fatal accident and non-fatal accident, to the training data and then inferred the labels of the other half of the documents. The results (**Table 3**) show that documents can be classified with high accuracy despite whether they are about fatal accidents or non-fatal accidents. In our experiments, the attacks on non-fatal accident documents did not succeed, and *SetKeywords* could change the parameter KW_i depending on the label. The labeling mechanism should vary by document type, but it is expected to improve the algorithm in terms of utility and processing speed by introducing variable parameters according to the labels.

6.4 Relationship between the Number of Metadata and the Risk of Re-identification

In this paper, we linked the name of the victim and the corresponding documents beforehand and generated an evaluation function, where this function outputs 1 when it finds a relevant name from the searched document and 0 otherwise. However, this function can be generated by sanitizers that have the name of the victim and we need to consider the case where a sanitizer does not know this name. For example, when a data owner does not have sanitization skills, this owner might conclude an agreement with an organization that does possess these skills. Before the data owner sends the dataset to the proxy sanitizer, the data owner may delete identifiers. Therefore, we need to consider a more flexible evaluation function. In other words, we need to create an index for the privacy risk of a document without using the name of the victim. We focus on the metadata as a way to handle this problem. Let w be the number of metadata that are also present in a searched document. Intuitively, the possibility that the searched document and the actual document indicate the same contents increases as w becomes large. We confirm the rela-

Table 3 Labeling result.

	Positive	Negative
True	650	646
False	10	6

© 2021 Information Processing Society of Japan

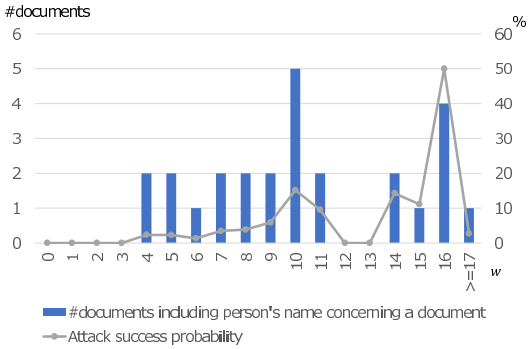


Fig. 3 Relationship between the number of w and attack success probability.

tionship between w and risk manually. In the experiment, we set all documents as d, which are the 3,000 SADs, and we applied *WebSearchAttack* to the data while focusing on the documents satisfying $w \geq 4$. The horizontal and vertical axes in **Fig. 3** represent w and `Attack` success probability, respectively. For example, we can see that the documents with $w \leq 3$ are not at risk of being linked to the name of the victim, 50% of the documents with $w = 16$ have the risk of being identified, and there are four documents in d. Worthy of attention here is that the attack success probability increases as w increases, and so the results support the hypothesis that there is a correlation between w and risk. In other words, when a document has many words that are the same as words included in another document, there is a high probability that the documents represent the same things. The results also show that the number of w is available for privacy risk evaluation. In actual operation, *WebSearchAttack* and *RiskEvaluation* are not independent and run alternately. *WebSearchAttack* attacks a document using $kw \subseteq KW_i$ and is the bottleneck. Thus, we can change the algorithm to improve the processing speed as follows; *RiskEvaluation* outputs w instead of 0 or 1 and when w exceeds a threshold, the algorithm moves on to *Reconstruct*, or aborts to attack and rejects to output the document immediately.

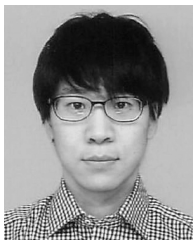
7. Conclusion

In this paper, we proposed a sanitization framework for documents and applied it to public accident reports from schools and in court documents. We first defined a generalized adversary model for re-identification and applied the model to a privacy issue on documents. Owing to this, we do not need to predefine what is sensitive information and it leads to a cost reduction compared with previous works. Furthermore, we do not need to calculate PMI, so that we can handle long articles such as court documents. We considered a web search engine is used for a simulation attack and implemented an attack algorithm based on the model and confirmed that attacks on actual sanitized documents occurred in our experiments. The results show that some documents remain at risk to be re-identified, even though they were sanitized by humans, and we thus proved that manual sanitization might not be sufficient for defeating attackers who use web search engines. Furthermore, we proposed a sanitizing algorithm against the attacks, and this algorithm sanitizes or generalizes only words that cause re-identification precisely. The situation assumed in our experiments is realistic, and even if documents do not have metadata, we can apply the framework by setting the words with a large volume of information as metadata.

Acknowledgments A part of this work was partly supported by JST CREST grant number JPMJCR1404, Japan.

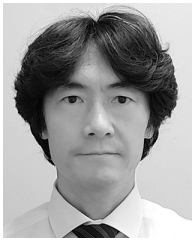
References

[1] Mimoto, T., Kiyomoto, S., Kitamura, K. and Miyaji, A.: A Practical Privacy-Preserving Algorithm for Document Data, *Proc. TrustCom 2020*, pp.1376–1383 (2020).
[2] Samarati, P. and Sweeney, L.: Generalizing data to provide anonymity when disclosing information, *Proc. PODS 1998* (1988).
[3] Dwork, C.: Differential Privacy, *Proc. ICALP 2006*, LNCS, Vol.4052, pp.1–12 (2006).
[4] Machanavajjhala, A., Gehrke, J. and Kifer, D.: *l*-Diversity: Privacy Beyond *k*-Anonymity, *Proc. ICDE '06*, pp.24–35 (2006).
[5] Machanavajjhala, A., Gehrke, J. and Kifer, D.: *t*-Closeness: Privacy Beyond *k*-Anonymity and *l*-Diversity, *Proc. ICDE '07*, pp.106–115 (2007).
[6] Sun, X., Wang, H., Li, J., Truta, T.M. and Li, P.: (p^+, α)-Sensitive *k*-anonymity: A new enhanced privacy protection model, *Proc. CIT '08*, pp.59–64 (2008).
[7] Zheng, Y.: Trajectory data mining: An overview, *ACM Trans. Intelligent Systems and Technology* (*TIST*), Vol.6, No.3 (2015).
[8] Mendes, R. and Vilela, J.P.: Privacy-preserving data mining: Methods, metrics, and applications, *IEEE Access*, Vol.5, pp.10562–10582 (2017).
[9] Sweeney, L.: Achieving *k*-anonymity privacy protection using generalization and suppression, *J. Uncertainty, Fuzziness, and Knowledge-Base Systems*, Vol.10, No.5, pp.571–588 (2002).
[10] LeFevre, K., DeWitt, D.J. and Ramakrishnan, R.: Incognito: Efficient Full-Domain *k*-anonymity, *Proc. SIGMOD 2005*, pp.49–60 (2005).
[11] LeFevre, K., DeWitt, D.J. and Ramakrishnan, R.: Mondrian Multidimensional K-Anonymity, *Proc. 22nd International Conference on Data Engineering* (*ICDE '06*), pp.25–35, IEEE (2006).
[12] Dwork, C.: Differential Privacy: A Survey of Results, *Proc. TAMC 2008*, LNCS, Vol.4978, pp.1–19 (2008).
[13] Sánchez, D. and Batet, M.: C-sanitized: A privacy model for document redaction and sanitization, *Journal of the Association for Information Science and Technology*, Vol.67, No.1, pp.148–163 (2016) (Wiley Online Library).
[14] Chakaravarthy, V.T., Gupta, H., Roy, P. and Mohania, M.K.: Efficient techniques for document sanitization, *Proc. 17th ACM Conference on Information and Knowledge Management*, pp.843–852 (2008).
[15] Bier, E., Chow, R., Golle, P., King, T.H. and Staddon, J.: The rules of redaction: Identify, protect, review (and repeat), *Journal of IEEE Security & Privacy*, Vol.7, No.6, pp.46–53, IEEE (2009).
[16] Anandan, B., Clifton, C., Jiang, W., Murugesan, M., Pastrana-Camacho, P. and Si, L.: t-Plausibility: Generalizing Words to Desensitize Text, *Journal of Trans. Data Privacy*, Vol.5, No.4, pp.505–534 (2012).
[17] Cumby, C. and Ghani, R.: A machine learning based system for semi-automatically redacting documents, *23rd IAAI Conference* (2011).
[18] Joulin, A., Grave, E., Bojanowski, P., Douze, M., Jégou, H. and Mikolov, T.: FastText.zip: Compressing text classification models, arXiv preprint arXiv:1612.03651 (2016).
[19] Church, K.W.: Word2Vec, *Journal of Natural Language Engineering*, Vol.23, No.1, pp.155–162 (2017).



Tomoaki Mimoto received his bachelors degree in engineering from Osaka university, Japan, in 2012, and received his master degree (Outstanding Performance Award) in information science from Japan Advanced Institute of Science and Technology in 2014. He joined KDDI in 2014, and was with the KDDI research, Inc. from 2015 to 2020. He is currently an research engineer in Advanced Telecommunication Research Institute International (ATR).

© 2021 Information Processing Society of Japan



Masayuki Hashimoto received his B.E., M.E. and Ph.D. degrees in communication engineering from Osaka University, Osaka, Japan, in 1995, 1997 and 2007, respectively. He received his M.B.A. from Graduate School of Management, GLOBIS University, Tokyo, Japan, in 2017. After he joined KDDI R&D Laboratories in 1997, he engaged research on digital image transmission, developed a mobile medical image transmission system and commercialized the system in medical fields in Japan. Since 2019, he has been a head of the Department of Advanced Security, Adaptive Communication Research Laboratories, Advanced Telecommunications Research Institute International. His research interests include data privacy, vulnerability-information extraction, group signature and supply chain security.

Shinsaku Kiyomoto received his B.E. in engineering sciences and his M.E. in Material Science from Tsukuba University, Japan, in 1998 and 2000, respectively. He joined KDD (now KDDI) and has been engaged in research on stream ciphers, cryptographic protocols, mobile security, and privacy protection. He is currently an executive director of KDDI Research, Inc. He was a visiting researcher of the Information Security Group, Royal Holloway University of London from 2008 to 2009. He received his doctorate in engineering from Kyushu University in 2006. He received the IEICE Young Engineer Award in 2004 and IEICE Distinguished Contributions Awards in 2011. He is a member of IEICE and JPS.

Koji Kitamura is a senior researcher of Artificial Intelligence Research Center, National Institute of Advanced Industrial Science and Technology. He received his Ph.D. degrees in mechanical engineering from Tokyo University of Science. His research topics are childhood injury prevention, elderly people injury prevention and supporting elderly care using IoT and AI technologies. He is a member of the executive board of NPO Safe Kids Japan and an expert member of the Consumer Safety Investigation Commission in Consumer Affairs Agency.

Atsuko Miyaji received her B.Sc., her M.Sc., and her Dr.Sci. degrees in mathematics from Osaka University, in 1988, 1990, and 1997 respectively. She is an IPSJ fellow. She joined Panasonic Co., Ltd. from 1990 to 1998 and engaged in research and development for secure communication. She was an associate professor at the Japan Advanced Institute of Science and Technology (JAIST) in 1998. She joined the computer science department of the University of California, Davis from 2002 to 2003. She has been a professor at Japan Advanced Institute of Science and Technology (JAIST) since 2007. She has been a professor at Graduate School of Engineering, Osaka University since 2015. Her research interests include the application of number theory into cryptography and information security. She received Young Paper Award of SCIS '93 in 1993, Notable Invention Award of the Science and Technology Agency in 1997, the IPSJ Sakai Special Researcher Award in 2002, the Standardization Contribution Award in 2003, the AWARD for the contribution to CULTURE of SECURITY in 2007, the Director-General of Industrial Science and Technology Policy and Environment Bureau Award in 2007, DoCoMo Mobile Science Awards in 2008, Advanced Data Mining and Applications (ADMA 2010) Best Paper Award, Prizes for Science and Technology, the Commendation for Science and Technology by the Minister of Education, Culture, Sports, Science and Technology, International Conference on Applications and Technologies in Information Security (ATIS 2016) Best Paper Award, the 16th IEEE Trustocm 2017 Best Paper Award, IEICE milestone certification in 2017, the 14th Asia Joint Conference on Information Security (AsiaJCIS 2019) Best Paper Award, Information Security Applications - 20th International Conference (WISA 2020) Best Paper Gold Award, and Distinguished Educational Practitioners Award in 2020. She is a member of the International Association for Cryptologic Research, the Institute of Electrical and Electronics Engineers, the Institute of Electronics, Information and Communication Engineers, the Information Processing Society of Japan, and the Mathematical Society of Japan.

© 2021 Information Processing Society of Japan