ARTICLE IN PRESS

Information Sciences xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Utility-preserving sanitization of semantically correlated terms in textual documents

David Sánchez *, Montserrat Batet, Alexandre Viejo

UNESCO Chair in Data Privacy, Department of Computer Science and Mathematics, Universitat Rovira i Virgili, Avda. Països Catalans, 26, 43007 Tarragona, Spain

ARTICLE INFO

Article history:
Received 8 July 2013
Received in revised form 27 January 2014
Accepted 20 March 2014
Available online xxxx

Keywords:
Data privacy
Document redaction
Document sanitization
Information theory

ABSTRACT

Traditionally, redaction has been the method chosen to mitigate the privacy issues related to the declassification of textual documents containing sensitive data. This process is based on removing sensitive words in the documents prior to their release and has the undesired side effect of severely reducing the utility of the content. Document sanitization is a recent alternative to redaction, which avoids utility issues by generalizing the sensitive terms instead of eliminating them. Some (semi-)automatic redaction/sanitization schemes can be found in the literature; however, they usually neglect the importance of semantic correlations between the terms of the document, even though these may disclose sanitized/redacted sensitive terms. To tackle this issue, this paper proposes a theoretical framework grounded in the Information Theory, which offers a general model capable of measuring the disclosure risk caused by semantically correlated terms, regardless of the fact that they are proposed for removal or generalization. The new method specifically focuses on generating sanitized documents that retain as much utility (i.e., semantics) as possible while fulfilling the privacy requirements. The implementation of the method has been evaluated in a practical setting, showing that the new approach improves the output's utility in comparison to the previous work, while retaining a similar level of accuracy.

1. Introduction

Legislations, economic pressures or the increasing trend toward outsourcing information to the Cloud have brought a risky scenario where thousands of documents containing potentially sensitive information related to individuals (e.g., identifiable data, personal information like diseases or economic status, etc.) or organizations (e.g., sale operations, commercial partners, etc.) are distributed and declassified daily.

Redaction is a well-known approach that tries to avoid (or at least mitigate) the privacy issues inherent to this scenario. This process is mainly based on blacking-out, obscuring or eliminating sensitive words in the documents prior to their release. Redaction schemes can be generally classified according to the level of supervision required by its users (i.e., manual, semi-supervised or fully-autonomous) and, also, according to the approach used to identify the sensitive elements of the text (e.g., use of lists of sensitive elements to be eliminated, trained classifiers, named entity recognition techniques, etc.).

* Corresponding author. Address: Departament d'Enginyeria Informàtica i Matemàtiques, Universitat Rovira i Virgili, Avda. Països Catalans, 26, 43007 Tarragona, Spain. Tel.: +34 977 559657; fax: +34 977 559710.

E-mail address: david.sanchez@urv.cat (D. Sánchez).

http://dx.doi.org/10.1016/j.ins.2014.03.103
0020-0255/© 2014 Elsevier Inc. All rights reserved.

The inherent problem in redaction methods is that they eliminate parts of the output document, thus reducing the *utility* of its content. In fact, in extreme cases, the redacted text can be no longer useful [8]. An additional problem of redacting documents is that the existence of obscured or blacked-out parts can raise the awareness of the document's sensitivity in front of possible attackers [3]. According to that, researchers have put their efforts into designing an alternative to redaction that preserves more utility while providing similar levels of privacy. These alternative methods, usually referred as *document sanitization*, are mainly based on *generalizing* the sensitive terms rather than directly eliminating them. The predominant advantages of sanitization over redaction is that the former pursuits to obtain a document that is less detailed than the original one but still provides enough utility, while no clues about the document's sensitivity are given.

Even though document sanitization addresses the utility reduction issues inherent to redaction, both suffer from a relevant problem that has received less attention from the scientific community: redaction/sanitization mechanisms generally evaluate the sensitivity of textual terms independently from each other. This situation is risky from the privacy point of view because the terms of any textual document are usually *semantically related* [2]. This fact may enable the re-identification of the redacted/sanitized elements from the presence of related terms left in clear forms. For example, a sanitization/redacting scheme that uses a list of diseases to detect sensitive elements in a document may identify the term *AIDS* as sensitive while other terms such as *blood transfusion* or *sexual transmission* may not be detected. These last two elements are apparently innocuous; however, by assuming that the adversary has a minimum knowledge of the domain [7], they can effectively re-identify *AIDS* by means of semantic inference [2] and hamper or even negate the whole redaction/sanitization process.

The prevention of the disclosure of sensitive information from the combination of, a priori, non-sensitive elements has been already addressed in the *Statistical Disclosure Control (SDC)* research field [10,15,16]. Nevertheless, the solutions which are proposed in that area deal with structured databases where the record attributes, whose combination of values may unequivocally identify an individual (i.e., quasi-identifiers), are defined beforehand. This strong requirement prevents SDC proposals from being applied to redact/sanitize unstructured textual documents in which *any* combination of terms may represent a disclosure risk depending on whether they are highly correlated or not.

In previous works [29,30], we have tackled the above problem by proposing an automatic redaction method that detects terms that are semantically correlated to sensitive ones, where the latter were identified with any redaction mechanism that analyses terms independently [1,6,27,28,37]. This method relies on the Information Theory to mathematically formulate the correlation between terms and to quantify the re-identification risk of a sensitive term caused by the presence of non-sanitized correlated terms. Since the method follows a pure *redaction* process, it assumes that both sensitive terms and those found to be semantically correlated with the former are *removed* prior publication. Given that the removal of sensitive data hampers the utility of the output, it is worth to mention that a redacting proposal such as [30], which is very likely to identify a large number of terms as sensitive (i.e., the sensitive elements and the semantically correlated ones), can incur in a high utility loss, which goes against the purpose of data publication.

To tackle this problem, in this paper, we extend the framework presented in [30] to enable an automatic and general-purpose utility-preserving sanitization of documents (regardless its domain of knowledge) that also considers semantically correlated terms. Our method acts as a complement to any redaction/sanitization method that detects sensitive terms independently. The main differences from the previous work are:

- It is able to quantify the disclosure risk of semantically correlated terms toward a sensitive term whether the latter is removed (redacted) or generalized (sanitized) prior publication.
- Terms that are found to cause a feasible disclosure of a sensitive one are generalized (rather than removed) coherently with the desired level of privacy.

To achieve the above goals, the present work offers the following new contributions:

- A general characterization from the perspective of the Information Theory of the disclosure risk caused by sensitive terms and their correlated terms, whenever they are removed [6,11,28,37,39] or generalized [1,14,27], prior publication.
- Exploitation of general-purpose knowledge and information sources to assist the disclosure risk assessment and the utility-preserving sanitization (regardless its domain of knowledge), which aims at retaining as much data semantics as possible while fulfilling the privacy requirements. As a result, and in comparison with approaches based on ad-hoc knowledge bases or trained classifiers [5,17,39], our method offers a domain independent solution that can be applied to textual documents regardless of their contents.
- A general, self-adaptive and automatic algorithm that enables the application of the theoretical framework in a practical scenario, regardless of the kind of sanitizer (manual/automatic, supervised/unsupervised, general/domain-specific, based on term removal or generalization) used to detect and hide sensitive terms.
- An evaluation of the improvements, in terms of utility and disclosure risk, brought by the present approach in comparison with previous works by using real and highly sensitive texts and a widely used sanitization mechanism based on *Named-Entity recognition* [13].

The rest of the paper is organized as follows. Section 2 reviews related works in document redaction/sanitization. Section 3 formalizes the theoretical framework that quantifies the disclosure risk of sensitive terms from a general perspective.