Efficient Techniques for Document Sanitization

V. Chakaravarthy
IBM India Research Lab,
New Delhi, India
vechakra@in.ibm.com

Himanshu Gupta
IBM India Research Lab,
New Delhi, India
higupta9@in.ibm.com

Prasan Roy
Aster Data Systems,
Redwood City, CA, USA
prasan.roy@in.ibm.com

Mukesh K. Mohania
IBM India Research Lab,
New Delhi, India
mkmukesh@in.ibm.com

ABSTRACT

Sanitization (syn. *redaction*) of a document involves removing sensitive information from the document, in order to reduce the document's classification level, possibly yielding an unclassified document. A document may need to be sanitized for a variety of reasons. Government departments usually need to declassify documents before making them public, for instance, in response to Freedom of Information requests. In hospitals, medical records are sanitized to remove sensitive patent information (patient identity information, diagnoses of deadly diseases, etc.). Document sanitization is also critical to companies who need to prevent malafide or inadvertent disclosure of proprietary information while sharing data with outsourced operations. In this paper, we propose ERASE (Efficient RedAction for Securing Entities), a system for performing document sanitization automatically.

1. INTRODUCTION

Sanitization (syn. *redaction*) of a document involves removing sensitive information from the document, in order to reduce the document's classification level, possibly yielding an unclassified document [8].

A document may need to be sanitized for a variety of reasons. Government departments usually need to declassify documents before making them public, for instance, in response to Freedom of Information requests. In hospitals, medical records are sanitized to remove sensitive patent information (patient identity information, diagnoses of deadly diseases, etc.). Document sanitization is also critical to companies who need to prevent malafide or inadvertent disclosure of proprietary information while sharing data with outsourced operations.

Traditionally, documents are sanitized manually by qualified reviewers. However, manual sanitization does not scale as the volume of data increases. The US Department of Energy's OpenNet initiative [7], for instance, needs to sanitize millions of documents each year. Given the amount of effort involved and limited supply of qualified reviewers, this is a tall order.

In this paper, we propose ERASE (Efficient RedAction for Securing Entities), a system for performing document sanitization automatically.

Contributions. Specific contributions in this paper are as follows.

- We present a principled approach to sanitization of unstructured text documents. While sanitization of structured relational databases has been addressed earlier [5, 3], we believe ERASE is the first work to provide a principled solution in an unstructured free-text domain.
- We device an algorithm that sanitizes a document while removing the minimum number of terms. We propose nontrivial pruning strategies that make the search practical.
- We present an alternative algorithm that achieves very reasonable performance even for large documents; the improvement in efficiency is achieved by relaxing the least-distortion property. We show that this alternative approach, in practice, is similar to the optimal algorithm in terms of the quality of the result.
- Our experimental study that shows that the proposed techniques are practical on realistic data.

2. OVERVIEW OF ERASE

ERASE models public knowledge as a database of entities (persons, products, diseases, etc.). Each entity in this database is associated with a set of terms related to the entity; this set is termed the *context* of the entity. For instance, the context of a person entity could include the firstname, the lastname, the day, month and year of birth, the street and city of residence, the employer's name, spouse's name, etc. This database could be structured or unstructured – wikipedia, a directory of employees and projects in an organization, a compendium of diseases, etc. all fit in the proposed notion of an entity database. The only requirement is that the database is able to provide the context of a given entity, and the set of entities containing a given term in their context. This context database need not be compiled manually – it can be a existing database, or can be extracted automatically using an information extraction system such as Snowball [1].

Some of the entities in the database are considered *protected*; these are the entities that need to be protected against identity disclosure. For instance, in a database of diseases, certain diseases (such as AIDS) can be marked as protected – we are interested in protecting the disclosure of these diseases, it does not matter if the any other disease (such as Influenza) is revealed. The set of protected entities is derived according to the access privileges of the adversary. The set of entities that need to be hidden from the adversary are declared protected.

ERASE assumes an adversary that knows nothing about an entity apart from what appears in the entity's context, and has bounded inference capabilities. Specifically, given a document, the adversary can match the terms present in the document with the terms present in the context of each of the protected entities. If the document contains a group of terms that appear together only in the context of a particular entity, then the adversary gets an indication that the entity is being mentioned in the given document. We term this a *disclosure*.

" Let's look at the immediate facts. You have a number of symptoms, namely **weight loss**, **insomnia**, **sweating**, **fatigue**, **digestive problems** *and* **headaches**. *These may or may not be related to* **sexually transmitted diseases**, *but you know you have been exposed to* **gonorrhea** *and you know you may have been exposed to* **hepatitis B** *and* **HIV**. *Your symptoms are significant and need full investigation in the near future. "*

Figure 1. Illustrative Example

ERASE attempts to prevent disclosure of protected entities by removing certain terms from the document – these terms need to be selected such that no protected entity can be inferred as being mentioned in the document by matching the remaining terms with the entity database.

A simplistic approach is to locate "give-away" phrases in the document and delete them all. To the best of our knowledge, most prior work on document sanitization has followed this approach, and has focused on developing more accurate ways of locating such phrases in the document's text [4, 2, 6]. We believe this is an overkill. For instance, in an intelligence report, removing all names, locations, etc. would probably leave the report with no useful content at all.

In contrast, ERASE makes an effort to sanitize a document while causing the least distortion to the contents of the document; indeed, this is considered one of the principal requirements of document sanitization [6]. Towards this goal, ERASE identifies the minimum number of terms in the document that need to be removed in order for the document to be sanitized.

3. ILLUSTRATIVE EXAMPLE

We illustrate our approach using an anecdotal real-life example. We created a database of 2645 diseases obtaining information from the website *wrongdiagnosis* (www.wrongdiagnosis.com). Each disease is an entity with the associated context consisting of symptoms, tests, treatments and risk factors. The website offers a classification of the diseases. We declared as protected entities the diseases under the following categories: sexual conditions, serious conditions (conditions related to heart, thyroid, kidney, liver, ovary), cancer conditions, and mental conditions. The number of protected entities were 550.

We created a document by taking a paragraph out of a communication between a doctor and a patient from another website (www.netdoctor.co.uk). The document is shown in Figure 1, where the relevant terms found in the entity contexts are shown in bold.

The document was sanitized using ERASE; the terms deleted as a result are shown underlined in Figure 1. Observe that the sanitization has differentiated between the generic symptoms and symptoms specific to the kind of diseases that appear in the protected entity set, and deleted the latter. Furthermore, though sweating is a common symptom associated with hundreds of diseases, the combination of sweating, weight loss and fatigue reveals a protected entity (in this case, HIV) and so, one of these terms must be deleted to sanitize the document; consequently, the system removed the term "sweating". We note that a different protected entity set might remove a different set of symptoms.

4. REFERENCES

[1] Eugene Agichtein, Luis Gravano, Jeff Pavel, Viktoriya Sokolova, and Aleksandr Voskoboynik. Snowball: A prototype system for extracting relations from large text collections. In SIGMOD, 2001.

[2] M.M. Douglass, G.D. Clifford, A. Reisner, W.J. Long, G.B. Moody, and R.G.Mark. De-identification algorithm for free-text nursing notes. In Computers in Cardiology, 2005.

[3] K. LeFevre, D.J. DeWitt, and R. Ramakrishnan. Incognito: Efficient full domain k-anonymity. In SIGMOD, 2005.

[4] Latanya Sweeney. Replacing personally-identifying information in medical records, the srub system. In Journal of the Americal Medical Informatics Association, 1996.

[5] Latanya Sweeney. K-anonymity: A model for protecting privacy. Intl Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10(5), 2002.

[6] Alex Tveit. Anonymization of general practitioner medical records. In HelsIT'04, Trondheim, Norway, 2004.

[7] U.S. Department of Energy. Department of energy researches use of advanced computing for document declassification. http://www.osti.gov/opennet.

[8] Wikipedia. Sanitization (classified information) — wikipedia, the free encyclopedia, 2006.