DETECTING TERM RELATIONSHIPS TO IMPROVE TEXTUAL DOCUMENT SANITIZATION

David Sánchez, Universitat Rovira i Virgili, UNESCO Chair in Data Privacy, Departament d'Enginyeria Informàtica i Matemàtiques. Av. Països Catalans 26, E-43007 Tarragona, Spain, david.sanchez@urv.cat

Montserrat Batet, Universitat Rovira i Virgili, UNESCO Chair in Data Privacy, Departament d'Enginyeria Informàtica i Matemàtiques. Av. Països Catalans 26, E-43007 Tarragona, Spain, montserrat.batet@urv.cat

Alexandre Viejo, Universitat Rovira i Virgili, UNESCO Chair in Data Privacy, Departament d'Enginyeria Informàtica i Matemàtiques. Av. Països Catalans 26, E-43007 Tarragona, Spain, alexandre.viejo@urv.cat

Abstract

Nowadays, the publication of textual documents provides critical benefits to scientific research and business scenarios where information analysis plays an essential role. Nevertheless, the possible existence of identifying or confidential data in this kind of documents motivates the use of measures to sanitize sensitive information before being published, while keeping the innocuous data unmodified. Several automatic sanitization mechanisms can be found in the literature; however, most of them evaluate the sensitivity of the textual terms considering them as independent variables. At the same time, some authors have shown that there are important information disclosure risks inherent to the existence of relationships between sanitized and non-sanitized terms. Therefore, neglecting term relationships in document sanitization represents a serious privacy threat. In this paper, we present a general-purpose method to automatically detect semantically related terms that may enable disclosure of sensitive data. The foundations of Information Theory and a corpus as large as the Web are used to assess the degree relationship between textual terms according to the amount of information they provide from each other. Preliminary evaluation results show that our proposal significantly improves the detection recall of current sanitization schemes, which reduces the disclosure risk.

Keywords: Privacy, Document sanitization, Term relationships, Information theory.

1 INTRODUCTION

In the last years, declassification of documents has become an essential tool to support information-intensive tasks such as scientific research, problem analysis or decision making. Medical or business environments are important scenarios where the publication of documents can make the difference. For example, declassifying medical records has significantly improved a wide range of procedures related to detecting, verifying and monitoring new diagnostic examinations and treatment methodologies (Jiang et al., 2009).

Nevertheless, the benefits of publishing textual documents are not provided without cost: on the one hand, many of these documents may contain confidential information about private entities; on the other hand, we live in a highly connected world where digital information can be easily copied and re-distributed. As a result of both convergent situations, potentially untrusted third parties might have access to large quantities of sensible data which would be used without the consent of their owners (Chen & Zhao, 2012; Mishra et al., 2011).

In order to avoid this dangerous situation, approaches designed to *sanitize* text documents before being declassified have been proposed in the past. Specifically, document sanitization techniques try to detect and remove or mask sensitive pieces of information from documents before being published, so that the risk of re-identification of individuals and/or of revealing confidential information is minimized.

First sanitization techniques were based on the manual application of certain rules or guidelines (Agency, 2005) detailing the correct procedures to ensure irreversible suppression or distortion of sensitive pieces of information. Nevertheless, manual mechanisms have been reported to be time-consuming (Dorr et al., 2006) and prone to disclosure risks (Cumby & Ghani, 2011). Moreover, they present serious scalability issues when the volume of data increases (Chakaravarthy et al., 2008). In order to overcome those problems, researchers have focused on designing automatic sanitization methods. Some schemes like (Chakaravarthy et al., 2008) use databases of entities (e.g., persons, products, diseases, etc) to detect sensitive elements. Other proposals like (Sweeney, 1996) base the sanitization on a set of patterns corresponding to sensitive information such as name, location, etc. More general mechanisms like (Abril et al., 2011) and (Cumby & Ghani, 2011) assume that named entities (i.e., proper nouns), due to their specificity, should be sanitized; to do that, automatic named entity recognition techniques are used.

However, in order to automatically sanitize documents from a general perspective, in which textual sources may cover a wide range of topics, refer to heterogeneous entities and for which no predefined structure may be assumed, most of the above approaches are not well-suited. More specifically, methods that are based on problem-specific and ad-hoc knowledge bases or trained classifiers are hardly generalizable. Methods based on the detection of named entities (NEs) offer a more general solution, however, as stated in (Sánchez et al., 2012), it usually happens that not all NEs reveal confidential information since they may refer to general entities (e.g. country names) and not all sensitive information is expressed with NEs (e.g. names of diseases, rare jobs, etc.). Moreover, NE recognition is usually limited to a set of categories (e.g. *persons*, *locations* or *organizations*). To tackle these problems, the authors in (Sánchez et al., 2012) propose a method based on Information Theory and the use of a large corpus, to assess the degree of sensitiveness of terms according to the amount of information that they provide. This approach relies on the fact that sensitive terms are those that, due to their specificity, provide more information than common terms. In this way, the textual terms that provide too much information (according to a predefined criterion) are detected and sanitized. To quantify the amount of information provided by each element, the proposed system uses the notion of Information Content (IC), computed from a corpus as large as the Web. A practical comparison against a state-of-the-art NE recognition package (i.e., Stanford Named Entity Recognizer (Finkel et al., 2005)), shown that the information theoretic approach significantly improved the detection recall while offer a more general and less constrained solution (Sánchez et al., 2012).

In any case, the above mechanisms share a fundamental limitation: they evaluate the sensitivity of textual terms considering them as independent variables. For example, a certain scheme that uses a database of diseases to detect sensitive elements in a document may identify the term *AIDS* as sensitive but other terms appearing within the same document such as *immunodeficiency*, *blood transfusion* and *sexual transmission* may not be detected. Nevertheless, those terms *are semantically related* and, hence, the latter may enable the re-identification of the former by means of semantic inference (Anandan & Clifton, 2011).

The prevention of the disclosure of sensitive information from the combination of, a priori, non-sensitive elements has been widely covered by the *Statistical Disclosure Control (SDC)* research field (Domingo-Ferrer, 2008; Martínez et al., 2012). However, the solutions which are proposed in that area deal with structured databases where *record attributes* whose combination of values may unequivocally identify an individual (named as quasi-identifiers) are defined a priori. Clearly, this requirement does not hold in a scenario with unstructured textual documents in which any combination of terms may potentially cause disclosure if a semantic relationship exists. Hence, new approaches that specifically focus on this problem should be provided to improve document sanitization methods.

1.1 Previous work on detecting term relationships when sanitizing unstructured documents

The existence of relationships between terms in unstructured documents has been barely addressed in the sanitization literature. Nevertheless, the relevance of this issue is stressed in (Anandan & Clifton, 2011). More specifically, the authors of that work show that sanitized terms could be re-identified given the presence of non-sanitized (or even generalized) ones in the same context. In that work, re-identification is enabled by a contingency table that quantifies the degree of correlation between each pair of textual terms and a taxonomy modelling term generalization. Unfortunately, the availability of such accurate contingency table and associated taxonomy in a general setting is quite unrealistic, hampering the applicability of the method that, on the other hand, provides a clear characterization of the disclosure risks inherent to presence of term relationships.

A more practical but hardly generalizable approach is presented in (Chakaravarthy et al., 2008). In that work, authors present a scheme that detects sensitive elements using a database of entities (e.g., persons, products, diseases, etc). Each entity in this database is associated to a certain context which contains a set of terms related to that entity (e.g., the context of a disease could include symptoms, treatments, risk factors, etc). Using this information, the proposed system detects terms to be sanitized in a straightforward manner by looking for sensitive entities and their contexts in the input database. Due to the cost of manually compiling such database covering a wide range of related terms, that proposal is mainly designed to work with domain-specific documents.

1.2 Contribution and plan of this paper

Due to the lack of practical and general-purpose schemes considering term relationships during the sanitization of documents, in this paper, we propose a method that automatically removes highly related terms in order to minimize disclosure risk of the sanitized output. The method is designed as an extension of the system proposed in (Sánchez et al., 2012) that, even though it provides more accurate and less constrained results than usual sanitization mechanisms found in the literature, it solely evaluates term sensitivity in an independent way.

Our contribution in the present paper consists on quantifying term relationships between terms detected as sensitive by (Sánchez et al., 2012) with those that were left in clear form. As a result, those terms for which a high enough relationship exists, according to a specific criterion that guides the sanitization process, are removed from the output. Our proposal also relies on Information Theory to automatically quantify the degree of term relationship and propose new terms to sanitize. Practical results reported in the evaluation section show that our proposal significantly reduces the disclosure risk of the outputs provided by (Sánchez et al., 2012).

The rest of this paper is organized as follows. Section 2 formalizes the sanitization scenario, describes the details of the individual sanitization proposed in (Sánchez et al., 2012) and how our proposal behaves as a complement to it. Section 3 evaluates the performance of the new scheme against baseline results using a set of real documents. The last section contains the conclusions and proposes some lines of future work.

2 DOCUMENT SANITIZATION AND TERM RELATIONSHIPS

The document sanitization scenario tackled in this paper is defined as follows:

- D. The original document to sanitize. It is viewed as an ordered sequence of textual terms (i.e., words or phrases).
- S_1. The sanitization method presented in (Sánchez et al., 2012), which detects sensitive terms independently.
- D'. The output of the sanitizer S_1 for an input document D, that is $D' = S_1(D)$. It is assumed that, in document D', sensitive terms detected by S_1 have been tagged but not removed yet.
- it_j. Each one of the sensitive terms contained in D', which have been identified by S_1.
- nt_j. Each one of the terms in D' which have not been identified as sensitive by S_1.
- S_2. This is our new proposal. It takes as input the sanitized document D' and identifies which non-sensitive terms, nt_j, may potentially re-identify any of the sensitive terms, it_j.
- rt_j. Each one of the, a priori, non-sensitive terms in D', which have been found to be highly related to a sanitized one by the proposed sanitization method S_2.
- D''. This is the output of the proposed sanitizer S_2, that is, $D'' = S_2(D')$, in which both sensitive terms detected by S_1, it_j, and highly related ones detected by S_2, rt_j, have been removed.

Basically, the new sanitization system S_2 takes as input a document D' that contains a set of terms identified as sensitive ($I = \{it_1, \dots, it_k\}$) and a remaining set of non-sensitive ones ($N = \{nt_1, \dots, nt_p\}$). Then, for each $it_j \in I$, S_2 evaluates its degree of relationship against any $nt_j \in N$. As a result, S_2 provides a subset $R \subseteq N$, where $R = \{rt_1, \dots, rt_q\}$, that contains all those highly related terms that might disclose any element of I. The sanitization output consists on a document D'' from which terms in I and also those in R have been removed.

The next two subsections describe both sanitizers: Section 2.1 briefly introduces the sanitization method S_1 which was presented in (Sánchez et al., 2012) whereas Section 2.2 details the new method S_2; finally, Section 2.3 presents an illustrative example showing how S_2 complements the sanitization offered by S_1 in order to minimize disclosure risk.

2.1 Sanitization method S_1: detecting sensitive terms

The method by (Sánchez et al., 2012) starts by extracting *noun phrases (NPs)* from the input document, since it is assumed that some of those may reveal sensitive data. To do so, S_1 relies on natural language processing tools such as OpenNLP[1], to perform the following steps:

- Sentence detection.
- Tokenization (i.e., word detection, including contraction separation).
- Part-of-speech tagging (POS) of individual tokens.
- Syntactic parsing of POS tagged tokens. As a result, NPs are extracted.
- Removal of *stop words* (e.g., determinants, prepositions or adverbs) from NPs.

After that, the degree of sensitiveness of each NP is quantified according the *amount of information* it provides. As stated in the introduction, sensitive terms are considered as those that provide *more*

[1] OpenNLP, Apache Software Foundation, http://opennlp.apache.org [last accessed: January 4th, 2013]

information than non-sensitive ones because the latter tend to be more general than the former. More specifically, S_1 relies on Information Theory and, in particular, on the notion of Information Content (IC) of textual terms to quantify the amount of information provided by a NP.

The IC of a term t is computed as the inverse of its probability of appearance in a corpus ($p(t)$) (see Equation 1). In this manner, infrequent terms (e.g. pancreatic cancer) are considered more informative than common ones (e.g. disease).

$$IC(t) = -log_2\, p(t) \quad (1)$$

In order to unambiguously compute term frequencies, classic proposals like (Resnik, 1995) use tagged textual data as corpora. Nevertheless, this solution has two main drawbacks: (i) such a corpus must be compiled and tagged manually, which represents an important cost; and (ii) it is unlikely to have a tagged corpora large and representative enough to cover a wide range of general and specific domains. This last situation may lead to data sparseness problems (i.e., the fact that not enough data is available to extract reliable conclusions from their analysis) when computing the IC of concrete terms (e.g., rare diseases), names or recent terms (e.g., netbook, tablet) (Sánchez, Batet, Valls, et al., 2010). Due to the fact that document sanitization focuses precisely on concrete (i.e., highly informative) terms, a wider corpus covering all those cases is required in order to obtain valid IC values.

In contrast to ad-hoc corpora, the Web can be considered a general-purpose corpus that covers almost any possible up-to-date term. According to (Cilibrasi & Vitányi, 2006), the Web is so large and heterogeneous that it tends to represent the true current distribution of terms at a social scale. As a result, the Web can be used to compute realistic IC values (Sánchez, Batet, Valls, et al., 2010). Moreover, Web Search Engines (WSEs), such as Google or Bing, can be directly queried to obtain term occurrences (i.e., page counts) at a web-scale (Sánchez, Batet, Valls, et al., 2010; Turney, 2001) in order to compute term probabilities.

Hence, using the Web as corpora, S_1 quantifies the IC of a potentially sensitive term t as follows:

$$IC_{web}(t) = -log_2\, p_{web}(t) = -log_2\, \frac{PageCounts(t)}{TotalWebs}, \quad (2)$$

where $PageCounts(t)$ stands for the number of results provided by a WSE when querying t; and $TotalWebs$ represents the total amount of web sites indexed by the WSE (e.g., around 11 billion in Bing[2]).

The final step of the sanitization method consists on assessing which NPs provide too much information according to their IC and, hence, should be sanitized. To achieve that, a *detection threshold* β stating the sanitization criterion is fixed by the user. This value, which is also expressed in terms of IC, represents the amount of information provided by the most general term that the user wants to hide in the sanitization output. For example, if the objective is to hide that a certain document (e.g. an electronic health care record) refers to *acquired immunodeficiency syndrome*, the detection threshold should be fixed as $\beta = IC_{web}("acquired\ immunodeficiency\ syndrome")$. Hence, S_1 marks as sensitive (in D') all NPs that are equally or more informative than β (i.e., $\forall\, it_j, IC_{web}(it_j) \geq \beta$). These NPs should be sanitized (i.e. removed) from the sanitized output.

2.2 Sanitization method S_2: detecting risky term relationships

The input of S_2 is the document D', which has been already sanitized by S_1. This document states the set of NPs detected as sensitive it_j by S_1, and maintains the rest of, a *priori*, non-sensitive ones, nt_j. Then, the purpose of S_2 is to identify which nt_j may disclose any of the already detected sensitive it_j.

[2] http://www.worldwidewebsize.com/ [last accessed: January, 2013]

In order to achieve that, we hypothesize that the *disclosure risk* derived from the presence of a certain term nt_j with regard to an already sanitized term it_j when both appear in the same document, can be measured according to the amount of information that nt_j, which would appear in clear form in the sanitized output, reveals about it_j. In terms of Information Theory, the amount of *Mutual Information* (MI) between terms, which measures the correlation between variables, can be used to extract such information. In particular, the instantiation of MI for two specific observations results in the well-known *Point-wise Mutual Information (PMI)*, which quantifies the difference between the probability of their co-occurrence given their join distribution and their marginal distributions (Church & Hanks, 1990). Applying PMI to measure the mutual information between a sanitized and a non-sanitized term we obtain the following expression:

$$PMI(it_j, nt_j) = log_2 \frac{p(it_j, nt_j)}{p(it_j) \times p(nt_j)} \quad (3)$$

PMI has been successfully applied in the past to evaluate different types of semantic correlations such as word collocation (Bouma, 2009), synonymy (Turney, 2001), taxonomic subsumption and similarity (Sánchez, 2010; Sánchez, Batet, & Valls, 2010; Vicient et al., 2013) and a variety of non-taxonomic relationships (Sánchez, 2010; Sánchez & Moreno, 2008). From these works, one can conclude that *PMI* is a suitable measure in the context of document sanitization, in which a high semantic correlation between term pairs may enable disclosure regardless the type of correlation (i.e., taxonomic or non-taxonomic).

Notice also that, in terms of Information Content, *PMI* fulfils that $PMI(it_j, nt_j) = IC(it_j) - IC(it_j|nt_j)$, where the *conditional information content* of it_j given the presence of nt_j is computed as $IC(it_j|nt_j) = -\log_2 p(it_j, nt_j)/p(nt_j)$. From the former expression, it turns out that $PMI(it_j, nt_j)$ can be used to measure how much information the presence of nt_j discloses about it_j (which is assumed to be removed in the sanitization process).

Numerically, if it_j and nt_j are completely independent (i.e., they co-occur in a textual context by chance), the result for Equation 3 is $PMI(it_j, nt_j) = 0$. This means that the presence of nt_j does not provide any particular evidence of it_j and, hence, there is no disclosure risk. On the contrary, if whenever nt_j occurs, it_j also occurs (i.e., $p(it_j, nt_j) = p(nt_j)$), the result of Equation 3 is $PMI(it_j, nt_j) = -\log_2 p(it_j) = IC(it_j)$. This means that, when $PMI(it_j, nt_j) = IC(it_j)$, it can be concluded that the presence of nt_j in a document completely discloses it_j and, hence, there is maximum disclosure risk.

Hence, negative or zero values for *PMI* will result in no disclosure risk, whereas positive values will indicate an increasing risk.

Considering the above theoretical background, S_2 quantifies the disclosure risk caused by the potential semantic relationship between any sanitized and non-sanitized term in D'. Like in S_1, to provide a general-purpose solution, we query WSEs to compute web-scale term probabilities (see section 2.1), obtaining the following expression:

$$PMI_{web}(it_j, nt_j) = IC_{web}(it_j) - IC_{web}(it_j|nt_j) =$$

$$= -log_2 \frac{PageCounts("it_j")}{TotalWebs} + log_2 \frac{PageCounts("it_j" \; AND \; "nt_j")/TotalWebs}{PageCounts("nt_j")/TotalWebs} \quad (4)$$

Notice the use of double quotes (“”) in web queries to force the search for the exact term and the use of the AND operator to force the co-occurrence of both terms within the same document.

After that, S_2 should assess which term relationships may incur in disclosure risk according to their *PMI* values. Whenever this happens, the preliminary non-sensitive term nt_j is proposed for sanitization (i.e., $nt_j \rightarrow rt_j$).

To do so, S_2 considers as risky those relationships whose *PMI* (computed using Equation 4) is higher or equal than a threshold τ. To define τ in a way that is coherent with the criterion implemented by S_1, we compute τ as a function of S_1's sanitization threshold β. Recall that β corresponded to the IC of the most general term considered as sensitive by S_1. We weight this value by a user-defined parameter w that represents the relative amount (in parts per unit) of information that the non-sanitized term $nt_j$is allowed to reveal about the sensitive term it_j, as follows:

$$\tau = w \cdot \beta \quad , \tag{5}$$

where w is defined in the interval [0..1].

Since w weights β, and the latter measures IC, the resulting τ is also expressed in terms of IC and reflects the maximum amount of information that a non-sanitized term may disclose about a sensitive one. In this way, for any pair of terms that obtains a *PMI* value equal or above τ, we consider that the non-sanitized term reveals too much information of the sensitive one and, hence, it must be sanitized (i.e., $nt_j \rightarrow rt_j$). Values of w close to 1, state that we allow revealing almost all the information provided by a sensitive term. Inversely, values of w close to 0, state that we allow revealing a very low percentage of sensitive information.

Formally, the set $R = \{rt_1, \ldots, rt_q\}$ of too highly related terms that may enable disclosure of elements in I from a document $D'$is obtained as follows:

$$R = \{nt_i \in N \mid \exists it_j \in I \; with \; PMI_{web}(it_j, nt_j) \geq \tau\} \tag{6}$$

As a result of this process, the final output of the system will be a document D''in which all terms in I(detected by S_1) and in R(detected by S_2) will be removed.

It is worth to mention that a sanitized document generally losses part of its usability as a result of term removal. Hence, the more strict the sanitization is (i.e., w value close to 0), the more the eliminated terms are and, hence, the more the document usability decreases. Nevertheless, eliminating terms also reduces disclosure risk. According to that, there is a clear trade-off between the level of privacy and the level of usability achieved by the sanitized document. Therefore, by setting w and, hence, τ, users can configure their desired level of trade-off.

2.3 Example

Consider that the text shown in Figure 1 is the document D that will be sanitized by the application of $S_1 + S_2$.

> The patient suffers from acquired immunodeficiency syndrome because of blood transfusion. He was diagnosed when his immune system responded poorly influenza.

Figure 1. Sample document to be sanitized.

Let us assume that the threshold for the first sanitization mechanism, S_1, was fixed to $\beta = IC_{web}(\text{"}acquired\ immunodeficiency\ syndrome\text{"})$. This means that the goal is to get a sanitized output that does not disclose this disease. First, S_1 detects the NPs of the input text. Later, their *IC* is computed and all those that provide an amount of information equal or above β are marked as sensitive. *IC* values of each NP are presented in Table 1.

Term	*IC*
patient	7.25
acquired immunodeficiency syndrome	14.33
blood transfusion	12.70
He	2.87
immune system	10.54
influenza	9.52

Table 1. IC values computed from Bing for each NP (once stop words are removed).

Note that, in this case, $\beta = 14.33$ and, hence, only the NP "acquired immunodeficiency syndrome" has been sanitized. The output of this process D', is shown in Figure 2.

> [*NP* The patient] suffers from [*NP* ~~acquired immunodeficiency syndrome~~] because of [*NP* blood transfusion]. [*NP* He] was diagnosed when [*NP* his immune system] responded poorly to [*NP* influenza].

Figure 2. Document D' outputted by S_1.

Next, the proposed method S_2 gets D' as input and evaluates all the possible relationships between the sensitive term "acquired immunodeficiency syndrome" and the rest of the NPs of the document. All relationships with PMIs above the threshold τ will be proposed for further sanitization. Let us assume that the parameter w is set to $w = 0.6$, which states that we allow revealing up to a 60% of sensitive information. Considering that $\beta = 14.33$, we obtain a threshold $\tau = 8.59$. According to this value, S_2 sanitizes all non-sensitive terms nt_j whose PMI is above τ. Table 2 depicts the different *PMI* values. As a result, the terms "blood transfusion" and "Immune system", which are closely related to "acquired immunodeficiency syndrome", are proposed for sanitization. Figure 3 shows the final sanitized output D''.

Sensitive term	*Non-sensitive terms*	*PMI*
acquired immunodeficiency syndrome	patient	6.60
	blood transfusion	9.19
	He	2.22
	Immune system	8.89
	influenza	7.43

Table 2. PMI values for each relationship between the sensitive term "acquired immunodeficiency syndrome" and other NPs in D'.

> The [*NP* patient] suffers from [*NP* ~~acquired immunodeficiency syndrome~~] because of a [*NP* ~~blood transfusion~~]. [*NP* He] was diagnosed when his [*NP* ~~immune system~~] responded poorly to [*NP* influenza].

Figure 3. Final sanitized document D''outputted by S_2.

3 EVALUATION

This section presents some results that show the degree of accuracy achieved by the proposed mechanism in detecting sensitive elements of raw textual documents. Since our method works as a second sanitization step for the method presented in (Sánchez et al., 2012), we evaluate the level of improvement provided by the new proposal by comparing the results obtained by the combination of $S_1 + S_2$ with the results provided by S_1 alone.

In order to test our method in a realistic setting and to enable a fair comparison with the results provided by S_1, we employ the same set of real texts used in (Sánchez et al., 2012). In particular, they

correspond to six biographical sketches about actors/actresses taken from English Wikipedia articles. Wikipedia descriptions of concrete entities usually contain a high amount of identifiable and correlated information, which makes the detection of sensitive data a challenging task.

To evaluate the results obtained by both methods, we requested two human experts to select and agree on which terms or term combinations (i.e. relationships) may enable disclosure of the described entity, considering that it is desired to hide the fact that each evaluated text refers to a *certain* actor/actress. In other words, under this sanitization requirement, an external observer who reads the *sanitized version* of the sketch linked to, for example, *Sylvester Stallone* would be able to learn that this document refers to a person or to an actor, but she would not be able to learn that it refers to *Sylvester Stallone* (the term "actor" is more general/provides less information than the term "Sylvester Stallone"). According to that, the threshold β used in S_1 to sanitize the text that refers to a particular actor is set to the actor's name, this is $\beta = IC_{web}("Actor\ name")$.

To compute term and relationship sensitiveness, the *Bing Web Search Engine*, which indexes approximately 11 billion of web sites, has been used. The detection performance is quantified by means of *precision*, *recall* and *F-measure* (Manning et al., 2008).

Precision is calculated as the percentage between the number of terms detected by the automatic sanitization method (let us name this set $AutomaticSensitiveTerms$) that have also been selected by the human experts (named $HumanSensitiveTerms$), and the total number of terms that have been detected by the automatic sanitization method. This is represented in Equation 7. Note that the higher the precision, the more accurate the sanitization output will be.

$$\text{Precision} = \frac{|AutomaticSensitiveTerms \cap HumanSensitiveTerms|}{|AutomaticSensitiveTerms|} * 100 \quad (7)$$

Recall is calculated as the percentage between the number of sensitive terms that are both in $AutomaticSensitiveTerms$ and $HumanSensitiveTerms$ sets and the total number of terms that have been detected by the human experts. This is represented in Equation 8. The higher the recall, the lower the disclosure risk because a lower number of non-detected sensitive terms (S_1) or a lower number of terms that may enable disclosure of a sensitive one (S_2) would remain in the sanitized text.

$$\text{Recall} = \frac{|AutomaticSensitiveTerms \cap HumanSensitiveTerms|}{|HumanSensitiveTerms|} * 100 \quad (8)$$

Finally, *F-measure* quantifies the harmonic mean of recall and precision, summarizing the detection accuracy of the automatic sanitization method. This is represented in Equation 9.

$$\text{F-measure} = \frac{2 * Recall * Precision}{Recall + Precision} \quad (9)$$

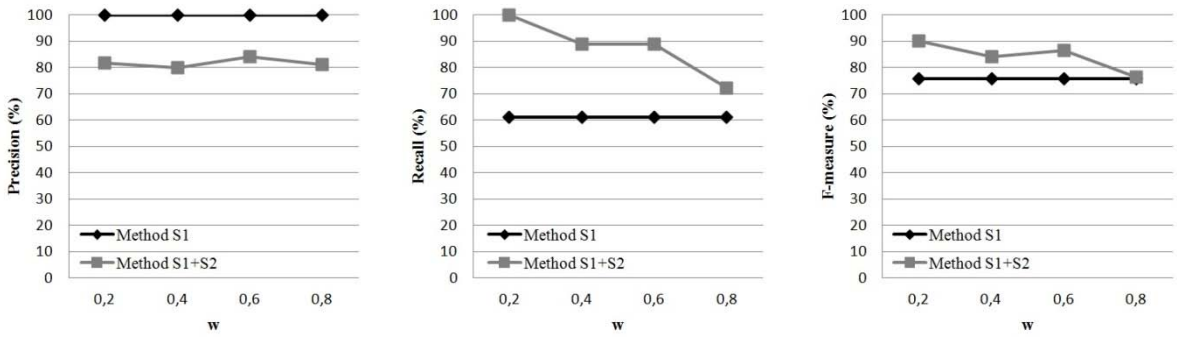


Figure 4. Precision, Recall and F-measure obtained by Method S_1 alone and for the combination of $S_1 + S_2$ for the entity "Antonio Banderas" with $\beta = IC_{web}("Antonio\ Banderas")$ and different values of w.

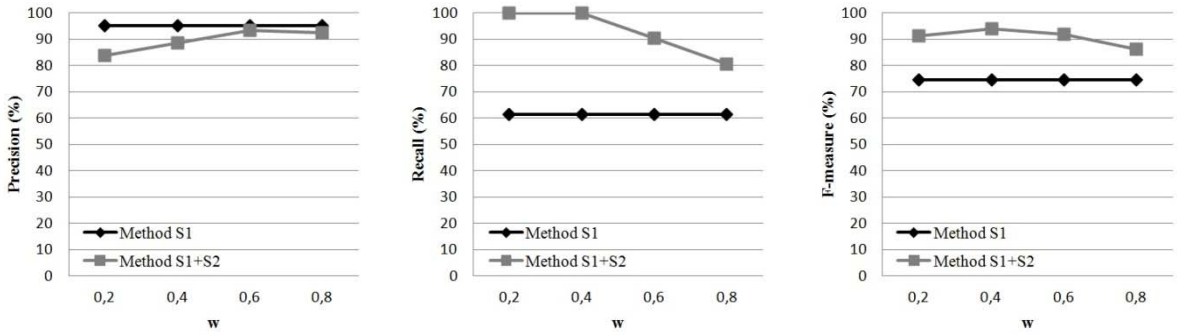


Figure 5. *Precision, Recall and F-measure obtained by Method S_1 alone and for the combination of $S_1 + S_2$ for the entity "Javier Bardem" with $\beta = IC_{web}("Javier\ Bardem")$ and different values of w.*

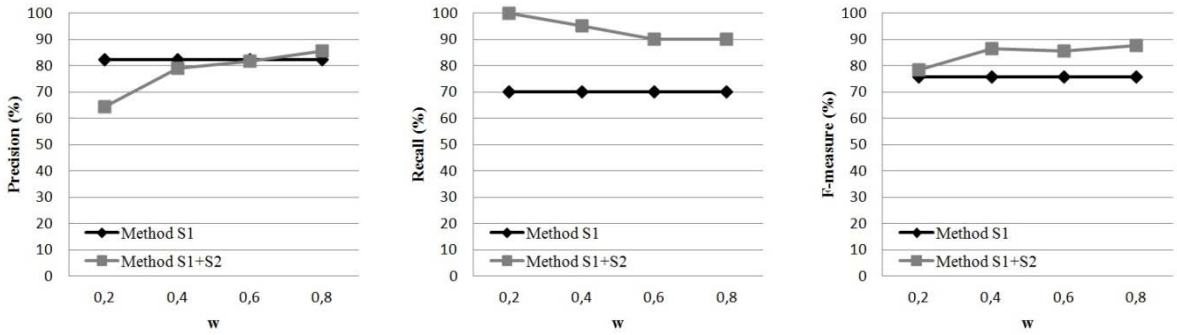


Figure 6. *Precision, Recall and F-measure obtained by Method S_1 alone and for the combination of $S_1 + S_2$ for the entity "Arnold Schwarzenegger with $\beta = IC_{web}("Arnold\ Schwarzenegger")$ and different values of w.*

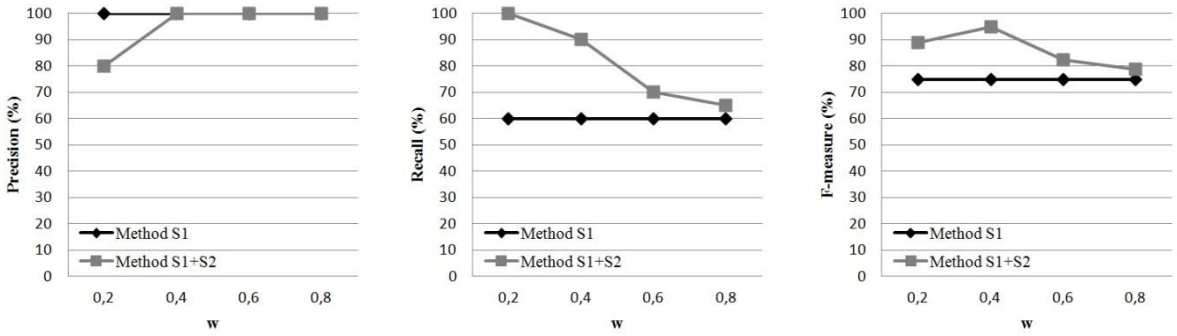


Figure 7. *Precision, Recall and F-measure obtained by Method S_1 alone and for the combination of $S_1 + S_2$ for the entity "Jordi Mollà" with $\beta = IC_{web}("Jordi\ Mollà")$ and different values of w.*

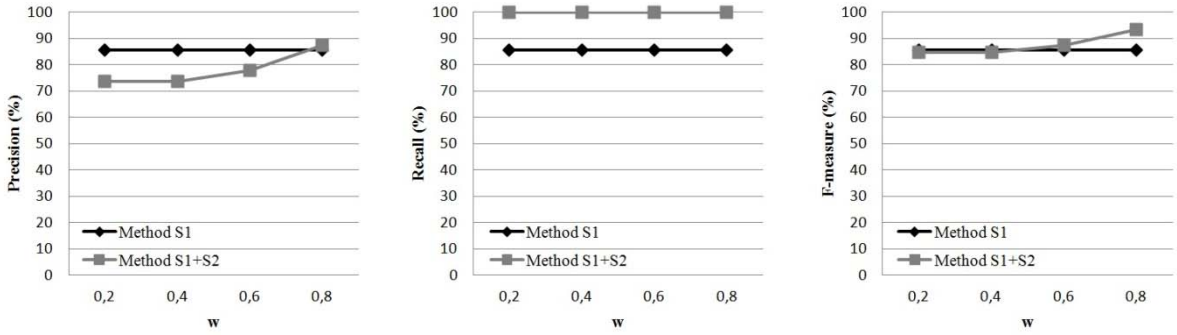


Figure 8. *Precision, Recall and F-measure obtained by Method S_1 alone and for the combination of $S_1 + S_2$ for the entity "Sylvester Stallone" with $\beta = IC_{web}("Sylvester\ Stallone")$ and different values of w.*

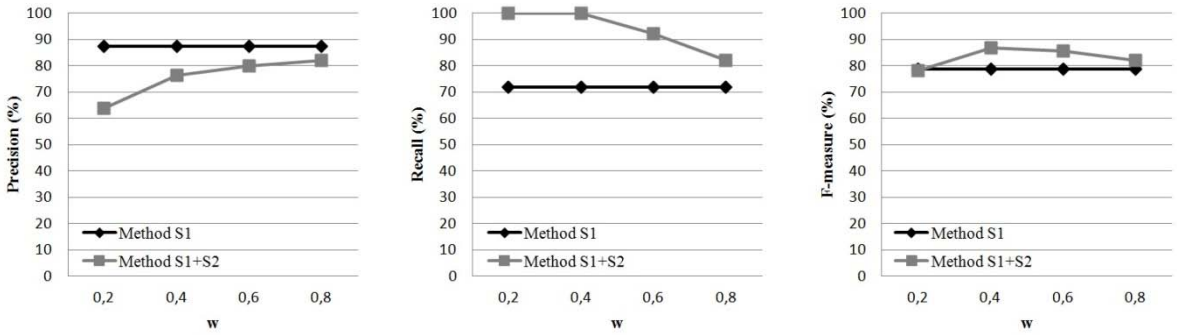


Figure 9. *Precision, Recall and F-measure obtained by Method S_1 alone and for the combination of $S_1 + S_2$ for the entity "Audrey Hepburn" with $\beta = IC_{web}("Audrey\ Hepburn")$ and different values of w.*

Results related to *precision* (the first graph in Figures 4-9) show that, generally, the method S_1 working alone achieves a better accuracy than $S_1 + S_2$. This was expected because precision mainly depends on the number of false positives (i.e., those terms considered as sensitive that were not according to the human criterion) and, hence, since the combination of $S_1 + S_2$ tends to sanitize *more* terms than S_1, the former will more likely achieve a lower precision. The level of precision obtained by $S_1 + S_2$ depends directly on the fixed parameter w. For values of w close to 1, the threshold τ is larger and more similar to the β criterion used by S_1. Hence, the number of detected terms by S_2 with respect to those detected by S_1 is smaller, resulting in better precision. On the other hand, for values of w close to 0, the threshold τ is smaller and the sanitization implemented by S_2 becomes more exhaustive since most of terms co-occurring in a document are related in some degree; thus the number of false positives (which may hamper document's utility) tends to increase.

Recall, on the other hand, represents a more important dimension than *precision* in the context of document sanitization, due to the fact that a method that achieves a low recall is more likely to generate sanitized documents that contain sensitive terms or that enable disclosure of sanitized terms. It is worth to mention that a single remaining sensitive term in a sanitized text might be able to fully disclose the information that the sanitization method tries to protect. The results related to *Recall* (the second graph in Figures 4-9) show that the use of the new proposal always achieves a better recall than method S_1 working alone. Moreover, recall results for $S_1 + S_2$ are better when the value w is closer to 0 (that is, a very low amount of information regarding the sensitive term is disclosed). This is coherent because in those cases the sanitization mechanism becomes more restrictive and more terms are detected as sensitive (*Recall* improves while *Precision* decays). From the point of view of the

information disclosure risk, these results show that the new step S_2, is a necessary addition to the sanitization process of a document.

Finally, the *F-measure* reflects the global accuracy of the tested methods. The results related to this performance aspect (the third graph in Figures 4-9) show that, in almost all the cases, the new proposal outperforms method S_1 working alone, showing a better trade-off between *precision* (which influences document's utility) and *recall* (which directly affects disclosure risk). Only in the texts related to *Sylvester Stallone* (Figure 8) and *Audrey Hepburn* (Figure 9), and for values of the *w* parameter closer to 0, the method S_1 obtains a better f-measure. In average, we observe that a *w* around 0.4 or 0.6 provides the best trade-off between precision and recall, even though this can be configured by the user to fit her sanitization needs (i.e. more utility or reduced disclosure risk).

4 CONCLUSIONS AND FUTURE WORK

Several sanitization mechanisms can be found in the literature; however, almost all of them evaluate the sensitivity of the textual terms considering them as independent variables. This situation is not ideal because works like (Anandan & Clifton, 2011) have shown that disclosure risk is inherently dependant on the existence of semantic relationships between terms. More specifically, it has been proved that sanitized terms could be re-identified given the presence of non-sanitized ones in the same context.

In this paper, an automatic text sanitization method focusing on discovering and quantifying term relationships has been proposed. This new scheme works as an extension to the system presented in (Sánchez et al., 2012). It gets as input a set of terms that have been detected as sensitive and verifies if they are semantically correlated with any other term of the document which has been left, a priori, in clear form. Our proposal relies on the foundations of the information theory and a corpus as global as the Web to offer a general-purpose solution that can be automatically applied to heterogeneous textual documents.

Evaluation results have shown that the proposed method is able to significantly increase the detection recall of the system presented in (Sánchez et al., 2012), which, in turn, was better than usual sanitization methods based on NE recognition (Abril et al., 2011). As a result, the disclosure risk of the sanitized output was minimized, while providing reasonable levels of accuracy (i.e. utility).

Regarding the future work, we plan to apply and evaluate the performance of the proposed method to other types of sanitizers (e.g. based on NE-detection (Abril et al., 2011; Cumby & Ghani, 2011), databases (Chakaravarthy et al., 2008), etc.) and other kind of general and domain-dependant documents. Moreover, since our approach is only meant to discover binary relationships, but disclosure may appear with combinations of *several* non-sanitized terms, we also plan to extend our method to support relationships of larger cardinalities.

Disclaimer and Acknowledgements

Authors are solely responsible for the views expressed in this paper, which do not necessarily reflect the position of UNESCO nor commit that organization. This work was partly supported by the European Commission under FP7 project Inter-Trust, by the Spanish Ministry of Science and Innovation (through projects eAEGIS TSI2007-65406-C03-01, CO-PRIVACY TIN2011-27076-C03-01, ARES-CONSOLIDER INGENIO 2010 CSD2007-00004, Audit Transparency Voting Process IPT-430000-2010-31, BallotNext IPT-2012-0603-430000 and ICWT TIN2012-32757) and by the Government of Catalonia (under grant 2009 SGR 1135).

References

Abril, D., Navarro-Arribas, G., and Torra, V. (2011). On the declassification of confidential documents *Modeling decision for artificial intelligence. 8th international conference, mdai 2011* (Vol. 6820, pp. 235–246): Springer.

Agency, N. S. (2005). Redacting with confidence: How to safely publish sanitized reports converted from word to pdf (Vol. Technical Report I333-015R-2005).

Anandan, B., and Clifton, C. (2011). Significance of term relationships on anonymization. Paper presented at the *IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology - Workshops*, Lyon, France.

Bouma, G. (2009). Normalized (pointwise) mutual information in collocation extraction. Paper presented at the *Biennial GSCL Conference 2009*, Tübingen, Germany.

Cilibrasi, R. L., and Vitányi, P. M. B. (2006). The google similarity distance. *IEEE Transactions on Knowledge and Data Engineering*, 19 (3), 370-383.

Cumby, C., and Ghani, R. (2011). A machine learning based system for semiautomatically redacting documents. Paper presented at the *Twenty-Third Conference on Innovative Applications of Artificial Intelligence*, San Francisco, California, USA.

Chakaravarthy, V. T., Gupta, H., Roy, P., and Mohania, M. K. (2008). Efficient techniques for document sanitization. Paper presented at the *17th ACM Conference on Information and Knowledge Management (CIKM'08)*, Napa Valley, California, USA.

Chen, D., and Zhao, H. (2012). Data security and privacy protection issues in cloud computing. Paper presented at the *2012 International Conference on Computer Science and Electronics Engineering (ICCSEE'12)*, Hangzhou, China.

Church, K. W., and Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16 (1), 22-29.

Domingo-Ferrer, J. (2008). A survey of inference control methods for privacy-preserving data mining. In C. C. Aggarwal & P. S. Yu (Eds.), *Privacy-preserving data mining* (pp. 53-80): Springer.

Dorr, D., Phillips, W., Phansalkar, S., Sims, S., and Hurdle, J. (2006). Assessing the difficulty and time cost of de-identification in clinical narratives. *Methods of Information in Medicine*, 45 (3), 246–252.

Finkel, J., Grenager, T., and Manning, C. (2005). Incorporating non-local information into information extraction systems by gibbs sampling. Paper presented at the *43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, Ann Arbor, USA.

Jiang, W., Murugesan, M., Clifton, C., and Si, L. (2009). T-plausibility: Semantic preserving text sanitization. Paper presented at the *International Conference on Computational Science and Engineering (CSE'09)*, Vancouver, Canada.

Manning, C. D., Raghavan, P., and Schtze, H. (2008). *Introduction to information retrieval*. Cambridge Cambridge University Press.

Martínez, S., Sánchez, D., and Valls, A. (2012). Semantic adaptive microaggregation of categorical microdata. *Computers & Security*, 31 (5), 653-672.

Mishra, R., Dash, S., Mishra, D., and Tripathy, A. (2011). A privacy preserving repository for securing data across the cloud. Paper presented at the *3rd International Conference on Electronics Computer Technology (ICECT'11)*.

Resnik, P. (1995, August 20 - 25). Using information content to evalutate semantic similarity in a taxonomy. Paper presented at the *14th International Joint Conference on Artificial Intelligence, IJCAI 1995*, Montreal, Quebec, Canada.

Sánchez, D. (2010). A methodology to learn ontological attributes from the web. *Data & Knowledge Engineering* 69 (6), 573-597.

Sánchez, D., Batet, M., and Valls, A. (2010). Web-based semantic similarity: An evaluation in the biomedical domain. *International Journal of Software and Informatics*, 4 (1), 39-52.

Sánchez, D., Batet, M., Valls, A., and Gibert, K. (2010). Ontology-driven web-based semantic similarity. *Journal of Intelligent Information Systems*, 35 (3), 383-413.

Sánchez, D., Batet, M., and Viejo, A. (2012). Detecting sensitive information from textual documents: An information-theoretic approach *Modeling decisions for artificial intelligence. 9th international conference, mdai 2012* (Vol. 7647, pp. 173-184): Springer.

Sánchez, D., and Moreno, A. (2008). Learning non-taxonomic relationships from web documents for domain ontology construction. *Data & Knowledge Engineering*, 63 (3), 600-623.

Sweeney, L. (1996). Replacing personally-identifying information in medical records, the scrub system. Paper presented at the *1996 American Medical Informatics Association Annual Fall Symposium*, Washington, DC, USA.

Turney, P. D. (2001). Mining the web for synonyms: Pmi-ir versus lsa on toefl. Paper presented at the *12th European Conference on Machine Learning, ECML 2001*, Freiburg, Germany.

Vicient, C., Sánchez, D., and Moreno, A. (2013). An automatic approach for ontology-based feature extraction from heterogeneous textual resources. *Engineering Applications of Artificial Intelligence*, 26 (3), 1092-1106