Utility-preserving sanitization of semantically correlated terms in textual documents

David Sánchez [1], 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)

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.

Keywords: data privacy, document redaction, document sanitization, information theory.

[1] Corresponding author. Address: Departament d'Enginyeria Informàtica i Matemàtiques. Universitat Rovira i Virgili. Avda. Països Catalans, 26. 43007. Tarragona. Spain
Tel.: +034 977 559657; Fax: +034 977 559710;
E-mail: david.sanchez@urv.cat.

1. Introduction

Legislations, economic pressures or the increasing trend towards 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.).

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 towards 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. Section 4 details the implementation of the framework and proposes a practical algorithm. Section 5 details the evaluation metrics and compares the results obtained for a collection of real documents against the previous work. The last section depicts the conclusions and proposes some lines of future work.

2. Related work

Document sanitization consists of two tasks: (i) detecting the sensitive terms; and (ii) generalizing the sensitive terms to reduce the information disclosure and fulfill a privacy criterion, but still keeping a certain utility level. In the literature, the first task has usually

received more attention than the latter [14]. For this reason, there are several redaction mechanisms (i.e., they focus on the detection step and directly remove the identified elements) but only a few purely sanitization ones (i.e., they address both steps). Then, if we focus on the sanitization schemes that consider the existence of semantically correlated terms in the document to be sanitized, the literature is even scarcer.

Traditionally, the detection of sensitive elements within a text has been tackled in a manual way. This approach requires a human expert who applies certain standard guidelines that detail the correct procedures to sanitize sensitive entities [18]. Manual redaction has proven to be a quite time-consuming process and does not scale with regard to the requirements of the Information Society [3, 6]. First, the industry and the academia tried to mitigate this issue by providing semi-automatic proposals that highlight potentially sensitive elements from input text and that leave the final decision about erasing, generalizing or keeping them to the human expert [5, 8, 19]. However, these schemes still require the interaction of knowledgeable users and, hence, they suffer similar issues to their fully-manual counterparts.

In the last years, some *automatic redaction* methods have been designed. On the one hand, schemes that use specific/tailored patterns to detect certain types of information (e.g., names, addresses, Social Security numbers, etc.) can be found in [11, 37, 39]. Instead of patterns, schemes such as [6] use databases that contain the entities (people, products, diseases, etc.) to be eliminated from the text. Due to the ad hoc design of both approaches, they can be hardly generalized to cover a wide range of document types and application domains. In order to provide a more general solution, the authors in [28] propose a general method that assumes that sensitive terms are those that, due to their specificity, provide too much information. Hence, by quantifying the amount of information given by textual terms, sensitive terms (i.e., the *too informative* ones) are identified and removed.

One of the main drawbacks of the schemes introduced in the last paragraph is that they focus on redacting and, hence, they may heavily reduce the utility of the resulting documents. Fortunately, *automatic sanitization* methods have been also proposed. Authors in [1] provide a quite general scheme that uses a trained Named Entity Recognition (NER) package (i.e., the Stanford NER [13]) to automatically recognize entities belonging to general categories such as person, organization and location names. This mechanism proposes generalizing the sensitive entities instead of removing them from the sanitized document. Its goal is to achieve a certain degree of privacy while preserving some of the semantics. The authors in [14] provide a theoretic measure ("t-plausibility") on the quality of sanitized documents from a privacy protection point of view. Their scheme tries to preserve the utility of sanitized documents by

means of generalizing terms using a general-purpose ontology/taxonomy such as WordNet [12]. Finally, [27] presents a system that relies on Information Theory to quantify the amount of information provided by each term of the document. The work is built on the basis presented in [28], although, [27] successfully addresses the generalization of the sensitive terms using WordNet and the Open Directory Project (ODP) as general-purpose knowledge bases.

An important limitation of most of the above-described methods is that the sensitivity of textual terms is evaluated independently. As discussed previously, this situation may produce disclosure of sensitive terms if semantically correlated ones appear in clear form in the resulting sanitized document [36], a situation that may render useless the whole sanitization process [2].

Schemes that consider *term correlations* usually focus on *redaction* instead of *sanitization.* For example, in [6], which has been introduced previously, authors detect sensitive elements by means of a database of entities and a linked context (a set of terms related to each entity). Entities and their contexts must be manually specified a priori. Therefore, it could only be applied to documents with very specific scopes since it suffers from the same scalability and generality problems of manual approaches. A more general and flexible approach is presented in [29, 30]. This work presents an automatic redaction method that gets as input a set of terms that have been already detected as sensitive by another redaction system, and assesses whether they are semantically correlated with any other terms of the document that have been left in clear form. This proposal relies on the foundations of the Information Theory to provide a general-enough solution that can be automatically applied to heterogeneous textual documents. As introduced in the first section, the work in this paper extends [30] to achieve the goals detailed in the Introduction.

Finally, few *sanitization* schemes consider *term correlations*. In [8] and [2] the authors present two schemes that fall into this category, even though both present some limitations. More specifically, [8] introduces a supervised method that uses the Web to detect possible term inferences, but it lacks a strong theoretical basis (the analysis is driven by ad-hoc parameters) and the sanitization criterion is left for the user. On the other hand, [2] uses a contingency table that quantifies the degree of correlation between each pair of textual terms and taxonomy modeling term generalizations. Unfortunately, the availability of such accurate contingency table and associated taxonomy in a general setting is quite unrealistic, thus hampering the applicability of the method.

3. An information theoretic framework for disclosure risk assessment

In order to minimize the risk of disclosing sensitive data by the presence of semantically-related terms, two tasks should be performed: (i) quantification of the disclosure risk as a function of the degree of correlation between sensitive terms and other terms appearing in the same context (e.g. sentence); and (ii) sanitization of those terms whose degree of correlation is high enough to produce a feasible disclosure (according to a privacy criterion). Regarding the first task, during the assessment of disclosure risk, we should consider the fact that sensitive terms may be either removed (redacted) or generalized (sanitized) prior publication. Obviously, term generalizations produce higher risks of disclosure (which should be adequately measured) since semantics of sensitive data have not been completely removed but made less detailed. Regarding the second task, we propose a utility-preserving sanitization method in which highly correlated terms are generalized (rather than removed) while fulfilling the privacy requirements.

This section presents a general framework that, by relying on the foundations of the Information Theory, it tackles the above-described tasks.

3.1. Formalization of the sanitization scenario

Since this work is designed as an extension of the approach presented in [30], our method is also designed as a complementary step to any sanitization mechanism in which sensitive terms are managed independently (such as [1, 6, 27, 28, 37]). For coherency, we use a similar notation as in [30] to formalize the sanitization scenario:

- D: It represents the input textual document. No particular structure is assumed.
- $C_i \subseteq D$: It corresponds to each of the textual contexts in D, which will bind the scope of the correlation analysis. Usual context lengths may cover immediate adjacent words (like in [2, 33]), sentences or paragraphs (as in[23, 30, 40]), or the whole document (like in [29, 38]).
- ζ_1: It represents the initial (manual or automatic) sanitization mechanism applied to D. It is assumed to be any of the approaches discussed in Section 2, which detect sensitive terms independently. According to the specific mechanism, sensitive terms can be proposed for either *removal/redacted* [6, 11, 28, 37, 39] or *generalization/sanitization* [1, 14, 27] prior publication.
- D': It is the output of the mechanism ζ_1, in which sensitive terms have been individually removed or generalized, that is, $D'=\zeta_1(D)$.

- $s_{ij} \in C_i$: They are each of the textual terms in C_i that have been found to be sensitive by the sanitization mechanism ζ_1. For each context C_i, the set of sensitive terms is defined as $S_i=\{s_{i1}, s_{i2}, \ldots, s_{in}\}$. If sensitive terms are proposed for generalization (instead of removal) by ζ_1, each s_{ij} will be associated with a suitable generalization: $g(s_{ij})$.
- $C_i' \subseteq D'$: They correspond to the sanitized versions of each textual context in D.
- $q_{ik} \in C_i'$: They are each of the textual terms appearing in the sanitized context C_i', which can be either (i) a generalization of a term detected as sensitive by ζ_1; or (ii) a non-sensitive term left in clear form by ζ_1. Notice that both terms may cause disclosure of any s_{ij} in S_i (i.e., in the same context), due to semantic correlation. The set of terms in C_i' is defined as $Q_i=\{q_{i1},\ldots, q_{im}\}$.
- ζ_2: It represents an implementation of the method proposed in this paper that, by taking D' as input, it identifies those terms in each context C_i' that may disclose any of the sensitive terms in the same context (i.e., in S_i).
- $qs_{ik} \in C_i'$: They correspond to any term in C_i' that has been found, by ζ_2, as potentially risky for any sensitive term in S_i. Private-enough generalizations of these risky terms, which will be used as replacement in the sanitized output, are denoted as $g(qs_{ik})$.
- D'': It is the output of the proposed mechanism ζ_2, that is, $D''= \zeta_2 (D')$, in which those terms qs_{ik}, which may cause disclosure of sensitive ones, have been replaced by appropriate generalizations, $g(qs_{ik})$.

3.2. Measuring the informativeness of textual terms

The cornerstone of our sanitization mechanism ζ_2 is the quantification of the amount of information given by any textual term that, in terms of Information Theory, corresponds to its Information Content (IC). The IC of a term (e.g. a sensitive one, s_{ij}) can be computed as the inverse of its probability of occurrence in corpora.

$$IC(s_{ij}) = -\log_2 p(s_{ij}) \quad (1)$$

In this manner, general terms provide less information that more specific ones, since the former are more likely to appear in a discourse.

Applied to the sanitization context, the quantification of the IC of a term proposed for sanitization is indeed measuring the amount of information that, according to the sanitization criterion of ζ_1, should be hidden because of its *sensitive* nature. If s_{ij} (e.g., sensitive diseases such as *breast cancer*, or the city of residency of an individual such as *Miami*) is proposed for

removal by ζ_1, none of this sensitive information will be left in the sanitized output. This obviously minimizes the disclosure risk of s_{ij}, but also hampers the utility of the output, since the reader would gain no information about s_{ij}. Moreover, the removal of certain pieces of text also raises awareness of potential attackers of document's sensitivity [3]. On the contrary, when the mechanism ζ_1 proposes replacing the sensitive term s_{ij} by an appropriate generalization $g(s_{ij})$ (e.g. *breast cancer -> disease, Miami -> city*), $IC(g(s_{ij}))$ measures the amount of information of s_{ij} that can remain in clear form. This last strategy also preserves a degree of utility of the sensitive information, while maintaining a reasonable level of disclosure according to the privacy criterion of ζ_1.

At a conceptual level, the information given by a generalization $g(s_{ij})$ is a strict subset of the information given by its specialization s_{ij} (see Figure 1). This is coherent to the extent that, if both the specialization and the generalization appear in the same context, the latter provides no additional information to the former.

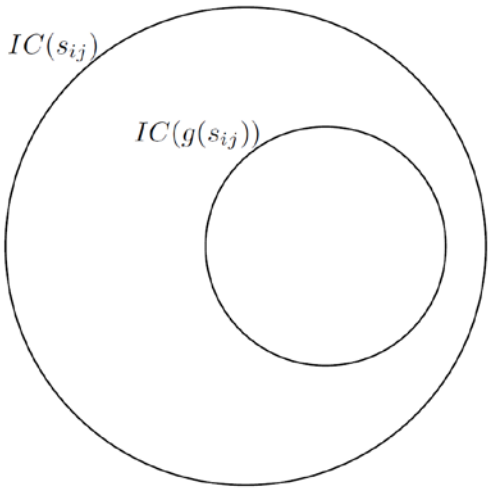

Figure 1. The amount of information of a generalization of a term is a strict subset of the information given by that term.

In terms of probability, the above relationship is fulfilled if term generalization occurrences are counted whenever any of the specializations occur [22], which results in monotonically increasing probabilities, that is, $p(g(s_{ij})) > p(s_{ij})$.

3.3. Measuring the disclosure risk of sensitive terms

If the information given by textual terms (either sensitive or not) co-occurring in a context is independent, the sanitization performed by a mechanism ζ_1, which evaluates terms

independently, will be optimal. This happens because no information of a sanitized term can be gained from another one. However, since human discourses are semantically coherent, textual terms are rarely independent (i.e., they do not co-occur by chance) and most of them are semantically related up to a degree [2] (e.g. a *disease* and its *treatments* or *symptoms*, a *city* and its relevant *points of interest* or *citizens*). As a consequence of the presence of semantic relations, the information given by a pair of terms usually overlaps. In terms of Information Theory, the information overlap between two variables can be measured as their *Mutual Information* (MI). The instantiation of MI for two specific outcomes (i.e., textual terms, in this case) is their *Point-wise Mutual Information* (PMI), whose meaning is graphically represented in Figure 2.

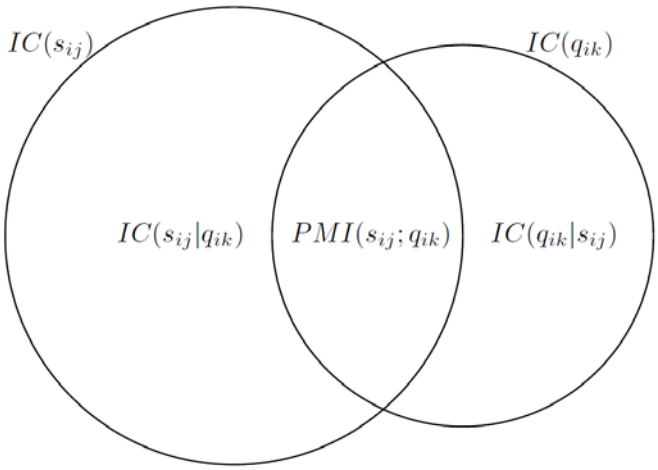

Figure 2. Graphical representation of the amount of information corresponding to the IC of a pair of terms and to their PMI and conditional ICs.

PMI measures the degree of information overlap between two terms as the difference between their normalized probability co-occurrence (e.g., between a sensitive term, s_{ij}, and any another term q_{ik} appearing in the same context), given their join distribution and marginal distributions [9]:

$$PMI(s_{ij};q_{ik}) = \log_2 \frac{p(s_{ij},q_{ik})}{p(s_{ij})\,p(q_{ik})} \quad (2)$$

From a probabilistic perspective, whenever s_{ij} occurs, q_{ik} tends to occur, so the information given by both terms will significantly overlap. Hence, given a sensitive term s_{ij} (e.g. *breast cancer*, *Miami*), the fact that another term, q_{ik}, with a tendency to co-occur (e.g. *radiotherapy*, *Florida*), appears in the same context means that q_{ik} is revealing an *amount of sensitive information* of s_{ij}, which can be measured as $PMI(s_{ij};q_{ik})$. By relying on this notion, in [30] it is

proposed to measure the *disclosure risk* of a sensitive term (assumed to be removed prior publication) with regards to other terms co-occurring in the same context, according to their PMI.

$$DR(s_{ij}; q_{ik}) = PMI(s_{ij}; q_{ik}) \quad (3)$$

Numerically, whenever q_{ik} occurs, s_{ij} also occurs, that is, if $p(s_{ij},q_{ik})=p(q_{ik})$, their PMI value maximizes to $PMI(s_{ij};q_{ik})=IC(s_{ij})$. Hence, the disclosure risk of s_{ij} is maximal because the presence of q_{ik} in a context completely discloses s_{ij}. On the contrary, if s_{ij} and q_{ik} are independent, that is, if they only co-occur by chance, then $p(s_{ij},q_{ik})=p(s_{ik})\cdot p(q_{ik})$ and, hence, $PMI(s_{ij};q_{ik})=0$. In this case, the presence of q_{ik} does not provide any information of s_{ij} and, thus, the disclosure risk of the latter is null. PMI may also produce negative values if s_{ij} and q_{ik} are mutually exclusive, thus resulting in a minimal value of PMI= $-\infty$ if they never co-occur (i.e., $p(s_{ij},q_{ik})=0$). However, since most textual terms are semantically correlated up to a degree [2], negative PMI values are rare. In fact, low or even null co-occurrences between terms are usually attributed to data sparseness of probability calculus, rather than to a real exclusiveness [26].

Since $p(s_{ij},q_{ik})$ can be rewritten in terms of conditional probability ($p(s_{ij}|q_{ik})p(q_{ik})$), PMI can be also expressed as:

$$PMI(s_{ij}; q_{ik}) = \log_2 \frac{p(s_{ij} \mid q_{ik})\, p(q_{ik})}{p(s_{ij})\, p(q_{ik})} = \log_2 \frac{p(s_{ij} \mid q_{ik})}{p(s_{ij})} \quad (4)$$

Since the *conditional information content* provided by s_{ij} given the presence of q_{ik} is $IC(s_{ij}|q_{ik})=-log_2(p(s_{ij}|q_{ik}))$, PMI can be expressed in terms of Information Content as follows:

$$PMI(s_{ij}; q_{ik}) = \log_2 \frac{p(s_{ij} \mid q_{ik})}{p(s_{ij})} = \log_2(p(s_{ij} \mid p_{ik})) - \log(p(s_{ij})) = IC(s_{ij}) - IC(s_{ij} \mid q_{ik}) \quad (5)$$

The expression $IC(s_{ij})$ - $IC(s_{ij}|q_{ik})$ emphasizes the fact that $PMI(s_{ij};q_{ik})$ measures how much –sensitive- information q_{ik} is telling us about s_{ij} (recall Figure 2).

In [30], a theoretical and empirical study shows the adequacy of PMI to measure the disclosure risk of sensitive terms that are *removed* prior publication.

3.4. Assessing the disclosure risk of generalized terms

The above characterization of disclosure risk assumes that sensitive terms s_{ij} (e.g. *breast cancer*, *Miami*), detected by the initial sanitizer ζ_1, will be removed prior publication, and that other terms q_{ik} (e.g. *radiotherapy*, *Florida*) co-occurring in the same context will be left in clear form. Thus, an attacker would only gain information about s_{ij} from q_{ik}. Figure 3 illustrates this gain of information in grey, which corresponds to the PMI of s_{ij} and q_{ik}.

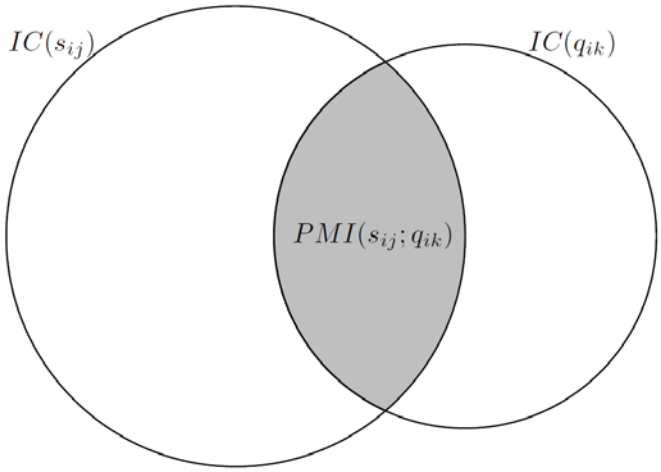

Figure 3. Gain of information (in grey) of s_{ij} from the presence of a correlated term q_{ik}.

Even though the removal of sensitive terms is the usual strategy implemented by many redacting mechanisms [6, 11, 28, 37, 39], as discussed in Section 2, this strategy hampers the utility of the output. This is why more sophisticated sanitizers [1, 14, 27] opt to replace sensitive terms by less detailed generalizations (e.g. *breast cancer -> disease*, *Miami-> city*). As discussed in Section 3.2, generalizations embrace a *strict* subset of the information carried by the generalized term. From the perspective of disclosure risk, if sensitive terms are generalized rather than removed, an attacker would gain information about a sensitive entity s_{ij} both from its generalization ($g(s_{ij})$) and from its overlapping, with semantically related terms (q_{ik}). This is shown in grey in Figure 4.

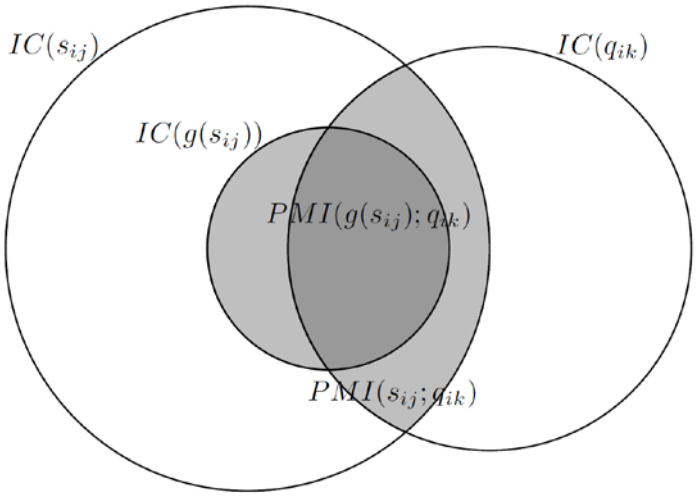

Figure 4. Areas in grey show the information gained of s_{ij} from its generalization ($g(s_{ij})$) and the presence of a correlated term q_{ik}.

Due to the additional information given by $g(s_{ij})$, this setting results in a higher risk of disclosure than in the scenario discussed in the previous section. To obtain realistic DR figures, the calculi implemented by ζ_2 should be extended to also consider the information provided by $g(s_{ij})$. To do so, the IC of the generalization (i.e., $IC(g(s_{ij}))$ should be added to the PMI of correlated terms (i.e., $PMI(s_{ij};q_{ik})$). Moreover, because the IC of $g(s_{ij})$ is a subset of s_{ij}, as shown in Figure 4 in darker grey, it is likely that the IC of $g(s_{ij})$ and the IC of the correlated term q_{ik} also overlap. The amount of information corresponding to this overlap specifically corresponds to $PMI(g(s_{ij});q_{ik})$. In order to avoid counting this overlap of information twice when adding $PMI(s_{ij};q_{ik})$ and $IC(g(s_{ij}))$, $PMI(g(s_{ij});q_{ik})$ must be subtracted from the DR calculus, as follows:

$$DR(s_{ij}; q_{ik}) = PMI(s_{ij}; q_{ik}) + IC(g(s_{ij})) - PMI(g(s_{ij}); q_{ik}) \quad (6)$$

This expression can be extended according to the equivalence stated in eq. (5):

$$DR(s_{ij}; q_{ik}) = PMI(s_{ij}; q_{ik}) + IC(g(s_{ij})) - (IC(g(s_{ij})) - IC(g(s_{ij}) \mid q_{ik})) = \\ = PMI(s_{ij}; q_{ik}) + IC(g(s_{ij}) \mid q_{ik}) \quad , \quad (7)$$

which emphasizes the fact that, in this scenario, the disclosure risk is the sum of the information that q_{ik} (left in clear form) provides about $s_{ij,}$ and the additional information that $g(s_{ij})$ (also left in clear form) provides given that we know q_{ik}.

Notice that eq. (6) results in eq. (3) if sensitive terms are removed rather than generalized by ζ_1 since, in that case, $IC(g(s_{ij}))=0$ and $PMI(g(s_{ij});q_{ik})=0$. Thus, eq. (6) represents a generalization of the DR assessment, regardless of the sanitization strategy implemented by ζ_1.

3.5. Defining the sanitization criterion

Once the disclosure risk of sensitive terms has been quantified by our mechanism ζ_2, we need to detect which ones cause *too much* risk. In such case, appropriate sanitization measures over q_{ik} should be taken by ζ_2 to reduce that risk. Thus, we need to set a *threshold* value for DR according to a privacy criterion.

As proposed in [30], the disclosure risk threshold (t_{DR}) considered by ζ_2 should be coherent to the sanitization criterion used by the initial sanitization mechanism ζ_1 to detect sensitive terms. In this manner, our method will seamlessly adapt its behavior to different input sanitizers and privacy criteria. To do so, we propose defining t_{DR} according to the evidences given by ζ_1 on its sanitization criterion. On the one hand, we assume that the set of terms tagged as sensitive by ζ_1 are those that provide *too much information* according to ζ_1's criterion (either fixed or manually defined) [27, 28]. For example, methods based on NER assume that Named Entities should be sanitized because, in general, they provide more information (due to their higher concreteness) than regular words [1]. Other works specifically state that those terms that provide more information than a user-defined value should be sanitized [27, 28]. In these cases, the IC of the *least informative sensitive term* (i.e., s_{least}) from all sets of sensitive terms S_i of all textual contexts C_i in D reflects the lower bound of the sanitization criterion implemented by ζ_1. On the other hand, sanitizers in which sensitive terms s_{ij} are replaced by their generalizations $g(s_{ij})$ implicitly state that, even though s_{ij} is too informative to be left in clear form, $g(s_{ij})$ is general enough to not revealing *too much* information. In this scenario, the IC of the *most informative generalization* (i.e., g_{most}) from all S_i represents the upper bound of the sanitization criterion of ζ_1. Thus, we propose to automatically define a sanitization threshold (t_{DR}) for our mechanism ζ_2 that is in coherency with ζ_1, as either:

$$t_{DR} = IC(s_{least}) = \min_{\forall S_i}(\min_{\forall s_{ij} \in S_i}(IC(s_{ij}))) \quad (8)$$

or

$$t_{DR} = IC(g_{most}) = \max_{\forall S_i}(\max_{\forall s_{ij} \in S_i}(IC(g(s_{ij})))) \quad (9)$$

The suitability of one value or the other will depend on the sanitization technique implemented by ζ_1 and also on the availability of term generalizations. Notice that we define one t_{DR} for all sets of sensitive terms S_i (one per textual context C_i), since we assume that the sanitization criterion of ζ_1 is the same for the entire document.

It is important to note that, since t_{DR} measures the *amount of –sensitive– information*, it can be coherently compared with DR values. Thus, any term q_{ik} in C_i', for which the DR with regard to any s_{ij} in S_i is equal or above t_{DR}, states a higher-than-desired risk of disclosure according to the implicitly privacy criterion of ζ_1; such terms, which we denote as qs_{ik}, should thus be sanitized by ζ_2.

3.6. Utility-preserving sanitization of semantically correlated terms

Once the set of risky terms QS_i has been identified by ζ_2, the final step consists in sanitizing each qs_{ik} from the output, so that the disclosure risk is lowered up to acceptable levels. In approaches like [7, 29, 30], qs_{ik} will be simply removed from the output. Again, this severely hampers the utility of sanitized documents, a problem that is aggravated by the potentially large amount of semantically correlated terms that usually co-occur in tight and focused discourses.

To improve the utility of the output, we propose replacing each qs_{ik} by a generalization. To enable an automatic sanitization, term generalizations should be retrieved from available knowledge bases (KB) that offer a taxonomic structure of textual terms; that is, for a given qs_{ik} (e.g. *radiotherapy*, *Florida*), an ordered set of generalizations $H_{ik}=h_1$->… -> h_n (e.g. *radiotherapy -> therapy -> treatment* or *Florida -> U.S. State -> State -> territorial division*) can be obtained.

Thus, given a sensitive term s_{ij} (and its corresponding generalization $g(s_{ij})$ if available), the optimal generalization of a correlated term qs_{ik} with respect to s_{ij} will be such from those in H_{ik} that, while fulfilling the desired level of privacy towards s_{ij}, they preserve as much semantics of qs_{ik} as possible; we denote the generalization of qs_{ik} with respect to s_{ij} as $g_{sij}(qs_{ik})$. To fulfill the privacy criterion (i.e. the sanitization threshold t_{DR}), $g_{sij}(qs_{ik})$ must decrease enough the disclosure risk towards s_{ij} (considering also its generalization $g(s_{ij})$, if available) so that:

$$DR(s_{ij}; g_{s_{ij}}(qs_{ik})) = PMI(s_{ij}; g_{s_{ij}}(qs_{ik})) + IC(g(s_{ij})) - PMI(g(s_{ij}); g_{s_{ij}}(qs_{ik})) < t_{DR} \quad (10)$$

Note that the expression is the same as eq. (6) but replacing qs_{ik} by a potential generalization, $g_{sij}(qs_{ik})$, which would replace qs_{ik} in the sanitized output.

From an Information Theoretic perspective, each generalization step of qs_{ik} reduces the amount of disclosed information and, hence, (i) the utility of the output and (ii) the degree of overlapping with the sensitive term s_{ij} and consequently, the associated disclosure risk as well.

Then, the optimal generalization from those in H_{ik} towards s_{ij}, that is, $g_{sij}(qs_{ik})$, will provide the *maximum information* while fulfilling t_{DR} towards s_{ij}.

$$g_{s_{ij}}(qs_{ik}) = \underset{\forall h_l \in H_{ik} | DR(s_{ij};h_l) < t_{DR}}{\arg\max} (IC(h_l)) \tag{11}$$

Figure 5 shows, in concentric circles, the amount of information (measured as $PMI(s_{ij};g_{sij}(qs_{ik}))$) given by progressive generalizations of qs_{ik}; an intermediate generalization (non-dashed inner circle) is taken as the one that fulfills the sanitization threshold (i.e., it decreases enough the amount of disclosed information –areas in grey– towards s_{ij}) and maximizes the utility.

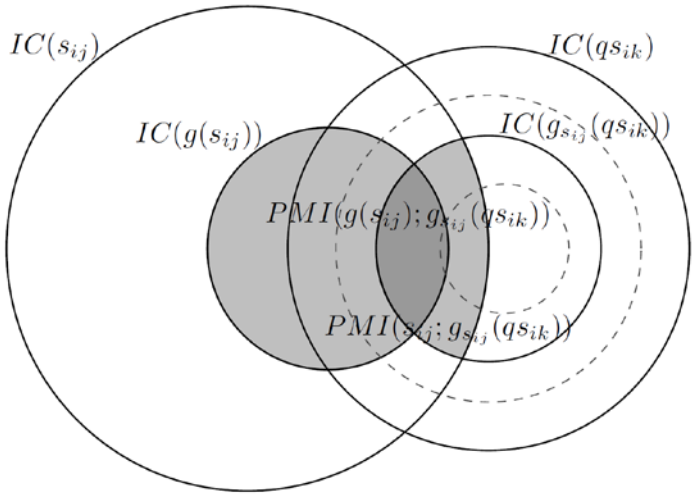

Figure 5. Successive generalizations of qs_{ik} decrease the disclosed amount of information of s_{ij}; an intermediate generalization (non-dashed inner circle) is taken as the optimal $g_{sij}(qs_{ik})$.

Since qs_{ik} may produce disclosure towards *several* sensitive terms s_{ij} in a certain context, different optimal generalizations $g_{sij}(qs_{ik})$ may be obtained for each s_{ij}. The generalization of (11), leads to the final generalization of qs_{ik} that will be used as a replacement in the output (referred as just $g(qs_{ik})$) will be such that provides the *maximum information* while fulfilling the privacy criterion for *all* s_{ij}.

$$g(qs_{ik}) = \underset{\forall h_l \in H_{ik} | DR(s_{ij};h_l) < t_{DR}}{\arg\max} (IC(h_l)) \quad \forall s_{ij} \in S_i \tag{12}$$

4. Framework implementation

In this section, the practical implementation of the theoretical framework is detailed. We discuss which techniques are used to extract terms from a text, how to calculate term probabilities for IC/PMI calculi, and which available knowledge bases can be exploited to obtain term generalizations. Finally, the proposed sanitization mechanism, ζ_2 is depicted in an algorithmic way.

4.1. Term extraction

In order to perform the analysis, the proposed framework requires the extraction of textual contexts in the input document (e.g. sentences, paragraphs, the whole document, etc.) and of terms within that context. In this work, we specifically focus on terms referring to concepts, since those are the ones that carry most of the semantics of the discourse; those can be referred in text as nouns (e.g. *melanoma*) or, more generally, as noun phrases (NPs) (e.g. *breast cancer*).

To extract contexts and NPs, we rely on a set of natural language-processing tools[2] by which it is possible to (i) *detect sentences*, (ii) *identify tokens* (i.e., individual words within a sentence), (iii) perform a *part-of-speech (POS) tagging* of each token and (iv) *syntactically parse* POS tagged tokens to detect NPs. Moreover, to avoid altering subsequent probability calculus (see Section 4.2), we also remove *stop words* from extracted NPs. Stop words are domain independent words like determinants, prepositions or adverbs whose removal do not affect the conceptualization to which the NP refers (e.g., *the Florida State -> Florida State*). Finally, since knowledge bases rarely include the different lexicalizations of concept labels (e.g. plurals), individual terms are *stemmed* [21] to remove derivative morphological suffixes (e.g. *cancers -> cancer*); this eases the matching between extracted terms and concept labels included in knowledge bases from which term generalization should be retrieved (see Section 4.3).

4.2. Computation of term probabilities

Disclosure risk assessment extensively relies on IC/PMI calculi, which are in turn a function of the join and marginal probabilities of terms. To compute probabilities, authors usually rely on tagged textual data in which words are semantically linked to their conceptual abstractions in a taxonomy/ontology. Even though this approach solves ambiguity problems that are inherent to words, it is also severely hampered by the limited coverage of tagged data. Thus, data sparseness usually affects the probability calculus, especially for domain-specific terms (e.g.

[2] OpenNLP, Apache Software Foundation. Available at: http://opennlp.apache.org

chemical compounds), NEs (e.g., organization names) or newly coined terms (e.g., a new drug) [26]. Unfortunately, due to their informativeness, these terms are a usual target of document sanitization.

An alternative to tagged corpora is offered by the Web. The use of the Web as corpora provides interesting benefits:

- Thanks to its size, heterogeneity and dynamicity, it has been argued that it represents the current distribution of terms at a social scale [4]. Thus, it enables general-purpose probability calculus with minimal data sparseness.
- To compute these probabilities, Web Search Engines (WSEs) can be used as proxies, since the page count provided by a WSE when querying a term, divided by the total number of web resources indexed by the WSE, reasonably approximates term probability at a Web scale [4]. In fact, WSEs have already been exploited to compute term probabilities in a variety of tasks [31, 32, 34, 38, 40, 41].
- From the perspective of document sanitization, some authors have considered the Web (accessed via WSEs) as an excellent proxy for public knowledge and, thus, it represents a reasonable approximation of the knowledge that a potential attacker may have and use to disclose sensitive data [7, 36].

Given the above, we compute the probability of a term *t* from the *page count* provided by WSE when querying the *t*:

$$p(t) = \frac{page\_count("t")}{total\_webs}, \tag{13}$$

where *total_webs* is the number of web sites indexed by the WSE.

Join probabilities of a set of terms $t_1 \ldots t_n$ are computed from their number of co-occurrences in the Web, as follows:

$$p(t_1,...,t_n) = \frac{page\_count("t_1" \; AND \; ... \; AND \; "t_n")}{total\_webs} \tag{14}$$

There are, however, some issues that limit the accuracy of the Web-based probability calculus. Indeed, the page count provided by a WSE for a specific term may be affected by language ambiguity (e.g. polysemy, synonymy or ellipsis). As a result, the informativeness of a term may be underestimated (i.e. if the term is polysemous and it thus appears in a different context, each one with a different sense) or overestimated (i.e. due to synonymy or ellipsis, appearances of a

term may not be considered by the strict string-matching analysis implemented by WSEs). These imperfections are however mitigated by the size and heterogeneity of Web data, which contributes to provide robust statistics [4], and also because of its high redundancy, which helps to compensate inaccuracies caused by synonymy or polysemy [23].

Another aspect to consider is related to the number of queries performed to the WSE. Due to the cost of online access to the Web, it is interesting to minimize the number of queries performed to the WSE that are required to compute the DR of a term. If we expand eq. (6) in terms of probabilities, we obtain:

$$DR(s_{ij};q_{ik}) = PMI(s_{ij};q_{ik}) + IC(g(s_{ij})) - PMI(g(s_{ij});q_{ik}) = \\ = \log\frac{p(s_{ij},q_{ik})}{p(s_{ij})\,p(q_{ik})} - \log p(g(s_{ij})) - \log\frac{p(g(s_{ij}),q_{ik})}{p(g(s_{ij}))\,p(q_{ik})} \quad , \qquad (15)$$

which relies on *five* probabilities and, thus, *five* queries to be performed to the WSE, to measure the DR caused by q_{ik}: $p(s_{ij},q_{ik})$, $p(s_{ij})$, $p(q_{ik})$, $p(g(s_{ij}))$ and $p(g(s_{ij}), q_{ik})$).

However, by rearranging the eq. (15), we can obtain the following simplified expression in terms of probabilities:

$$DR(s_{ij};q_{ik}) = \log\frac{p(s_{ij},q_{ik})}{p(s_{ij})\,p(q_{ik})} - \log p(g(s_{ij})) - \log\frac{p(g(s_{ij}),q_{ik})}{p(g(s_{ij}))\,p(q_{ik})} = \\ = \log\frac{\dfrac{p(s_{ij},q_{ik})}{p(s_{ij})\,p(q_{ik})}}{p(g(s_{ij}))\dfrac{p(g(s_{ij}),q_{ik})}{p(g(s_{ij}))\,p(q_{ik})}} = \log\frac{p(s_{ij},q_{ik})}{p(s_{ij})\,p(g(s_{ij}),q_{ik})} \quad , \qquad (16)$$

which requires just *three* probabilities/queries to compute the DR: $p(s_{ij},q_{ik})$, $p(s_{ij})$, $p(g(s_{ij}), q_{ik})$). As a result, the efficiency of DR assessment is significantly improved.

The same simplified result can be achieved for the expression that evaluates the suitability of a generalization of q_{ik} (eq. (10)), but replacing q_{ik} by $g(q_{ik})$.

4.3. Knowledge bases

As detailed in Section 3.6, our approach replaces those q_{ik} that may cause disclosure of sensitive data by an appropriate generalization. To do so, a KB (e.g., structured thesaurus, folksonomies, ontologies, etc.) offers a suitable taxonomy from which extract term generalizations are needed.

In this section, we discuss what makes a KB well-suited for that purpose and which ones are used in the implementation of our approach.

First, the KB should provide a high *recall* of terms contained in the input document because, if a term is not covered by the KB, the only options left are (i) to replace it by the root of the taxonomy (e.g. World) [27], (ii) to replace it by a random entity [1] or (iii) to remove it [6], which will produce a significant loss of information. Moreover, a KB with a *fine grained taxonomic tree* will be desirable in order to reduce the loss of information resulting from each generalization step, and to better approximate the optimal generalization (i.e., the generalization that, while maximizing the amount of given information, fulfills the sanitization threshold).

In a general-purpose scenario, several methods [5, 14] have used WordNet [12] to generalize sensitive data. WordNet is a domain-independent knowledge base that describes more than 100,000 concepts linked by means of semantic relationships; it thus offers a detailed taxonomy created by knowledge experts. Unfortunately, its coverage of NEs and highly specific NPs is very limited [27].

An alternative to WordNet is the Open Directory Project (ODP)[3]. ODP is the largest, most comprehensive human-edited directory of the Web. It contains more than 1 million of manually created categories organized in a taxonomical form and edited by thousands of contributors. Due to its dynamicity, ODP offers a high recall for NEs and newly coined terms; however, due to its distributed editing, its taxonomic structure is not as coherent and as correct as WordNet's [27].

In the implementation of our framework, we use both WordNet and ODP, with preference to the former, in cases in which a term is found in both KBs. Moreover, in order to improve the recall of the conceptual mapping, if the NP corresponding to a qs_{ik} (e.g. *curable breast cancer*) is not found in any KB, we try simplified versions of the NP by iteratively removing nouns/adjectives starting from the one most on the left (e.g. *curable breast cancer -> breast cancer*). After removing each word, we also check if the simplified version (e.g. *breast cancer*) already fulfills the sanitization threshold and, thus, we directly use it as a replacement. Only if the simplest form of qs_{ik} is not found in any KB, we replace it by the most abstract generalization of the taxonomies (i.e., *world*).

[3] http://www.dmoz.org/docs/en/about.html

4.4. Algorithm

In this section, we detail the practical implementation of the sanitizer, ζ_2, which implements the framework presented in Section 3. The inputs of the algorithm are: (i) the original document D and (ii) the output of the initial sanitizer ζ_1 in which sensitive terms are detected independently: $D' = \zeta_1(D)$. As parameters, the set of KBs used to retrieve generalizations (i.e. WordNet and OPD) and the length of textual contexts (e.g. sentences, documents) are considered. The output of the algorithm is a document D'' in which both sensitive terms and those that may disclose the former are replaced by generalizations.

Algorithm 1. Sanitizer ζ_2

```
Input:  D           //the original document
        D'=ζ1(D)    //the output of the initial sanitizer
Output: D''         //final sanitized document

1   D''=empty;
2   t_DR=compute_threshold(D'); //According to eq. (8) or (9)
3   for each (C_i' ⊆ D') do //According to the length of textual contexts
4     S_i=getSanitizedTerms(C_i',D); //Terms sanitized by ζ1 in the context C_i'
5     GS_i=getSanitizedGeneralizations(C_i',D); //The generaliz. proposed by ζ1
6     NS_i=getNon-sanitizedTerms(C_i',D);
7     Q_i= GS_i ∪ NS_i; //The set of terms in the context C_i' to analyze
8     QS_i=empty; //The set of risky terms for the context C_i'
9     for each (s_ij ∈ S_i) do
10      Q_iToRemove=empty;
11      for each (q_ik ∈ Q_i) do
12        if (DR(s_ij, q_ik, g(s_ij)) > t_DR) then //According to eq. (6)
13          qs_ik= q_ik; //The term is considered as risky for s_ij
14          g(qs_ik)=empty;
15          p=0;
16          H_ik=getGeneralizations(qs_ik, KBs); //Terms generaliz. from the KBs
17          while (g(qs_ik)=empty) do //Look for a suitable generalization
18            if (DR(s_ij, H_ik[p], g(s_ij))) < t_DR, ∀s_ij ∈ S_i) then //As in eq. (10)
19              g(qs_ik)=H_ik[p];
20              put(<qs_ik, g(qs_ik)>, QS_i); //store the term and its generaliz.
21              if (qs_ik ∈ GS_i) then //qs_ik is a generaliz. of a sensitive term
22                replace(qs_ik, g(qs_ik), GS_i); //replace by the new generaliz.
23              end if
24              put (qs_ik, Q_iToRemove); //qs_ik is already treated
25            end if
26            p++;
27          end while
```

```
28        end if
20      end for
30      removeTerms(Qi ToRemove, Qi); //remove already treated terms from Qi
31    end for
32    D''=D''+sanitize(Ci',QSi); //Attach the context and sanitize risky terms
33  end for
34  return D'';
```

The first step of the algorithm is the computation of the sanitization threshold (t_{DR}) as detailed in Section 3.5 (line 2). Recall that, in order to compute the threshold, all the terms already sanitized by ζ_1 in *D'* are considered.

Then, the algorithm individually extracts (according to the context length) and analyses each context C_i' in *D'* (line 3). The already sanitized terms S_i of C_i', their generalizations GS_i proposed by ζ_1, if available, and the non-sanitized terms NS_i of C_i' are obtained (lines 4-6). The set of terms Q_i to be analyzed by our method is composed by both the generalizations sensitive terms detected by ζ_1 and the set of non-sensitive terms left in clear form (line 7, as defined in Section 3.1).

Next, the algorithm computes the DR of each sensitive term s_{ij} in S_i for each term q_{ik} in Q_i (line 12). If the DR value is higher than the threshold t_{DR}, the term q_{ik} is considered risky (qs_{ik}) with regard to s_{ij} (line 13). To reduce this risk, qs_{ik} will be sanitized in the output. To do so, the set of generalizations of qs_{ik}, ordered from the most specific to the most general one, is obtained from the KBs (line 16, as explained in Section 4.3). Then, the DR of each generalization $H_{ik}[p]$ is computed as in eq. (10) (line 18). If the DR is lower than the threshold t_{DR} for all sensitive terms s_{ij} in S_i, that generalization is considered the most suitable one for qs_{ik} (line 19) and it is added together with qs_{ik} to the vector of terms to sanitize QS_i (line 20). It is important to note that disclosure risk may happen due to a generalization of a term already sanitized by ζ_1. In such case (line 21), the generalization is replaced by a more abstract one fulfilling the threshold (line 22), which will be considered in the next iterations of the algorithm. Moreover, since the generalizations of risky terms are picked so that they fulfill the threshold with regard to *any* sensitive term in S_i, there is no need to consider them in further iterations (i.e., other s_{ij}). Thus, they are stored (line 24) and removed from Q_i (line 30) before analyzing the next sensitive term s_{ij}.

The process is repeated until there are no elements in Q_i that can disclose *any* term s_{ij} in S_i. As a result, for each context C_i', a vector QS_i of risky terms and their suitable generalizations are

obtained. This vector is used to sanitize risky terms in the current context in the output document *D''* (line 32).

5. Evaluation

In this section, we evaluate the performance of the implementation of our method, ζ_2, from two perspectives: (i) the accuracy in detecting terms that may disclose sensitive data and (ii) the preservation of utility of the sanitized output. The approach presented in this paper has been directly compared with the previous work from [30] (which already reported an improvement over works based on Named Entity recognition like [1, 5]), whereas the utility evaluation has reproduced some of the usual protection strategies implemented by related works [1, 6, 11, 14, 27, 28, 30, 37, 39]. As input sanitizer to detect sensitive terms, ζ_1, we used the state-of-the-art *Stanford Named Entity Recognizer* (SNER) [13]. Notice that the detection of NEs is one of the most usual and less constrained automatic methods for document sanitization [1, 41]. SNER evaluates input terms individually, and relies on a trained classifier to detect NEs (which are tagged as sensitive) and to classify them as *persons*, *locations* and *organizations*. These classification labels are used as generalizations, $g(s_{ij})$, for sensitive terms.

5.1. Evaluation data

In order to enable a fair comparison against the previous work, we used the same evaluation data as in [30]. It consists of a set of Wikipedia English articles describing entities of different domains. Given the high informativeness and tight discourse of Wikipedia articles, these documents represent a challenging scenario for document sanitization, with a potentially large number of semantically correlated terms. As detailed in [30], articles have been picked up so that they describe *persons*, *organizations* and *locations* in order to offer a favorable scenario for the input NER-based sanitizer.

As described in Section 3.5, our proposal is able to adapt its behavior to the sanitization criterion of ζ_1 by defining a sanitization threshold (t_{DR}), according to the IC of the least informative sensitive term (eq. (8)) or the IC of the most informative generalization (eq. (9)). Given that NE-based generalizations are quite constrained (only three different categories are considered), we opted for the first option, which enables a finer-grained assessment of the sanitization threshold. Moreover, this also enables a direct comparison against the results reported in [30], which employs the same strategy.

The set of evaluated articles are listed in Table 1. For each entity we also show the term that, according to the strategy detailed in Section 3.3, acts as sanitization threshold.

Table 1. Wikipedia articles used for evaluation with associated threshold terms.

Wikipedia Article	*Threshold term*
Steve Wozniak	Steve Jobs
Steven Spielberg	Spielberg
Tom Cruise	Magnolia
Arnold Schwarzenegger	California
Sylvester Stallone	Stallone
Audrey Hepburn	London
Antoni Gaudi	Spain
Antonio Banderas	Antonio Banderas
Javier Bardem	Boca
Jordi Mollà	United States
Dreamworks	LLC
Microsoft	United States
Apple	United States
Aston Martin	England
Volkswagen	Audi
Port Aventura	Europe
Yellowstone	North America
Barcelona	Europe
Tarragona	Spain
Salou	Spain

5.2. Evaluating the detection accuracy

The first evaluation measures the accuracy of our method in detecting terms that may enable disclosure of the sensitive ones identified by ζ_1. These results have been compared to those reported in [30] in which, for the same entities, sensitive terms detected by ζ_1 are assumed to be removed/redacted prior publication (recall DR assessment from Section 3.3). On the contrary in this paper (recall the extended DR assessment from Section 3.4), sensitive terms can be replaced by their corresponding NE generalizations (*person*, *location* or *organization*, as detailed above), thus contributing to improve the utility of the output. In both cases, textual contexts (C_i) have been set to the whole article length, which is coherent to the tight discourse of Wikipedia articles.

In order to fairly compare the results of both scenarios, the evaluation has been done in the same manner as in [30]. Three human experts were requested to select and agree on which terms

(either removed or generalized) appearing in the same context as sensitive terms would feasibly reveal any of the latter. Hereinafter, we refer to this set of terms as *Human_QS*. By comparing *Human_QS* with the output of each test, we measure their performance according to the usual metrics of *precision*, *recall* and *F-measure*.

Precision (eq. (17)) quantifies the percentage of automatically identified terms that may cause disclosure (*QS*), which have been also identified by human experts (*Human_QS*) from the total number of detected terms (*QS*). The higher the precision is, the better the utility of the output, because the number of non-necessary sanitizations is lower.

$$Precision = \frac{|QS \cap Human\_QS|}{|QS|} \times 100 \tag{17}$$

Recall (eq. (18)) measures the percentage of terms in *Q*S that have been also identified by human experts from the total amount of terms in *Human_QS*. Recall states the amount of risky terms (according to the human criterion) that the automatic methods were able to detect. Thus, it reflects the degree of privacy of the output.

$$Recall = \frac{|QS \cap Human\_QS|}{|Human\_QS|} \times 100 \tag{18}$$

F-measure (eq. (19)) provides an aggregation (harmonic mean) of precision and recall that summarizes the accuracy of each method.

$$F - measure = \frac{2 \times Recall \times Precision}{Recall + Precision} \tag{19}$$

Notice that recall plays a more important role than precision in document sanitization. A low recall implies that a number of terms that may disclose sensitive data will appear in clear form in the sanitized output. Given that the disclosure of a sole sensitive term may negate the sanitization [2], a high recall figure is crucial to provide robust privacy guarantees.

Table 2 shows the evaluation results when sensitive terms detected by ζ_1 are: (i) *removed* from the output, as in [30], or (ii) replaced by their *generalized* versions, as supported in the current approach. Recall that, as ζ_1, we used the state-of-the-art *Stanford Named Entity Recognizer* (SNER) [13].

Table 2. Evaluation of terms that may enable disclosure of sensitive data for different Wikipedia articles assuming (i) the removal of sensitive terms (as in [30]) or (ii) their replacement with a generalization, as proposed in this work.

Wikipedia Article	*Precision*		*Recall*		*F-Measure*	
	Removal	*Generalization*	*Removal*	*Generalization*	*Removal*	*Generalization*
Steve Wozniak	100%	100%	75%	75%	85,71%	85,71%
Steven Spielberg	71,42%	80%	55,5%	66,66%	62,5%	72,72%
Tom Cruise	100%	85%	82,35%	94,44%	90,32%	89,47%
Arnold Schwarzenegger	68,75%	60,86%	100%	100%	81,48%	75,66%
Sylvester Stallone	50%	50%	66,66%	66,66%	57,14%	57,14%
Audrey Hepburn	66,66%	62,5%	100%	100%	80%	76,92%
Antoni Gaudi	51,28%	53,33%	90,90%	92,3%	65,57%	67,6%
Antonio Banderas	80%	75%	57,14%	100%	66,66%	85,71%
Javier Bardem	77,77%	75%	100%	100%	87,5%	85,71%
Jordi Mollà	53,84%	50%	100%	100%	70%	66,67%
Dreamworks	35%	35%	100%	100%	51,85%	51,85%
Microsoft	48,57%	48,57%	100%	100%	65,38%	65,38%
Apple	58,33%	61,7%	84%	100%	68,85%	76,92%
Aston Martin	55,55%	50%	100%	100%	71,42%	66,67%
Volkswagen	80%	71,42%	66,66%	83,33%	72,72%	76,92%
Port Aventura	57,89%	52,17%	100%	100%	73,33%	68,57%
Yellowstone	71,42%	60%	90,90%	100%	80%	75%
Barcelona	31,88%	42,68%	88%	94,6%	46,80%	58,82%
Tarragona	50%	50%	100%	100%	66,66%	66,66%
Salou	52,94%	47%	100%	100%	69,23%	63,95%
Average	*63,06%*	*60,51%*	*87,86%*	*93,65%*	*70,65%*	*71,67%*

Results obtained in both settings are quite similar (averaged F-measures do not show significant differences). However, with the approach presented in this paper, where sensitive terms are generalized rather than removed, precision is, in average, slightly lower; on the other hand, recall is equal or higher in all cases. In fact, a perfect recall is obtained for 13 articles (only 10 for the previous method based on term removal), thus reflecting a more robust sanitization than that of [30]. This is especially relevant, given that the approach in [30] provided high recall figures (87,36%, in average) and that the experiments conducted in [30] already showed a much higher recall than that of related works based on Named Entity recognition (like [1, 5]). The relative differences are coherent with the behavior of each approach. When removing sensitive terms [30], DR is measured only from the amount of information disclosed by a non-sanitized term (eq. (3), Section 3.3). On the other hand, when generalizing sensitive terms, DR is the result of the sum of information given by the generalization and the information disclosed by other terms (generalized or not) appearing in the same context. Thus, in this latter case, DR figures tend to be higher. Given that the same threshold is used in both scenarios, the latter case

results in a more exhaustive sanitization, which is reflected in the higher recall figures but also in the larger amount of false positive that produces a lower precision. Certainly, the slightly lower precision may affect the utility of the output, since more terms will be unnecessarily sanitized. However, as it will be evaluated in the next section, this issue will be more than compensated by the fact that the approach presented in this paper replaces risky terms by appropriate generalizations, rather than removing them. There are, however, some cases in which the stricter sanitization behavior of the current approach produces an improvement in precision due to the correct identification of a number of false negatives (which also improves recall).

The analysis of individual results reflects notable differences. For some of them (e.g. *Steve Wozniak*, *Tom Cruise*), very high precisions are obtained, whereas for other ones (e.g. *Barcelona*, *Microsoft*), precision falls below 50%. These differences are related to the term used to define t_{DR}, that is, the threshold value that guides the sanitization process (see Table 1). For entities with a very general threshold term (e.g. *Europe* for *Barcelona*), results are less accurate than for those with more concrete ones (e.g. *Steve Jobs* for *Steve Wozniak*). The variability of the degree of generality of the threshold term (and, hence, of the t_{DR} value) is caused by the sanitization criterion implemented by ζ_1. Since it is based on NE recognition, all terms identified as NEs are systematically proposed for sanitization. Even though NEs are usually more informative (and hence, more sensitive) than normal words, many NEs are so general (e.g. country names) that they hardly pose any risk. Since ζ_2 behavior depends on the sanitization performed by ζ_1, when very general NEs are identified, ζ_2 is forced to implement a stricter sanitization that may produce a number of false positives and a lower precision. This, however, illustrates the adaptability of ζ_2 with regard to the –even though imperfect- sanitization implemented by ζ_1.

5.2. Evaluating output's utility

The second part of the evaluation measures the degree of utility of the sanitized output and compares different strategies implemented by related works to sanitize sensitive and correlated terms.

To quantify output's utility, we calculate the ratio between the *amount of information* provided by the sanitized document with regard to the original version. The amount of information given by a document D (i.e., $IC(D)$) is measured as the sum of the IC of all terms appearing in D.

$$IC(D) = \sum_{\forall q_k \in D} IC(q_k) \quad (20)$$

Thus, the utility of the sanitized output *D'* is expressed as the percentage of information from the original document *D* that is preserved in the output *D'* [27], as follows:

$$Utility(D') = \frac{IC(D')}{IC(D)} \times 100 \quad (21)$$

In Table 3, utility figures obtained for the entities and thresholds presented in the previous section are detailed. To contextualize and compare our results against related works, the following sanitization strategies have been implemented:

- *S1*: only those terms tagged as sensitive by ζ_1 are removed. Correlated terms that may cause disclosure are not considered and, hence, they will appear in clear form in the output. This reproduces the usual strategy implemented by redacting methods [6, 11, 28, 37, 39], as discussed in Section 2.
- *S2*: the same as *S1*, but sensitive terms are replaced by generalized versions; this corresponds to the behavior of sanitization methods that focus on preserving document's utility [1, 14, 27]. In our case, these generalizations correspond to NER classes.
- *S3*: both sensitive terms tagged by ζ_1 and correlated terms detected by ζ_2 that may cause disclosure of the former are removed. This corresponds to the behavior of the previous work in [30].
- *S4*: sensitive terms tagged by ζ_1 are removed, whereas correlated terms detected by ζ_2 are replaced by generalized versions. This setting simulates a simplified version of the scenario tackled in this paper, in which the generalizations of sensitive terms are not considered in the DR assessment.
- *S5*: both terms detected by ζ_1 and by ζ_2 are replaced by generalizations. This corresponds to the complete scenario described in this paper, which aims at preserving as much utility as possible while fulfilling the privacy criterion.

Table 3. Degree of utility preserved in the sanitized output according to the sanitization strategy implemented for sensitive terms (by ζ_1) and for correlated ones (by ζ_2).

Wikipedia Article	*S1*	*S2*	*S3*	*S4*	*S5*
Steve Wozniak	59,67%	73,11%	22,02%	47,47%	61,29%
Steven Spielberg	68,43%	75,52%	20,79%	48,81%	51,52%
Tom Cruise	74,11%	83,86%	32,4%	53,96%	55,63%
Arnold Schwarzenegger	70,51%	77,08%	7,22%	37,39%	43,25%
Sylvester Stallone	43,64%	56,76%	13,57%	28,01%	39,7%
Audrey Hepburn	51,51%	65,72%	5,46%	26,89%	31,53%
Antoni Gaudi	69,85%	77,34%	1,89%	41,65%	38,59%
Antonio Banderas	44,77%	63,29%	21,74%	29,51%	28,36%
Javier Bardem	59,9%	71,99%	20,65%	42,92%	57,05%
Jordi Mollà	47,02%	68,45%	9,37%	32,17%	47,84%
Dreamworks	46,14%	62,83%	6,27%	26,63%	42,81%
Microsoft	70,84%	81,34%	0,43%	35,13%	46,54%
Apple	85,13%	90,94%	8,11%	50,23%	47,06%
Aston Martin	36,06%	55,51%	2,52%	17,24%	32,86%
Volkswagen	29,01%	54,54%	2,24%	16,47%	32,07%
Port Aventura	52,41%	69,34%	12,82%	31,51%	44,53%
Yellowstone	70,38%	83,05%	17,91%	47,16%	56,8%
Barcelona	60,41%	73,99%	2%	34,21%	39,21%
Tarragona	70,08%	83,55%	28,31%	59,14%	59,4%
Salou	57,91%	72,53%	18,21%	43,25%	46,52%
Average	*58,39%*	*72,04%*	*12,7%*	*37,49%*	*45,13%*

First, we observe that strategies *S1* and *S2* are the ones that result in the more useful sanitization outputs, since only sensitive terms are sanitized (an average of 58,39% and 72,04%, respectively). Obviously, *S2* retains more utility, since terms are generalized rather than removed. However, as discussed through this paper, the individual sanitization of sensitive terms is not enough to avoid disclosure due to the presence of semantically correlated terms. In fact, the empirical evaluation carried in [30] for the same entities showed that the practical privacy of related works corresponding to the strategies *S1* and *S2* is almost half of the privacy achieved when correlated terms are also considered. Hence, even though these approaches produce usable outputs, privacy guarantees are weak. Utility figures for these scenarios are however interesting, because they define the upper bound of utility preservation when correlated terms are also considered (strategies *S3*, *S4* and *S5*).

When both sensitive and correlated terms are removed (*S3*), which corresponds to the approach detailed in [30], privacy is significantly improved, but the utility of the output is drastically reduced (12,7% in average). Individual figures are well below a 10% of utility preservation for many entities (e.g. *Arnold Schwarzenegger*, *Audrey Hepburn*, *Antoni Gaudi*, *Dreamworks*,

Microsoft, *Aston Martin*, *Volkswagen*, *Barcelona*), stating that sanitized outputs will be hardly usable both for human readers and also for data analysis. The reason for such low figures is the tight and highly informative discourse that usually characterizes Wikipedia articles, in which almost all terms are highly correlated.

Finally, utility is significantly better preserved when correlated terms are generalized, both when sensitive terms are removed (*S4*, with and average preservation of 37,49%) or generalized (*S5*, with an average preservation of 45,13%). Moreover, these figures are just a 21%-27% below the upper bounds defined by *S1* and *S2* (37,49% vs. 58,39% and 45,13% vs. 72,04%, respectively). These differences quantify the cost in utility preservation derived from the sanitization of correlated terms (*S4* and *S5*), that is, the cost of the more robust privacy guarantees. Also notice that, the utility improvement shown by *S5* (45,13%) came at no cost in output's privacy, as shown in Table 2. In fact, according to recall figures reported in Table 2, privacy is even better for *S5* than for *S3* and *S4* (93,54% vs. 87,86%, in average) due to the stricter sanitization criterion implemented by *S5*, in which sensitive terms are generalized rather than removed. Moreover, as discussed in the previous section, the slightly lower precision achieved by our generalization-based approach in the previous test (60,02% vs. 63,06%, in average as shown in Table 2), which may negatively affect utility due to the larger number of false positives, is more than compensated by the fact that the proposed strategy *S5* better preserves the semantics of correlated terms.

6. Conclusions and future work

This paper tackles the document redaction/sanitization problem with a special focus on two aspects that are commonly neglected in available solutions: (i) the preservation of the output's utility and (ii) the detection and sanitization of terms that may cause disclosure of sensitive data due to semantic correlation. The proposed theoretical framework offers a general model, grounded in the Information Theory, to measure the disclosure risk of term occurrences/co-occurrences regardless of the fact that they are proposed for removal or generalization. The framework can be exploited to implement general-purpose solutions regardless the domain of knowledge (such as the one described in Section 4), and also domain-specific applications (i.e., using domain-specific KBs and/or corpora). The evaluation illustrated the suitability of the approach for a set of heterogeneous entities, which showed a significant improvement of output's utility (in comparison to the previous work) while achieving a level of privacy stronger than that of related works.

As future work, some lines of research can be defined:

- The current model considers semantic correlations between term pairs. Notwithstanding this is enough to detect most risky terms [29] in many cases, disclosure may still happen by the combination of *several* sanitized and non-sanitized terms. To tackle this problem, a generalization of the expressions proposed in Section 3 is required, so that the DR of term triples/quartets/etc. is appropriately measured. Notice, however, that the larger the cardinality of the set to evaluate is, the harder the compilation of the required probabilities will be, due to data sparseness.
- As stated in Section 4.2, the accuracy of the probabilities acquired by querying a WSE is hampered by language ambiguity. On the one hand, polysemic queries may overestimate term appearance frequency whereas, on the other hand, synonymy and ellipsis may underestimate it. These issues are caused by the lack of semantic disambiguation that characterizes keyword-based WSEs. To tackle them, the ambiguity of WSE queries can be reduced by adding more terms that help contextualize the query. For example, by querying a term together with an appropriate conceptual generalization (e.g. “cancer” + “disease”), the effect of polysemy in the page count can be minimized [26]. As an alternative to corpora-based probability calculus, intrinsic IC models that estimate concept’s IC unambiguously according to the taxonomic knowledge modeled in the KBs can be used [24, 25].
- Additional evaluations can be carried out by implementing domain-specific sanitizers, based on the proposed framework and comparing them with ad-hoc solutions. The biomedical domain seems specially suited for this purpose due to the sensitive nature of clinical data, the availability of the resources in which the sanitizer relies (i.e., medical ontologies [35] and tagged corpora [20]), and the amount of ad-hoc redaction systems [17].

Acknowledgements and disclaimer

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, ICWT TIN2012-32757, CO-PRIVACY TIN2011-27076-C03-01, ARES-CONSOLIDER INGENIO 2010 CSD2007-00004, Audit Transparency Voting Process IPT-430000-2010-31 and BallotNext IPT-2012-0603-430000), by the Spanish Ministry of Industry, Commerce and Tourism (through projects eVerification2 TSI-020100-2011-39 and SeCloud TSI-020302-2010-153) and by the Government of Catalonia (under grant 2009 SGR 1135).

References

[1] D. Abril, G. Navarro-Arribas, V. Torra, On the declassification of confidential documents, in: Modeling Decision for Artificial Intelligence. 8th International Conference, MDAI 2011, Springer, 2011, pp. 235–246.

[2] B. Anandan, C. Clifton, Significance of term relationships on anonymization, in: IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology - Workshops, Lyon, France, 2011, pp. 253–256.

[3] E. Bier, R. Chow, P. Golle, T. H. King, J. Staddon, The Rules of Redaction: identify, protect, review (and repeat), IEEE Security and Privacy Magazine, 7 (2009) 46-53.

[4] R.L. Cilibrasi, P.M.B. Vitányi, The Google Similarity Distance, IEEE Transactions on Knowledge and Data Engineering, 19 (2006) 370-383.

[5] C. Cumby, R. Ghani, A machine learning based system for semiautomatically redacting documents, in: Twenty-Third Conference on Innovative Applications of Artificial Intelligence, San Francisco, California, USA, 2011, pp. 1628–1635.

[6] V.T. Chakaravarthy, H. Gupta, P. Roy, M.K. Mohania, Efficient techniques for document sanitization, in: 17th ACM Conference on Information and Knowledge Management (CIKM'08), Napa Valley, California, USA, 2008, pp. 843–852.

[7] R. Chow, P. Golle, J. Staddon, Detecting Privacy Leacks Using Corpus-based Association Rules, in: 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, Las Vegas, NV, 2008, pp. 893-901.

[8] R. Chow, I. Oberst, J. Staddon, Sanitization's Slippery Slope: The Design and Study of a Text Revision Assistant, in: 5th Symposium on Usable Privacy and Security, New York, USA, 2009.

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

[10] J. Domingo-Ferrer, A Survey of Inference Control Methods for Privacy-Preserving Data Mining, in: C.C. Aggarwal, P.S. Yu (Eds.) Privacy-Preserving Data Mining, Springer, 2008, pp. 53-80.

[11] M. Douglass, G. Cliffford, A. Reisner, W. Long, G. Moody, R. Mark, De-identification algorithm for free-text nursing notes, in: Computers in Cardiology, 2005, pp. 331–334.

[12] C. Fellbaum, WordNet: An Electronic Lexical Database, MIT Press, Cambridge, Massachusetts, 1998.

[13] J. Finkel, T. Grenager, C. Manning, Incorporating non-local information into information extraction systems by gibbs sampling, in: 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05), Ann Arbor, USA, 2005, pp. 363–370.

[14] W. Jiang, M. Murugesan, C. Clifton, L. Si, t-plausibility: Semantic preserving text sanitization, in: International Conference on Computational Science and Engineering (CSE'09), Vancouver, Canada, 2009, pp. 68–75.

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

[16] S. Martínez, D. Sánchez, A. Valls, A semantic framework to protect the privacy of electronic health records with non-numerical attributes, Journal of Biomedical Informatics, 46 (2013) 294-303.
[17] S.M. Meystre, F.J. Friedlin, B.R. South, S. Shen, M.H. Samore, Automatic de-identification of textual documents in the electronic health record: a review of recent research, BMC Medical Research Methodology, 10 (2010).
[18] National Security Agency, Redacting with confidence: How to safely publish sanitized reports converted from word to pdf, in, 2005.
[19] National Security Agency, Redaction of pdf files using Adobe Acrobat Professional X, in, 2011.
[20] T. Pedersen, S. Pakhomov, S. Patwardhan, C. Chute, Measures of semantic similarity and relatedness in the biomedical domain, Journal of Biomedical Informatics, 40 (2007) 288-299.
[21] M.F. Porter, An Algorithm for suffix stripping, in: Readings in Information Retrieval, 1997, pp. 313-316.
[22] P. Resnik, Using Information Content to Evalutate Semantic Similarity in a Taxonomy, in: C.S. Mellish (Ed.) 14th International Joint Conference on Artificial Intelligence, IJCAI 1995, Morgan Kaufmann Publishers Inc., Montreal, Quebec, Canada, 1995, pp. 448-453.
[23] D. Sánchez, A methodology to learn ontological attributes from the Web, Data & Knowledge Engineering 69 (2010) 573-597.
[24] D. Sánchez, M. Batet, A New Model to Compute the Information Content of Concepts from Taxonomic Knowledge, International Journal on Semantic Web and Information Systems, 8 (2012) 34-50.
[25] D. Sánchez, M. Batet, D. Isern, Ontology-based Information Content computation, Knowledge-based Systems, 24 (2011) 297-303.
[26] D. Sánchez, M. Batet, A. Valls, K. Gibert, Ontology-driven web-based semantic similarity, Journal of Intelligent Information Systems, 35 (2010) 383-413.
[27] D. Sánchez, M. Batet, A. Viejo, Automatic general-purpose sanitization of textual documents, IEEE Transactions on Information Forensics and Security, 8 (2013) 853-862.
[28] D. Sánchez, M. Batet, A. Viejo, Detecting sensitive information from textual documents: an information-theoretic approach, in: Modeling Decisions for Artificial Intelligence. 9th International Conference, MDAI 2012, Springer, 2012, pp. 173-184
[29] D. Sánchez, M. Batet, A. Viejo, Detecting Term Relationships to Improve Textual Document Sanitization, in: 17th Pacific Asia Conference on Information Systems, Jeju Island, South Korea, 2013, pp. Paper 105.
[30] D. Sánchez, M. Batet, A. Viejo, Minimizing the disclosure risk of semantic correlations in document sanitization, Information Sciences, 249 (2013) 110-123.
[31] D. Sánchez, J. Castellà-Roca, A. Viejo, Knowledge-based scheme to create privacy-preserving but semantically-related queries for web search engines, Information Sciences, 218 (2013) 17-30.
[32] D. Sánchez, D. Isern, Automatic extraction of acronym definitions from the Web, Applied Intelligence, 34 (2011) 311-327.
[33] D. Sánchez, A. Moreno, Pattern-based automatic taxonomy learning from the Web, AI Communications, 21 (2008) 27-48.

[34] D. Sánchez, A. Moreno, L.D. Vasto-Terrientes, Learning relation axioms from text: An automatic Web-based approach, Expert Systems with Applications, 39 (2012) 5792-5805.
[35] K. Spackman, SNOMED CT milestones: endorsements are added to already-impressive standards credentials, Healthcare Informatics, 21 (2004) 54-56.
[36] J. Staddon, P. Golle, B. Zimmy, Web-based inference detection, in: 16th USENIX Security Symposium on USENIX Security Symposium, 2007, pp. Article No. 6.
[37] L. Sweeney, Replacing personally-identifying information in medical records, the scrub system, in: 1996 American Medical Informatics Association Annual Fall Symposium, Washington, DC, USA, 1996, pp. 333–337.
[38] P.D. Turney, Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL, in: L. De Raedt, P. Flach (Eds.) 12th European Conference on Machine Learning, ECML 2001, Springer-Verlag, Freiburg, Germany, 2001, pp. 491-502.
[39] A. Tveit, O. Edsberg, T.B. Rost, A. Faxvaag, O. Nytro, T. Nordgard, M.T. Ranang, A. Grimsmo, Anonymization of general practioner medical records, in: second HelsIT Conference, 2004.
[40] C. Vicient, D. Sánchez, A. Moreno, An automatic approach for ontology-based feature extraction from heterogeneous textual resources, Engineering Applications of Artificial Intelligence, 26 (2013) 1092-1106.
[41] A. Viejo, D. Sánchez, J. Castellà-Roca, Preventing automatic user profiling in Web 2.0 applications, Knowledge-Based Systems, 36 (2012) 191-205.