Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Utility-preserving privacy protection of textual healthcare documents

David Sánchez *, Montserrat Batet, Alexandre Viejo

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

ARTICLE INFO

Article history:
Received 5 December 2013
Accepted 25 June 2014
Available online xxxx

Keywords:
Privacy-protection
Document redaction
Data sanitisation
Healthcare data
Semantics
Information theory

ABSTRACT

The adoption of ITs by medical organisations makes possible the compilation of large amounts of healthcare data, which are quite often needed to be released to third parties for research or business purposes. Many of this data are of sensitive nature, because they may include patient-related documents such as electronic healthcare records. In order to protect the privacy of individuals, several legislations on healthcare data management, which state the kind of information that should be protected, have been defined. Traditionally, to meet with current legislations, a manual redaction process is applied to patient-related documents in order to remove or black-out sensitive terms. This process is costly and time-consuming and has the undesired side effect of severely reducing the utility of the released content. Automatic methods available in the literature usually propose ad-hoc solutions that are limited to protect specific types of structured information (e.g. e-mail addresses, social security numbers, etc.); as a result, they are hardly applicable to the sensitive entities stated in current regulations that do not present those structural regularities (e.g. diseases, symptoms, treatments, etc.). To tackle these limitations, in this paper we propose an automatic sanitisation method for textual medical documents (e.g. electronic healthcare records) that is able to protect, regardless of their structure, sensitive entities (e.g. diseases) and also those semantically related terms (e.g. symptoms) that may disclose the former ones. Contrary to redaction schemes based on term removal, our approach improves the utility of the protected output by replacing sensitive terms with appropriate generalisations retrieved from several medical and general-purpose knowledge bases. Experiments conducted on highly sensitive documents and in coherency with current regulations on healthcare data privacy show promising results in terms of the practical privacy and utility of the protected output.

1. Introduction

New technologies have played a crucial role on improving healthcare delivery. Data digitalisation, in particular, has paved the way for the extensive adoption of Electronic Health Records (EHR) systems that have enabled clinicians and researchers to access, manage and exploit large amounts of valuable patient data easily [23]. Nevertheless, as data becomes more accessible and easily copied and transferred, the confidentiality of the patients is more likely to be jeopardised.

Redaction is a privacy-preserving method that aims to avoid (or at least mitigate) the disclosure of raw confidential data, such as textual documents (in contrast with specific privacy protection methods focusing only on relational databases [12,22]). Redaction is based on blacking-out, obscuring or eliminating sensitive terms in the documents prior to their release. Selecting which elements of the document have to be *redacted* is crucial, because a weak redaction process may disclose sensitive data. On the other hand, a too restrictive redaction may destroy the utility of the document, a situation that goes against the purpose of data releasing.

Official regulations have been developed at this respect within the medical context. For example, the Health Insurance Portability and Accountability Act (HIPAA) [11], states safe harbour rules about the kind of personally identifiable information which should be removed in medical documents prior allowing their publication. More specifically, the HIPAA requires 18 data elements (called PHI: Protected Health Information) to be removed from a redacted document. The goal of such regulation is to maintain individual's *anonymity* while preserving healthcare outcomes, which are useful for medical research, intact.

In other scenarios, such as when medical records are released to insurance companies or legal counsel [34] to be used as a support for legal claims (e.g. workers' compensation claims and motor vehicle accident claims), privacy protection regulations are focused towards ensuring the *confidentiality* of individual's data. In those

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

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

http://dx.doi.org/10.1016/j.jbi.2014.06.008
1532-0464/© 2014 Elsevier Inc. All rights reserved.

cases, documents have to be still linked to a patient but any information that may impair her dignity and/or be the cause of discrimination must be removed. Regarding this last point, many US state and federal laws stipulate that elements such as HIV status, drug or alcohol abuse and mental health conditions must be *redacted* before releasing medical records to third parties [10,20,37].

The main side effect of document redaction is that it significantly reduces the *utility* of the protected content [8]. Another important drawback is that the existence of obscured or blacked-out parts can raise the awareness of the document's sensitivity in front of potential attackers [4]. This is especially problematic in documents linked to a specific knowledge area such as healthcare, because the number of different textual elements that usually appear on the documents is relatively limited. Therefore, leaving blacked-out parts increases the probabilities for an attacker to determine the redacted terms by means of some characteristics such as their context and their length [21]. A more suitable alternative to document redaction consists on *generalising* sensitive content instead of removing it, a measure that preserves more content utility [8]. This is usually referred to as *document sanitisation* [4]. As a result of content generalisation a less detailed but still useful document is obtained, while no explicit clues about the document's sensitivity are given.

In any case, document redaction/sanitisation is significantly hindered by the need to deal with textual data and because of the existence of *semantic relationships* between the textual terms in the document. The fact that many sensitive terms lack a regular structure (e.g. disease names, in contrast to e-mail addresses or social security numbers) limits the effectiveness of automatic methods based on pattern matching or trained classifiers [23]. Semantic relationships, on the other hand, may enable the re-identification of redacted/sanitised elements from the presence of related terms left in clear form. For example, let us consider the following text in which the sensitive term *AIDS* has been redacted (i.e. replaced by XXX).

> "*The patient suffers from XXX that was transmitted because of an unprotected sexual intercourse. He was diagnosed when his immune system responded poorly to influenza.*"

Even though the terms left in clear form "unprotected sexual intercourse", "immune system" or "influenza" may seem apparently innocuous, an external observer with a certain knowledge on the domain [7] may effectively re-identify the redacted term *AIDS* by semantic inference; for example, by simply querying the three terms in a Web search engine, most of the resulting web pages are related to AIDS/HIV.

Providing privacy protection against an external observer who exploits semantic inferences to re-identify redacted/sanitised elements is a quite challenging task. According to [34], with the help of a Web search engine, an external observer can easily infer facts, reconstruct events and piece together identities from fragments of information collected from different sources. This situation is assumed to be even worse when dealing with medical records, whose tight focus would result quite commonly in the presence of *strong* relationships between the different elements (diseases, symptoms, treatments, medication, etc.).

Due to the above problems, redaction/sanitisation of medical documents is usually performed in a manual way by a human expert (or a team of human experts), who follows regulations and redaction guidelines detailing the correct procedures to sanitise sensitive entities [24]. This task has proven to be burdensome, very time-consuming [16] and prone to disclosure risks [6]. For example, the authors in [4] interviewed the medical records manager for a 10,000+ patient healthcare provider in California, who stated that the act of redacting records takes approximately 20% of her time while the remaining 80% is consumed by the more difficult task of deciding what to redact. In order to deal with textual terms and their potential semantic relationships, this expert maintains lists of names of medications, treatments, etc., which are related to sensitive diseases to be protected, such as HIV.

Given the burden of manual document redaction/sanitisation, the development of automatic schemes capable of dealing with textual healthcare data and avoiding the disclosure caused by semantic inference while retaining the utility of the output document, is a clear need for medical organisations. Unfortunately, as it will be discussed in Section 2, methods available in the literature are limited and they only partially cover some of those features.

1.1. Contributions and plan

In this paper, we describe an *automatic* scheme designed to *sanitise textual medical documents* (e.g. electronic healthcare records) without assuming any kind of structural regularity in the entities to protect. Moreover, it puts special emphasis in the detection and sanitisation of terms that are semantically related to sensitive entities, in order to avoid disclosure via semantic inference. The present work offers the following contributions:

- It applies an information theoretic notion of term sensitivity [30] to develop a sanitisation method for textual medical data, which is well-suited to fulfil with the privacy requirements stated by the current legislations on healthcare data protection.
- It applies and extends the notion of disclosure risk introduced in [31,32] to detect risky semantically related terms. Moreover, it exploits several general-purpose (WordNet [14] and ODP[1]) and healthcare (SNOMED-CT [33]) knowledge bases to preserve data utility.
- It proposes a more accurate way of computing term disclosure risk using the Web as corpora and a knowledge base to minimise semantic ambiguity.
- The proposed scheme is evaluated and compared against related works using documents describing highly sensitive medical concepts and realistic use cases based on existing regulations on medical data privacy.

The rest of the paper is organised as follows. Section 2 reviews related works in document redaction/sanitisation. Section 3 presents our proposal. Section 4 details the evaluation metrics and compares the results obtained for a collection of highly sensitive documents against previous works. The final section provides the conclusions.

2. Related work

As explained in the introduction, depending on the privacy requirements, protection of healthcare documents may pursue two different goals: (i) preserve the individual's *anonymity*; or (ii) ensure the *confidentiality* of her data. For the first case, the HIPAA rules [11] specify which identifying elements should not appear in any protected medical document in order to avoid re-identification. Regarding the second case, governmental legislations stipulate [10,20,37,38] which sensitive elements should be masked from any medical record (to avoid potential discrimination) before releasing it to a third party.

Most of the literature on the sanitisation/redaction related the medical domain focus on identifying terms. In a recent survey of redacting methods for electronic health records [23], all the 18 reviewed schemes focused on protecting those identifying

[1] http://www.dmoz.org (last accessed: November, 2013).

Please cite this article in press as: Sánchez D et al. Utility-preserving privacy protection of textual healthcare documents. J Biomed Inform (2014), http://dx.doi.org/10.1016/j.jbi.2014.06.008

elements stated in the HIPAA rules (e.g. ages, e-mails, locations, dates, social security numbers, etc.). A relevant characteristic of HIPAA elements is that they tend to present a regular structure (e.g. numbers of a certain length, dates, e-mails, etc.) or can take values from a finite set (e.g. locations). Thus, redaction schemes focused on these elements exploit these regularities to detect them in an automatic way. Because of this, as stated by the survey, they can be classified into schemes based on machine learning (e.g. [2,17,19,36,41]), which train classifiers to detect specific types of elements, and those based on pattern matching, which rely on the manual definition of regular expressions (e.g. [3,13,15,18,25,35,40]).

Because of their inherent design, the above systems can hardly support other kind of sensitive elements, especially in those cases in which no regular structure can be exploited. In fact, most sensitive elements related to patient's *confidentiality*, such as names of diseases, do not have a regular structure. In such cases, the disclosure risk is a function of their inherent semantics (i.e. the fact that they refer to a specific and sensitive matter), rather than of their type [7]; for example, diseases such as *flu* and *AIDS* in a medical record may appear to be mere plain words, but the latter is highly sensitive because of being potentially discriminatory [10,20,37]: AIDS has traditionally entailed social discrimination because of its usual transmission mechanism (which is related to sexual habits) and severity. Moreover, because of the lack of a proper semantic analysis, the above approaches cannot detect the existence of apparently innocuous terms (e.g. symptoms, treatments) that may re-identify a sensitive one by means of semantic inference, thus negating the redaction process.

Approaches that consider term semantics during document redaction/sanitisation are mainly *manual*. In [4,16] it is discussed the implications of a manual redaction process performed by an expert on the healthcare field (i.e. the medical records manager in a Hospital). In order to detect sensitive terms and risky semantic relationships, this expert uses her own knowledge on the field together with long lists of names of diseases, treatments, etc., which are also manually compiled and maintained. This results in a burdensome and time-consuming process [16], which is not exempt from disclosure risks [6].

The need of *automatic* sanitisation/redaction schemes that can be general enough to be applied to textual documents without assuming any kind of term structure have been acknowledged in the past in works like [4,7,8,34]. In these schemes, the idea is to rely on several semantic analyses that can help to estimate the sensitivity of textual terms in an unsupervised manner, similarly to what is done by a human expert. To retrieve and analyse term semantics, different information repositories can be exploited. More in detail, in [7] the authors provide a practical model of inference detection using a reference corpus and by considering word co-occurrences. As the authors acknowledge, inferences extracted from a large corpus can be used both to assist the document redaction process and to attack redacted outputs. Unfortunately, these works suffer from two limitations: (i) they only automatize the *redaction* process, that is, the detection and removal of sensitive terms; as a result, either the utility of the protected output will suffer due to term removal or the supervision of an expert will be required to propose appropriate generalisations; and (ii) the automatic method designed to detect sensitive terms relies on a set of ad-hoc parameters (e.g. absolute number of term occurrences/co-occurrences) that should be carefully tuned for each redaction scenario and that lack a sound theoretical justification.

3. Proposal

The goal of the proposed method is to automatically sanitise a textual medical document according to certain privacy requirements, which would be usually specified by current legislations on medical data privacy; even though, if needed, other sources of privacy requirements may be incorporated. This will be done in a way that (i) sensitive elements (defined by the privacy requirements), and also those apparently innocuous terms than can effectively re-identify the former by means of semantic inference, will be sanitised, and (ii) the sanitisation process will try to preserve the utility of the resulting document as much as possible. In the following, we call *terms* to (noun) phrases designating a concept (e.g. *AIDS* and *acquired immune deficiency syndrome* are terms referring to the same medical concept).

The general workflow of the proposed method is depicted in Fig. 1. It consists of two preliminary steps (*Step-0a* and *Step-0b*) that take the input privacy requirements (e.g. current legislations) and automatically set the method parameters; these steps are expected to be executed only once, just previously to the first sanitisation process. Then, two main steps (*Step-1* and *Step-2*), which are executed for each document to sanitise, will be in charge of detecting and protecting sensitive terms and those that are semantically related, according to the input privacy requirements. These steps are explained in detail in the next subsections.

3.1. Step-0a: Obtain the list of sensitive terms

This preliminary step takes as input current legislations on medical data privacy, such as the U.S. legislation [10,20,37], specifying the kind of entities that should be protected because they are potentially discriminatory (e.g. AIDS/HIV, mental diseases, Sexually Transmitted Diseases and drug or alcohol abuse). In any case, other privacy requirements may also be incorporated. Then, it uses a knowledge base (KB) to compile a list of terms that will be considered as sensitive and, thus, should not appear in any publicly disclosed medical document. We use SNOMED-CT as knowledge base. SNOMED-CT (Systemized Nomenclature of Medical Clinical Terms) covers more than 360,000 medical concepts that are taxonomically modelled in 18 partially overlapping hierarchies; individual concepts are associated to lists of equivalent terms (synonyms) [33].

The list of sensitive terms will include the different synonyms and lexicalisations provided in SNOMED-CT for each entity to protect (e.g. for STDs: sexually transmitted disease, venereal disease, VD, etc.) and also all of its taxonomic specialisations, which inherit the semantics of their –sensitive- ancestors (e.g. for STDs: gonorrhoea, syphilis, chlamydia, etc.). Hereinafter, we refer to the resulting list of sensitive terms as *S*. Notice that we do not explicitly consider the post-coordinated expressions that can be created from SNOMED-CT concepts to refer to other more complex concepts. In such cases, we would only refer to the parts of the expression that are considered as sensitive by the privacy criterion.

3.2. Step-0b: Compute the sanitisation threshold

The proposed method relies on the Information Theory and, in particular, on the notion of Information Content (IC) of textual terms to automatically guide the sanitisation process. The idea is to use the notion of IC to quantify the *amount of sensitive information* provided by any of the terms to protect. Given that terms in *S* define the privacy requirements of the sanitisation process, we assume that the amount of information (IC) they provide states the *baseline amount of sensitive information* that should not be revealed in the protected output in order to avoid disclosure [30]. As it will be detailed in the following sections, this baseline will act as *sanitisation threshold* to decide up to which level sensitive terms should be generalised in the sanitised output to meet with the privacy requirements and, also, which semantically

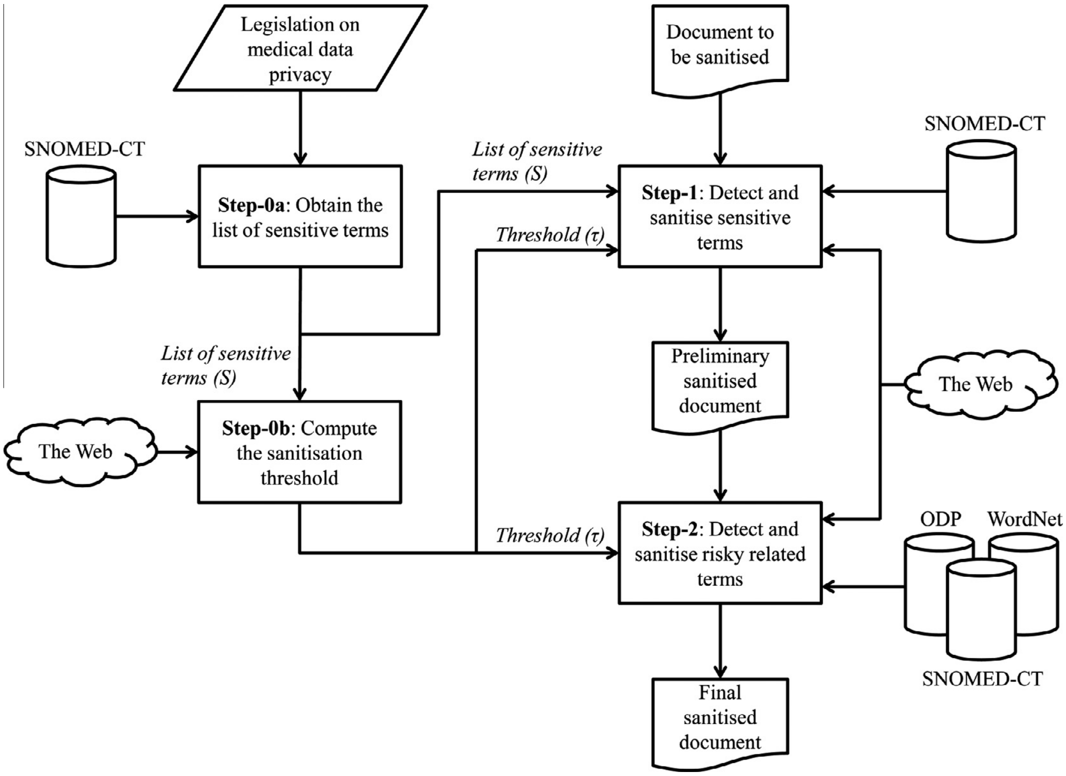
Legislation on medical data privacy
Document to be sanitised
SNOMED-CT
SNOMED-CT
List of sensitive terms (S)
Step-0a: Obtain the list of sensitive terms
Threshold (τ)
Step-1: Detect and sanitise sensitive terms
List of sensitive terms (S)
Preliminary sanitised document
The Web
The Web
Step-0b: Compute the sanitisation threshold
Threshold (τ)
Step-2: Detect and sanitise risky related terms
ODP
WordNet
SNOMED-CT
Final sanitised document

Fig. 1. General workflow of the proposed method.

related terms are risky because they disclose too much information about a sensitive entity.

Numerically, the *IC* of a term *t* is computed as the inverse of its probability of appearance, $p(t)$, in a corpus.

$$IC(t) = -\log_2 p(t) \tag{1}$$

In this manner, general terms (e.g. *disease*) will be assumed to provide less information than specialised terms (e.g. *Chlamydia*), because the former are more likely to be referred in a discourse.

To compute term probabilities, classical methods such as [27] used tagged textual data as corpora, which provided accurate probabilities when applied to general terms at the cost of manually compiling and tagging text. However, this approach suffers from data sparseness when computing the IC of concrete terms (e.g. rare diseases or specific medical terms) or recently minted terms (e.g. names of new drugs), due to the lack of enough data to compute robust probabilities. In general, the accuracy on the probability calculus and, hence, of the IC assessment, closely depends on the size and heterogeneity of the corpus used to retrieve term occurrences [28]. Nowadays, the largest electronic repository is the Web. In fact, the Web is so large and heterogeneous that it is said to be a faithful representation of the information distribution at a social scale [5], an argument that has been supported by recent works focusing on privacy-protection [7,30–32], which considered the Web as a realistic proxy for social knowledge. Moreover, thanks to its dynamicity, the Web covers any possible up-to-date term. Thus, it constitutes an ideal general-purpose source to compute realistic IC values.

In order to compute term probabilities from the Web in an efficient manner, several authors [28,39] have used the *hit count* returned by a Web search engine (e.g. Bing, Google) when querying the term *t*. In our approach, term probabilities are computed in this way.

$$IC_{Web}(t) = -\log_2 \frac{hits(t)}{N} \tag{2}$$

where *N* is the number of web resources indexed by the Web search engine.

By applying the IC calculus to the list of sensitive terms in *S*, we can compute the amount of sensitive information they carry out. Given that our goal is to avoid disclosing any term in *S*, we can reformulate this requirement in terms of Information Theory as *avoiding revealing equal or more information about any term in S than the informativeness of any term in S*. This baseline value can be formulated as a *sanitisation threshold* τ as the IC of the *least informative s* ∊ *S*:

$$\tau = \min_{\forall s \in S}(IC_{Web}(s)) \tag{3}$$

Notice that, by definition, τ corresponds to the IC of the most general term in *S*, which is the one that imposes the stronger privacy constraint; this is because it defines the maximum amount of sensitive information that is allowed to be revealed in the sanitised output without incurring in disclosure. The more general the terms in *S* are (i.e. the lower their IC and, thus, τ), the stricter the sanitisation becomes because a lower degree of information disclosure is allowed.

It is worth to mention that, ideally, this step (as well as *Step-0a*) will be run the first time the system is deployed. After that, recalculations should be only needed when the privacy requirements (i.e. the list of sensitive elements defined by the legislation) or the medical knowledge base (SNOMED-CT) change.

3.3. Step-1: Detect and sanitise sensitive terms

In this step, the system takes as input the medical document to be sanitised and the list of sensitive terms *S*. The goal is twofold: first, all the terms of the document that appear in *S* are detected and marked as sensitive; after that, the system uses the KB to sanitise the formerly detected terms by replacing them with suitable generalisations.

Given that no regular structure is assumed for the document and for its textual content, our method uses several natural language processing tools to detect sensitive terms. Specifically, because sensitive terms are those referring to concepts or instances and these are referred in text by means of nouns or, more generally, noun phrases (NPs) [30], we focus on the detection of NPs. To detect NPs we use several natural language tools (OpenNLP[2]) performing sentence detection, tokenisation (i.e. word detection, including contraction separation), part-of-speech tagging (POS) and syntactic parsing of input text. As a result of this process, POS-tagged words are put together according to their role, by obtaining, among others, the noun phrases (NPs). Once NPs are detected, those that contain any of the terms in S are marked as sensitive. Hereinafter we refer to this set of elements as T. To improve the recall of this matching process with regard to the different morphological derivations of the same word (e.g. singular/plural forms), a stemming algorithm [26] is applied both to the terms in S and to the words of the NP. Notice that linguistic parsing is language-dependant; thus, linguistic tools appropriate for the language in which the input document is written are needed. Currently, natural language processing tools such as OpenNLP support many different languages, including English, German, Spanish and Portuguese.

Afterwards, terms t in T should be protected in a way that they do not disclose sensitive information or, more desirably, that the amount of sensitive information they disclose is low enough. The latter approach is better suited to retain the utility of the protected data, for which the optimal sanitisation is such that protects sensitive information while minimising information loss. To do so, terms t in T are substituted by generalisations $g(t)$ (e.g. $t = HIV \rightarrow g(t) = Virus$), which are extracted from the KB. In this manner, the sanitised document still retains part of its semantics and, hence, a degree of utility.

To select the appropriate generalisation with respect to disclosure risk and data utility, we employ *an information theoretic* approach [30]. More specifically, this step uses the sanitisation threshold τ to guarantee that any generalisation $g(t)$ proposed as replacement of any t discloses less information than τ. As explained in the previous section, this threshold is a numerical value that represents the amount of information provided by the least informative element in S. In this way, the system ensures that any valid generalisation discloses less information than any of the elements to protect. To do so, the hierarchy of generalisations of each sensitive term t, $H = h_1 \rightarrow \cdots \rightarrow h_l$, is obtained from the KB. The optimal generalisation $g(t)$ (from the data utility point of view) will be such that provides the maximum information while fulfilling τ.

$$g(t) = \arg \max_{\forall h_j \in H | IC_{Web}(h_j) < \tau} (IC_{Web}(h_j)) \quad (4)$$

The large and detailed taxonomic structure of SNOMED-CT is especially suited for this purpose, because generalisation steps are fine grained, and this allows retrieving generalisations that accurately fit the sanitisation criterion (i.e. fulfilling the threshold but retaining maximum information).

Fig. 2 illustrates this process. The large white circle represents the IC of a sensitive term $t = HIV$ found in the input document. By looking at the generalisation hierarchy provided by SNOMED-CT for the term *HIV*, we obtain: *HIV* → *Lentivirus* → *Retroviridae* → *Retro-transcribing Virus* → *Virus* → *Microorganism* → *Organism*. Assuming that *Virus* is the first generalisation to fulfil the sanitisation threshold τ, the term *HIV* will be replaced in the sanitised document by *Virus*, whose IC is represented in grey. Notice that, by definition of the taxonomical subsumption [27], the informativeness of a term generalisation constitutes a strict subset of the informativeness of the term, because the former strictly subsumes the semantics of the latter. In this way, the sanitisation process decreases the amount of disclosed information to fulfil the privacy criterion, while retaining a degree of utility that, in this example, specifically corresponds to the area of the grey circle.

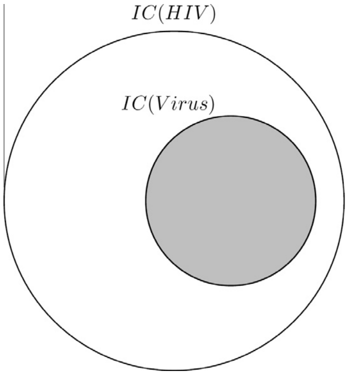

Fig. 2. Graphical representation of the amount of information provided by the term *HIV* and by its generalisation *Virus*.

The output of *Step-1* is a *preliminary sanitised document* that contains the terms marked as sensitive and their suitable generalisations.

3.4. Step-2: Detect and sanitise risky related terms

With the previous step, sensitive terms are analysed and detected independently, which is similar to what is done in many sanitisers [13,23,30,35,40]. For example, when sanitising *Sexually Transmitted Diseases* (*STD*) from a document, the term *Chlamydia* will be marked as sensitive because it is a specialisation of *STD* (according to SNOMED-CT) whereas other terms that are not specialisations like *sexual contact* or *genitals* will remain in clear form. However, these last terms are highly related with *STD* and they may enable its re-identification by means of semantic inference [34]. In fact, most terms appearing in a discourse are semantically related up to a degree [1] and, thus, they may negate sanitisation efforts in which terms are managed independently.

To tackle this problem, we incorporate the final sanitisation step (*Step-2*), in which terms that were not found to be sensitive in the previous step, but that may disclose any of the already detected ones, are also protected. In order to achieve this, we rely on an information theoretic estimation of the risk of disclosing sensitive terms [31] due to the presence, in the same document, of semantically related ones, which can be strictly medical or not. Hereinafter we refer to each of these related terms as q.

According to [31], the disclosure risk (*DR*) that a semantically related term q causes with regard to a sensitive term t can be estimated according to the *amount of information* that q reveals about t. In terms of Information Theory this can be measured as the *Mutual Information* (MI) of two variables. The instantiation of the MI for two specific values (q and t, in this case) corresponds to their *Point-wise Mutual Information* (PMI) [9].

$$DR(t;q) = PMI(t;q) = \log_2 \frac{p(t,q)}{p(t)p(q)} \quad (5)$$

Numerically, if t and q are completely independent (i.e. they co-occur in a textual context just by chance), their PMI is 0 and, thus, there is no disclosure risk for t with regard to q (i.e. $DR(t;q) = 0$). This means that the presence of q in a document does not provide any particular evidence of t. On the other hand, positive PMI values reflect the amount of information of t disclosed by the presence of q. Particularly, if whenever q occurs, t also occurs, then $p(t,q) = p(q)$ and, thus, their PMI is equal to the amount of information provided

[2] http://opennlp.apache.org/ (last accessed: November, 2013).

by t (i.e. $PMI(t;q) = IC(t)$). In this case, there is maximum disclosure risk because the presence of q completely discloses t.

In the same way as in the previous section, PMI probabilities can be computed from the Web hit count provided by a Web search engine:

$$PMI_{Web}(t;q) = \log_2 \frac{\frac{hits(p\ \ AND\ \ q)}{N}}{\frac{hits(t)}{N} \times \frac{hits(q)}{N}} \tag{6}$$

The characterisation of disclosure risk presented in Eq. (5) assumes that *all* the information that an attacker may gather about t is given by q, that is, t is assumed to be removed in the output document. However, since in our approach the terms t are replaced by generalisations $g(t)$, rather than removed, the presence of $g(t)$ in the sanitised output provides *additional* information about t that would produce a higher risk of disclosure.

Fig. 3 illustrates this scenario following to the sanitisation process depicted in Fig. 2, in which the sensitive term $t = HIV$ was replaced by the less informative generalisation $g(t) = Virus$, so that the information that remains in the output about *HIV* is $IC(Virus)$, which is shown in grey. In Fig. 3, the term $q = Infected\ immune\ cells$, which was left in clear form in the first step, is also considered. Given that $t = HIV$ and $q = Infected\ immune\ cells$ are semantically related and, thus, they have a tendency to co-occur, $q = Infected\ immune\ cells$ is revealing an amount of information about $t = HIV$ that specifically corresponds to $PMI(HIV;\ Infected\ immune\ cells)$, which is also represented in grey. Thus, the real disclosure of $t = HIV$ is, in this case, the result of the *union* of the information given by $g(t) = Virus$ and the information that $q = Infected\ immune\ cells$ is revealing about $t = HIV$, that is, $DR = PMI(t;q) + IC(g(t))$, which corresponds to the whole grey area in Fig. 3. Moreover, in the same manner as $q = Infected\ immune\ cells$ is revealing some information about $t = HIV$, it is also likely to reveal information about $g(t) = Virus$, which corresponds to $PMI(Virus;\ Infected\ immune\ cells)$. In order to not to count this amount of information twice, which is represented in the figure as the darker grey area that corresponds to the overlap between $IC(Virus)$ and $PMI(HIV;\ Infected\ immune\ cells)$, we should subtract $PMI(g(t);q)$ from the final expression of *DR*.

Given the above discussion, the final *DR* expression that considers the fact that sensitive terms t are replaced by generalisations $g(t)$ and also the presence of semantically related terms q is the following:

$$DR(t;q) = PMI(t;q) + IC(g(t)) - PMI(g(t);q) \tag{7}$$

In the same manner as in the previous step, a threshold for *DR* values should be defined, both to detect which semantically related terms produce *too much disclosure* and, if that is the case, up to *which level* those should be protected (i.e. sanitised). Given that the threshold τ (see Eq. (3)) quantifies the baseline amount of information that should not be revealed about the entities to protect, any q that, in addition to $g(t)$, reveals equal or more information than τ about any t will be considered as risky.

Like in the previous step, in order to preserve the utility of the output as much as possible, those q found to be risky will be replaced by appropriate generalisations $g(q)$ that provide the maximum information ($IC(g(q))$) while fulfilling the privacy requirements stated by τ, that is $DR(t;g(q)) < \tau$. Moreover, given that the same q may disclose different amounts of information for each sensitive term t in T, its generalisation should be such that the privacy criterion is fulfilled for *all* t in T. To do so, a knowledge base is queried to retrieve the set of generalisations of q ($H = h_1 \rightarrow \cdots \rightarrow h_l$) and the most appropriate one, by considering all t in T, is taken:

$$g(q) = \underset{\forall h_j \in H | DR(t_i;h_j) < \tau}{\arg\max} (IC_{Web}(h_j)) \quad \forall t_i \in T \tag{8}$$

It is worth to mention that, given that semantically related terms q may or may not correspond to medical concepts, general knowledge structures covering domains other than the medical one will be needed to retrieve generalisations. In our case, we use the taxonomies provided by WordNet [14], an structured thesauri covering more than 100,000 concepts, and ODP (Open Directory Service), whose taxonomy structures web resources in more than 1,000,000 categories, in addition to SNOMED-CT. In case of overlap (i.e. a term which is contained in several knowledge bases), the strict order is (i) SNOMED-CT, (ii) WordNet and (iii) ODP, because the former provide a more detailed structuring of medical concepts. Only in such cases in which q is not found in any knowledge base, it will be removed from the output.

Following the graphical example used above, Fig. 4 shows the reduction of disclosure risk resulting from the sanitisation of the semantically related term $q = Infected\ immune\ cells$ with regard to the sensitive one $t = HIV$. In this case, assume that the generalisation retrieved from SNOMED-CT of $q = Infected\ immune\ cells$ that fulfils the privacy criterion (Eq. (8)) is $g(q) = Cells$. Since $IC(Cells)$ is a strict subset of $IC(Infected\ immune\ cells)$, its information disclosure about $t = HIV$ (i.e. the grey area corresponding to $PMI(HIV;\ Cells)$) is also smaller than for the original $q = Infected\ immune\ cells$ (see the grey area that corresponds to $PMI(HIV;\ Infected\ immune\ cells)$ in Fig. 3). As a result, the total disclosure (union of greyed areas) is also smaller than in Fig. 3, even though an amount of information/utility about q is still preserved in the output, which strictly corresponds to $IC(Cells)$.

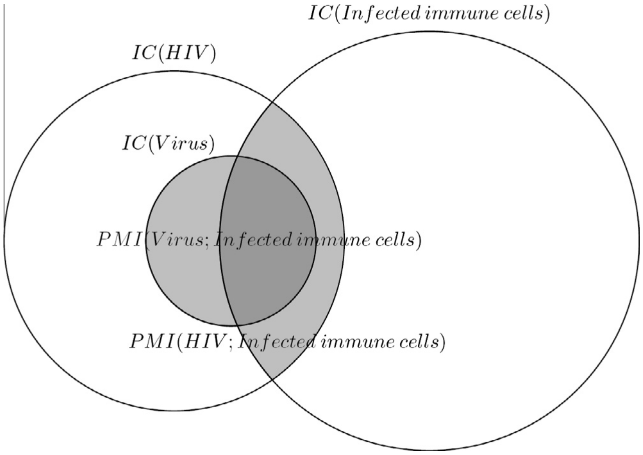
IC(Infected immune cells)
IC(HIV)
IC(Virus)
PMI(Virus; Infected immune cells)
PMI(HIV; Infected immune cells)

Fig. 3. Graphical representation of the disclosure risk caused by the presence of the semantically related term *Infected immune cells* with respect to the sensitive term *HIV* and its generalisation *Virus*.

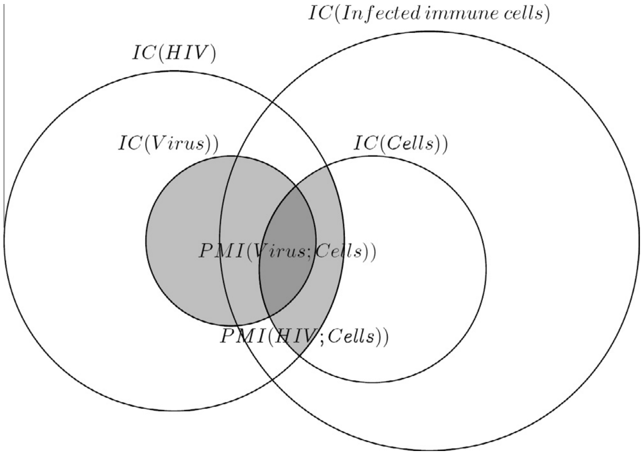

Fig. 4. Graphical representation of the disclosure risk reduction (with respect to *HIV*) resulting from the replacement of *Infected immune cells* by its generalisation *Cells*.

Please cite this article in press as: Sánchez D et al. Utility-preserving privacy protection of textual healthcare documents. J Biomed Inform (2014), http://dx.doi.org/10.1016/j.jbi.2014.06.008

Notice also that the disclosure of a certain t (e.g. *HIV*) may also happen because of the combination of *several* terms $\{q_1,\ldots,q_p\}$ (e.g. *sexual transmission*, *immune system*, etc.), even if each q_i did not caused disclosure individually. In such cases, the above method can be extended to evaluate the disclosure risk of t with regard to a set $\{q_1,\ldots,q_p\}$ and to generalise the elements of the set until the threshold τ is fulfilled. To do so, the PMI of the term t with regard to a set $\{q_1,\ldots,q_p\}$ can be computed as follows:

$$PMI(t;\{q_1,\ldots,q_p\}) == \log_2 \frac{p(t,q_1,\ldots,q_p)}{p(t)p(q_1,\ldots,q_p)} \tag{9}$$

As a result of this process, the *final sanitised document* is obtained in which both sensitive terms and those semantically related ones that were found to be risky are sanitised.

3.5. Step-2: Tackling language ambiguity

In Section 3.2 we argued about the suitability of using the Web as general corpora for the probability calculus needed for IC/PMI assessments, and of using Web Search Engines (WSEs) as proxies to obtain these probabilities. However, probabilities of sensitive entities computed from the number of explicit occurrences of terms referring to them, that is, the *hit count* provided by the WSE when querying the terms, may be negatively influenced by *language ambiguity*. Specifically, when the words used to refer an entity are polysemic (e.g. *crabs* can refer to a *sexually transmitted disease* or to a *crustacean*), the probability computed from the web hit count may overestimate the probability of the underling concept (e.g. *crabs* as a *STD*), because the web hit count includes *all* the appearances of the word, regardless of their meanings; thus, the entity will be considered as *less* informative than what it truly is. Likewise, if an entity can be referred by means of different synonyms (e.g. the terms *HIV* and *human immunodeficiency virus* refer to the same disease) or entities referred in a discourse are not explicitly mentioned in text due to ellipsis, the resulting probabilities will be lower than expected and, thus, the entity informativeness will be overestimated.

On the one hand, probability inaccuracies caused by synonymy can be minimised by obtaining the synonyms of each entity to be protected from SNOMED-CT, as it was stated in Section 3.1. On the other hand, to tackle the problems caused by polysemy and to minimise the effect of ellipsis, a suitable solution consists on contextualising WSE term queries within the scope of one of the term's generalisations (e.g. *crabs* AND *STD*), as proposed in [28]. The main idea is that, by forcing the co-occurrence of a term referring to an entity (e.g. *crabs*, as a type of *STD*) and the generalisation that is adequate to the meaning of the entity (e.g. *STD*), the effect of polysemy in the resulting hit count is minimised. This is because words rarely refer to different senses within the same document: if *crabs* and *sexually transmitted disease* appears together it is very likely that the *crabs* appearance does not refer to a *crustacean*. Likewise, since only explicit co-occurrences of terms and their generalisations are considered, the potential ellipses of the latter are omitted from the probability assessment. This contributes to improve the coherence of the IC assessment, that is, the fact that the informativeness of a generalisation should be a strict subset of the informativeness of its specialisation [27], as stated in Section 3.3. Thus, query contextualisation enables a more accurate probability assessment [28]. The downside is the fact that the explicit contextualisation of term occurrences constraints the size of the corpora considered in the probability calculus. However, the size and redundancy of the Web helps to minimise the effect of this handicap, which, in any case, is preferable to the negative influence of language ambiguity [28].

To contextualise the hit count resulting from term queries performed to the WSE, the generalisation will be attached to the term using a logic operator supported by the WSE, such as AND or +. We have applied this procedure to the queries evaluating the disclosure risk of t in T during *Step-2*. In this manner, the detection and sanitisation of related terms q will be less affected by the ambiguity associated to the evaluated t. Thus, whenever a sensitive entity is queried to the WSE to compute IC and PMI, its generalisation will be attached. Since this generalisation is retrieved from SNOMED-CT, we can be sure that it will correspond to the appropriate meaning of the entity, thus enabling the desired disambiguation (e.g. the generalisation of *crabs* retrieved from SNOMED-CT will refer to *a type of STD*). As stated above, this contextualisation is considered when querying t in the expression of DR (Eq. (7)):

$$\begin{aligned} DR(t;q) &= PMI_{Web}(t\ AND\ g(t);q) + IC_{Web}(g(t)) \\ &\quad - PMI_{Web}(g(t);q) \end{aligned} \tag{10}$$

where $PMI_{Web}(t\ AND\ g(t);q)$ is computed as follows:

$$PMI_{Web}(t\ AND\ g(t);q) = \log_2 \frac{\frac{hits(t\ AND\ g(t)\ AND\ q)}{N}}{\frac{hits(t\ AND\ g(t))}{N} \times \frac{hits(q)}{N}} \tag{11}$$

The contextualisation is also considered in the expression computing the threshold τ from the sensitive entities s in S (Eq. (3)):

$$\tau = \min_{\forall s \in S}(IC_{Web}(s\ AND\ g(s))) \tag{12}$$

where $IC_{Web}(s\ AND\ g(s))$ is computed as follows:

$$IC_{Web}(s\ AND\ g(s)) = -\log_2 \frac{hits(s\ AND\ g(s))}{N} \tag{13}$$

In this last equation $g(s)$ corresponds to the most specific generalisation of s that fulfils $IC(g(s)) < IC(s)$.

4. Evaluation

To simulate and evaluate the protection that our system would be able to achieve for textual medical documents (e.g. electronic healthcare records), we used the Wikipedia descriptions of the set of medical entities that are considered as sensitive by U.S. state and federal laws [10,20]. As stated in the introduction, these legislations mandate hospitals and healthcare organisations to redact some medical-related concepts that are considered of confidential nature before releasing patient records to, for example, insurance companies, in response to Worker's Compensation or Motor Vehicle Accident claims, or a judge, in case of malpractice litigation [4]. Usually, all references to potentially discriminating conditions like alcohol and substance abuse, Sexually Transmitted Diseases (STD), mental diseases or HIV/AIDS status should be redacted/sanitised. All these terms were feed as the input to our method in order to obtain the list of sensitive terms S from SNOMED-CT (*Step-0a*) and to compute the sanitisation threshold (*Step-0b*). A total of 6 Wikipedia articles each one describing the main entities from a medical perspective were taken: *STD*, *HIV*, *AIDS*, *Mental disorder*, *Alcohol abuse* and *Substance abuse*. As done in other works [7,30,34], Wikipedia descriptions were chosen as representatives of the kind of textual information that could be found in healthcare records due to being freely accessible and authoritative sources of information, and also because of their high informativeness and tight discourses, which configure a challenging scenario from the perspective of document redacting/sanitisation. SNOMED-CT, WordNet and ODP were used to retrieve term generalisations and Bing[3] was employed as the Web search engine to obtain term hit counts for probability calculus.

[3] http://www.bing.com/.

Please cite this article in press as: Sánchez D et al. Utility-preserving privacy protection of textual healthcare documents. J Biomed Inform (2014), http://dx.doi.org/10.1016/j.jbi.2014.06.008

The proposed method has been implemented in Java and run over an Intel Core2 Quad 2.66 GHz with 4 GB RAM, Windows 7 and a 100 Mb Internet connection. The average runtime for the sanitisation of the six Wikipedia sources was 16.7 min. Note that most of the runtime is devoted to perform queries to Web search engines and to wait for the results, but this waiting periods are highly parallelisable.

4.1. Detection accuracy

First, we evaluated the detection accuracy of sensitive and related terms. To do so, two human experts were requested to manually sanitise each Wikipedia article with the aim of detecting terms that, under their opinion, may help to disclose any of the entities stated as sensitive (i.e. *STD*, *HIV*, *AIDS*, *Mental disorder*, *Alcohol abuse* and *Substance abuse*). The initial inter-agreement was 0.88; however, they were requested to agree on the differences to obtain consensual results to which evaluate our method. By comparing the outputs of the automatic and manual detection procedures, *precision*, *recall* and *F-measure* can be measured.*Precision* quantifies the proportion of sensitive terms identified by our *method* that have also been identified by the *human experts*. The lower the precision is and, thus, the better the utility of the protected output will be because a lower amount of terms will be unnecessarily redacted or sanitised.

$$Precision = \frac{|Method \cap Human|}{|Method|} \times 100 \tag{14}$$

Recall quantifies the proportion of sensitive terms identified by the *human experts* that our *method* has been also able to identify. Thus, the higher the recall is, the higher the privacy of the output will be. In document redaction/sanitisation, recall usually plays a more important role than precision because a low recall implies disclosing data that may negate the whole sanitisation [1].

$$Recall = \frac{|Method \cap Human|}{|Human|} \times 100 \tag{15}$$

Finally *F-measure* provides the harmonic mean of precision and recall and, thus, summarises the accuracy of the detection process.

$$F\text{-}measure = \frac{2 \times Recall \times Precision}{Recall + Precision} \tag{16}$$

Table 1
Evaluation of sensitive terms detected only by the first step of the proposed method.

Wikipedia article	Precision (%)	Recall (%)	*F*-measure (%)
STD	100	9.52	17.38
HIV	100	12	21.43
AIDS	100	5.88	11.11
Mental disorder	100	17.39	29.63
Alcohol abuse	100	28.57	44.44
Substance abuse	100	6.97	13.03
Average	100	13.38	22.84

In the first experiment, as done in many methods [13,29,30,35,40], we did not considered semantically related terms, that is, we solely applied the first step of the proposed method. Evaluation results are shown in Table 1.

Since the detection process of *Step-1* just looks for sensitive terms and for their taxonomic specialisations (retrieved from SNOMED-CT), a large number of risky terms that are semantically related (other than taxonomically) with the entities to protect are left in clear. Indeed, since Wikipedia articles usually present a tight focus on the described entity, most of the terms they contain are semantically related to the entity to protect [31,32]. These results illustrate the importance of considering semantic relationships in document redaction/sanitisation and how an independent evaluation of terms is usually not enough to achieve enough protection in front of semantic inferences [1]. On the other hand, given that terms detected as sensitive are extracted from SNOMED-CT unambiguously, precision is perfect in all cases.

In the second test, the whole method was applied, which detects both sensitive terms and those that are semantically related. Given that the detection criteria of the second step depends on how probabilities are computed, we performed two runs for each article: one using the basic probability calculus and another one with the contextualised version proposed in Section 3.5 to minimise language ambiguity. Evaluation results are depicted in Table 2.

Recall figures in this second experiment are noticeably higher than those of the first due to the evaluation of semantically related terms (93–96% vs. 13.4%, in average). This is especially relevant given that recall figures directly measure the degree of privacy achieved in the output. It is interesting that, in around half of the cases, the recall reached a 100%. This suggests that the outputs are perfectly valid in a real setting.

Precision, on the other hand, is lower than in the previous experiment (54–75% vs. 100%, in average). This is caused by the larger number of false positives that result from the imperfections of the automatic assessment and, especially, of the probability calculus. In fact, we observe noticeable differences between contextualised on non-contextualised versions of the probability calculus. In the former case, precision is noticeably higher (75.3% vs. 54.5%, in average), which suggests that:

- The non-contextualised term probabilities resulted in a too strict sanitisation, that is, too many terms were unnecessarily sanitised, which is reflected in the lower precision. This was either because the baseline informativeness used as threshold was underestimated, or because the amount of information disclosure caused by related terms regarding the sensitive ones were overestimated.
- The procedure applied to minimise language ambiguity in the probability calculus (that is, polysemy and ellipsis) had a positive contribution in enabling a more precise detection, even when causing a slightly decrease in recall (93.45% vs. 96.31%).
- As a result, the global accuracy (i.e. *F*-measure) of the detection process with contextualised probabilities significantly surpasses that of its non-contextualised version (83% vs. 69.3%, in

Table 2
Evaluation of sensitive terms detected by the whole method with contextualised and non-contextualised versions of the probability calculus.

Wikipedia article	Precision		Recall		*F*-measure	
	No context (%)	Context (%)	No context (%)	Context (%)	No context (%)	Context (%)
STD	50.06	66.66	100	97.56	66.72	79.20
HIV	75.6	85.71	100	100	86.105	92.30
AIDS	50.08	75.6	91.17	91.17	64.65	82.66
Mental disorder	46.93	82.14	92	85.18	62.15	83.63
Alcohol abuse	51.61	66.66	100	100	68.08	79.99
Substance abuse	52.94	75	94.73	86.84	67.92	80.49
Average	54.54	75.29	96.31	93.45	69.27	83.05

Please cite this article in press as: Sánchez D et al. Utility-preserving privacy protection of textual healthcare documents. J Biomed Inform (2014), http://dx.doi.org/10.1016/j.jbi.2014.06.008

Table 3
Utility preservation according to the different redaction/sanitisation strategies.

Wikipedia article	Step-1: removal (%)	Step-1: generalisation (%)	Steps-1 & 2: removal		Steps-1 & 2: generalisation	
			No context (%)	Context (%)	No context (%)	Context (%)
STD	93.75	96.76	30.11	39.34	35.88	74.69
HIV	90.93	97.97	10.49	14.13	24.24	71.89
AIDS	93.15	98.87	21.30	39.27	29.88	78.21
Mental disorder	91.61	96.33	32.44	50.54	46.19	81.11
Alcohol abuse	83.56	93.55	8.04	20.10	14.58	63.65
Substance abuse	96.21	98.73	25.19	43.41	38.41	75.27
Average	91.53	97.o03	21.26	34.46	31.53	74.13

average). This suggests that contextualised probabilities better capture the informativeness and information disclosure of sensitive entities and related terms, respectively.

4.2. Utility evaluation

The second part of the evaluation quantifies the amount of utility that the protected output preserves with regard to the input document, and compares our method with different redaction/sanitisation strategies.

To measure the degree of utility preservation we first quantify the utility of a document as the sum of the *amount of information* provided by *all* the terms *w* appearing in a document *D* (i.e., noun phrases detected by means of the syntactic parsing, regardless of being latterly identified as sensitive or not).

$$IC(D) = \sum_{\forall w \in D} IC(w) \tag{17}$$

Then, the degree of utility preservation of the sanitised document D' is measured as the ratio between the utility of D' and of the original version D [30], as follows:

$$Utility(D') = \frac{IC(D')}{IC(D)} \times 100 \tag{18}$$

Utility figures are shown in Table 3. In order to put them into context, the following redaction/sanitisation strategies have been implemented:

- Only those terms detected as sensitive in the *Step-1* of the method are protected. This is done in two ways: (i) term *removal* (i.e. redacting), as in [13,35,40], and (ii) term *generalisation* (i.e. sanitisation) according to the privacy criteria, as in [30].
- All terms detected by the whole method (*Step-1* and *2*) are protected either by: (i) term *removal*, as done in [31], and (ii) term *generalisation*, as proposed by our method. Given that *Step-2* depends on the probability calculus used to detect and generalise terms, we include the results with the basic probability assessment and with the contextualised version proposed in Section 3.5.

Utility figures obtained when only the first step of the method is applied are coherent with the low recall figures reported in Table 1. Given that the individual analysis of terms omits a large number of semantically related ones (which are left in clear form), utility preservation is very high (above 90%), at the cost of a very low privacy protection (recall below 14% in average, as shown in Table 1). Obviously, the generalisation of sensitive terms preserves a larger amount of information than the straightforward term removal (97% vs. 91.5%, in average). This illustrates the advantages of the exploitation of a knowledge base to improve the utility of the protected output.

Utility preservation is significantly lower when both sensitive and semantically related terms are protected (*Step-1* and *2*). This suggests that a large percentage of terms appearing in the input document were indeed semantically related to the entities to be protected, which is coherent with the tight discourses that characterise Wikipedia articles. These results are also coherent with the near-perfect recall reported in Table 2, which suggests that most of the risky terms where properly identified and protected. Thus, the observed differences in utility preservation quantify the cost derived from the protection of risky related terms, that is, the cost of the more robust privacy guarantees.

There are, however, noticeable differences between the specific protection method and the probability calculus. First, the strategy based on term generalisation preserves around 100% more utility (34.5% vs. 21.3% for non-contextualised probabilities and 74.1% vs. 31.5% for contextualised probabilities, in average) than the one performing term removal. In fact, figures obtained by pure redaction (i.e. removal) suggest that the protected outputs would be hardly usable for human readers and also for data analysis. Moreover, the contextualisation of the probability calculus adds a comparable degree of utility improvement over the baseline protection strategy (31.5% vs. 21.7% for pure redaction and 74.1% vs. 34.5% for term sanitisation, in average). In this latter case, the observed improvement is motivated by:

- The number of false positives that are unnecessarily redacted/sanitised in *Step-2* is significantly lower when using contextualised probabilities, as suggested by the precision figures reported in Table 2. Thus, the utility of the protected output will be noticeably higher thanks to the more accurate detection process.
- As argued in Section 3.5, term ambiguity (i.e. polysemy and ellipsis) is minimised when using contextualised probabilities. Given that this ambiguity tends to underestimate baseline IC values and, thus, to force the method to replace terms by more abstract generalisations to fulfil the privacy criterion (i.e. those with lower IC), unambiguous probabilities contribute to retrieve more accurate and, thus, more utility-preserving generalisations.

At the end, utility figures obtained by our complete method (*Step-1* and *2*, with contextualised probabilities and term generalisations) are just a 17–23% lower than baseline approaches that do not consider semantic relationships (74% vs. 91.5–97%, in average), while providing much more robust privacy guarantees (93.5% vs. 13.4% of average term recall, as reported in Tables 1 and 2).

5. Conclusions

This paper presents an automatic method to sanitise textual medical data. Several aspects differentiate it from related works. First, several knowledge bases are exploited to retain, as much as possible, the semantics and, thus, the analytical utility and readability of the protected output. Second, on the contrary to methods focusing on specific types of sensitive entities (like e-mail addresses or social security numbers) [2,3,13,15,17–19,25,35,36,40,41], our approach does not make any assumptions on the structure of the terms to protect. As a result, it in can be applied to textual contents regardless the fact that terms (e.g. diseases, symptoms, treatment, etc.) present or not a regular structure.

Please cite this article in press as: Sánchez D et al. Utility-preserving privacy protection of textual healthcare documents. J Biomed Inform (2014), http://dx.doi.org/10.1016/j.jbi.2014.06.008

Finally, it carefully considers the disclosure risk caused by the presence of semantically related terms, which is especially critical in the medical context in which most terms appearing in a document are likely to be semantically related up to some degree.

To achieve those goals, the proposed method builds on an information theoretic characterisation of term sensitiveness and disclosure risk [30–32] and an accurate calculus of term probabilities from the Web. The fact that the privacy criterion can be defined by simply listing the set of entities that should be considered as sensitive makes our approach intuitive (from the perspective of the privacy guarantees that one may expect from the protected output), and especially suitable to be applied in coherency with legislations on medical data privacy, which are specified in the same manner.

The evaluation showed that: (i) the analysis of semantically related terms, (ii) the contextualisation of probability queries and (iii) the replacement (instead of removal) of sensitive terms by appropriate generalisations improved the detection recall of sensitive information (i.e. the practical privacy of the output) while contributing to preserve the output's utility.

As future work, we plan to improve and extend the linguistic analysis of texts by incorporating other terms that may cause disclosure (e.g. verbs in sentences like "he drinks too much"). Ontologies modelling verbs such as WordNet could be used to assist the sanitisation process. A deeper linguistic analysis may also contribute to improve the accuracy by detecting negated assertions (e.g. "AIDS negative") and thus avoiding unnecessary sanitisations. Further tests will be also performed with sources written in different languages in order to illustrate the applicability of our method given the availability of linguistic parsing tools for such languages. Finally, experiments with real medical data, which is the focus of our system, are also planned.

Acknowledgments

Authors are solely responsible for the views expressed in this paper, which do not necessarily reflect the position of UNESCO nor commit that organisation. 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). This work was also made possible through the support of a grant from Templeton World Charity Foundation (USA). The opinions expressed in this paper are those of the author(s) and do not necessarily reflect the views of Templeton World Charity Foundation.

References

[1] Anandan B, Clifton C. Significance of term relationships on anonymization. In: IEEE/WIC/ACM international joint conference on web intelligence and intelligent agent technology – workshops. Lyon, France; 2011. p. 253–6.

[2] Aramaki E, Imai T, Miyo K, Ohe K. Automatic deidentification by using sentence features and label consistency. In: i2b2 Workshop on challenges in natural language processing for clinical data; 2006.

[3] Beckwith BA, Mahaadevan R, Balis UJ, Kuo F. Development and evaluation of an open source software tool for deidentification of pathology reports. BMC Med Inform Decis Making 2006;6.

[4] Bier E, Chow R, Golle P, King TH, Staddon J. The rules of redaction: identify, protect, review (and repeat). IEEE Secur Privacy Mag 2009;7:46–53.

[5] Cilibrasi RL, Vitányi PMB. The Google similarity distance. IEEE Trans Knowl Data Eng 2006;19:370–83.

[6] Cumby C, Ghani R. A machine learning based system for semiautomatically redacting documents. In: Twenty-Third conference on innovative applications of artificial intelligence. San Francisco, California, USA; 2011. p. 1628–35.

[7] Chow R, Golle P, Staddon J. Detecting privacy leaks using corpus-based association rules. In: 14th Conference on knowledge discovery and data mining. Las Vegas, NV: ACM; 2008. p. 893–901.

[8] Chow R, Oberst I, Staddon J. 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] Church KW, Hanks P. Word association norms, mutual information, and lexicography. Comput Linguist 1990;16:22–9.

[10] Department for a Healthy New York. New York State Confidentiality Law; 2013.

[11] Department of Health and Human Services. The health insurance portability and accountability act of 1996; 2000.

[12] Domingo-Ferrer J, Sánchez D, Rufian-Torrell G. Anonymization of nominal data based on semantic marginality. Inform Sci 2013;242:35–48.

[13] Douglass M, Cliffford G, Reisner A, Long W, Moody G, Mark R. De-identification algorithm for free-text nursing notes. Comput Cardiol 2005:331–4.

[14] Fellbaum C. WordNet: an electronic lexical database. Cambridge, Massachusetts: MIT Press; 1998.

[15] Friedlin FJ, McDonald CJ. A software tool for removing patient identifying information from clinical documents. J Am Med Inform Assoc 2008;15:601–10.

[16] Gordon K. Medical record redacting: a burdensome and problematic method for protecting patient privacy. MRA: Health Information Services; 2013.

[17] Guo Y, Gaizauskas R, Roberts I, Demetriou G, Hepple M. Identifying personal health information using support vector machines. In: i2b2 Workshop on challenges in natural language processing for clinical data; 2006.

[18] Gupta D, Saul M, Gilbertson J. Evaluation of a deidentification (De-Id) software engine to share pathology reports and clinical documents for research. Am J Clin Pathol 2004;121:176–86.

[19] Hara K. Applying a SVM based Chunker and a text classifier to the deid challenge. In: i2b2 Workshop on challenges in natural language processing for clinical data; 2006.

[20] Health Privacy Project. State Privacy Protections; 2013.

[21] Lopresti D, Spitz A. Information leakage through document redaction: attacks and countermeasures. Document recognition and retrieval XII; 2005.

[22] Martínez S, Sánchez D, Valls A. A semantic framework to protect the privacy of electronic health records with non-numerical attributes. J Biomed Inform 2013;46:294–303.

[23] Meystre SM, Friedlin FJ, South BR, Shen S, Samore MH. Automatic de-identification of textual documents in the electronic health record: a review of recent research. BMC Med Res Methodol 2010;10.

[24] National Security Agency. Redacting with confidence. How to safely publish sanitized reports converted from word to pdf; 2005.

[25] Neamatullah I, Douglass MM, Lehman LH, Reisner A, Villarroel M, Long WJ, et al. Automated de-identification of free-text medical records. BMC Med Inform Decis Making 2008;8.

[26] Porter MF. An algorithm for suffix stripping. Read Inform Ret 1997:313–6.

[27] Resnik P. Using information content to evaluate semantic similarity in a taxonomy. In: Mellish CS, editor. 14th International joint conference on artificial intelligence, IJCAI 1995. Montreal, Quebec, Canada: Morgan Kaufmann Publishers Inc.; 1995. p. 448–53.

[28] Sánchez D, Batet M, Valls A, Gibert K. Ontology-driven web-based semantic similarity. J Intell Inform Syst 2010;35:383–413.

[29] Sánchez D, Batet M, Viejo A. Detecting sensitive information from textual documents: an information-theoretic approach. In: 9th International conference modeling decisions for artificial intelligence, MDAI 2012. Springer; 2012. p. 173–84.

[30] Sánchez D, Batet M, Viejo A. Automatic general-purpose sanitization of textual documents. IEEE Trans Inform Forensics Secur 2013;8:853–62.

[31] Sánchez D, Batet M, Viejo A. Minimizing the disclosure risk of semantic correlations in document sanitization. Inform Sci 2013;249:110–23.

[32] Sánchez D, Batet M, Viejo A. Utility-preserving sanitization of semantically correlated terms in textual documents. Inform Sci 2014;279:77–93.

[33] Spackman K. SNOMED CT milestones: endorsements are added to already-impressive standards credentials. Healthcare Inform 2004;21:54–6.

[34] Staddon J, Golle P, Zimmy B. Web-based inference detection. In: 16th USENIX security symposium; 2007. p. Article No. 6.

[35] Sweeney L. Replacing personally-identifying information in medical records, the scrub system. In: 1996 American medical informatics association annual fall symposium. Washington, DC, USA; 1996. pp. 333-7.

[36] Szarvas G, Farkas R, Busa-Fekete R. State-of-the-art anonymization of medical records using an iterative machine learning framework. J Am Med Inform Assoc 2007;14:574–80.

[37] Terry N, Francis L. Ensuring the privacy and confidentiality of electronic health records. Univ Illinois Law Rev 2007:681–735.

[38] The European Parliament and the Council of the EU. Data Protection Directive 95/46/EC; 1995.

[39] Turney PD. Mining the web for synonyms: PMI-IR versus LSA on TOEFL. In: De Raedt L, Flach P, editors. 12th European conference on machine learning, ECML 2001. Freiburg, Germany: Springer-Verlag; 2001. p. 491–502.

[40] Tveit A, Edsberg O, Rost TB, Faxvaag A, Nytro O, Nordgard T, Ranang MT, Grimsmo A. Anonymization of general practioner medical records. In: Second HelsIT conference; 2004.

[41] Wellner B, Huyck M, Mardis S, Aberdeen J, Morgan A, Peshkin L, et al. Rapidly retargetable approaches to de-identification in medical records. J Am Med Inform Assoc 2007;14:564–73.