Automatic General-Purpose Sanitization of Textual Documents

David Sánchez, Montserrat Batet, and Alexandre Viejo

***Abstract*—The advent of new information sharing technologies has led society to a scenario where thousands of textual documents are publicly published every day. The existence of confidential information in many of these documents motivates the use of measures to hide sensitive data before being published, which is precisely the goal of *document sanitization*. Even though methods to assist the sanitization process have been proposed, most of them are focused on the detection of specific types of sensitive entities for concrete domains, lacking generality and and requiring user supervision. Moreover, to hide sensitive terms, most approaches opt to remove them, a measure that hampers the utility of the sanitized document. This paper presents a general-purpose sanitization method that, based on information theory and exploiting knowledge bases, detects and hides sensitive textual information while preserving its meaning. Our proposal works in an automatic and unsupervised way and it can be applied to heterogeneous documents, which make it specially suitable for environments with massive and heterogeneous information-sharing needs. Evaluation results show that our method outperforms strategies based on trained classifiers regarding the detection recall, whereas it better retains the document's utility compared to term-suppression methods.**

***Index Terms*—Data publishing, document sanitization, information theory, privacy.**

I. Introduction

In THE context of the Information Society, thousands of documents potentially containing sensitive information are made public or available for third parties daily for a variety of reasons. Governments that publish documents in response to Freedom of Information requests [1] or medical data like electronic health care records, which are made available due to their usefulness for clinical research [2], are examples of this situation. Moreover, in recent years, the emergence of the Cloud has represented a fundamental change in the way information technology services are designed and deployed in business and governments. In fact, the use of cloud environments has predominantly focused on information sharing and communications [3]. More specifically, the use of document-sharing applications is one of the main opportunities for the cloud computing industry [4]. However, this environment represents a serious threat for data privacy, since information related to companies, clients or sales operations might be made available for potentially untrusted parties [5], [6].

Due to the confidential nature of many of the published/shared documents, measures should be taken to remove or hide sensitive information that may disclose identities of referred entities (e.g., names, social security numbers, addresses, etc.) or reveal their confidential data (e.g., medical diagnosis, outcomes, etc.). Official regulations have been developed at this respect. A recent EU bill will force any company which compromises the privacy of its clients by moving their confidential data to the Cloud to pay a fine which may amount to a substantial part of its revenue [7]. In the medical context, the *Health Insurance Portability and Accountability Act (HIPAA)* [8] states safe harbor rules about the kind of personally identifiable information which should be removed in medical documents prior allowing their publication.

Data sanitization pursuits the removal of sensitive information so that it may be distributed to a broader audience. In such process, there is a trade off between removing/hiding enough information to protect sensitive data and not to over-sanitize them to the point in which their utility for third parties (which is the main motivation of data publication) is eliminated [9].

In recent years, a lot of efforts have been put in privacy protection of disclosed/published data, even though most of them focus on structured data like relational databases [10]–[12]. In such scenarios, authors exploit the structure of data to anonymize attributes that are known to be potential identifiers (e.g., ID cards, names, addresses), making them nondistinguishable from other records in the same dataset. Much less attention has been paid to the development of methods for sanitizing unstructured data, like textual transactions (e.g., query logs [13]) or raw text documents [9], [14], which is the usual way in which data is transferred between parties. Sanitization of text documents has been traditionally done manually, making it expensive, time-consuming [15] and prone to disclosure risks [9]. Moreover, manual sanitization does not scale as the volume of data increases [16]. Considering the amount of digital textual information made available daily (e.g., the US Department of Energy's OpenNet initiative [17] requires of sanitizing millions of documents yearly), and the adaption of massive information-sharing technologies like the Cloud, one can realize of the need of automatic text sanitization methods. This need is manifested in initiatives from DARPA [18] or the Consortium for Healthcare Informatics Research

Manuscript received May 23, 2012; revised November 20, 2012; accepted January 07, 2013. Date of publication January 11, 2013; date of current version May 16, 2013. This work was supported in part by the European Commission under FP7 project Inter-Trust, in part by the Spanish Ministry of Science and Innovation (through projects eAEGIS TSI2007-65406-C03-01, CO-PRIVACY TIN2011-27076-C03-01, ARES-CONSOLIDER INGENIO 2010 CSD2007-00004, Audit Transparency Voting Process IPT-430000-2010-31, BallotNext IPT-2012-0603-430000 and ICWT TIN2012-32757), and in part by the Government of Catalonia (under Grant 2009 SGR 1135). The associate editor coordinating the review of this manuscript and approving it for publication was Jayant Haritsa.

The authors are with the Departament d'Enginyeria Informàtica i Matemàtiques, Universitat Rovira i Virgili, UNESCO Chair in Data Privacy, E-43007 Tarragona, Spain (e-mail: david.sanchez@urv.cat).

Digital Object Identifier 10.1109/TIFS.2013.2239641

(CHIR)[14] which aim at building new methods and tools for declassification of confidential documents.

Contrary to the anonymization of relational databases, raw text does not necessarily contain explicitly identified sensitive attributes [9]. Hence, document sanitization consists of two tasks: (i) detection of identifiable information within text; and (ii) information hiding, in a way that the disclosure risk is minimized and, ideally, the utility of the sanitized text is maximized. As it will be shown in the next section, the first task has usually received more attention than the later [19], which is commonly tackled by simply removing sensitive information. Note that this measure severely hampers the document's utility (which goes against the purpose of data publication), resulting in sanitized texts that are hardly readable and, even worse, not containing enough information to be used in research tasks.

A. Related Work

Manual document sanitization has a been widely used by governments and companies. In the literature, there are standard guidelines [20] detailing the correct procedures to ensure irreversible suppression or distortion of sensitive parts. Commercial applications like Adobe Acrobat Professional offers a semiautomatic sanitization [21] that recognizes certain sensitive entities using patterns (e-mail addresses, dates, etc.). Upon detection, users are asked to suppress them or leave them unmodified. Both manual and semiautomatic approaches require user interaction and, hence, their scalability is severely limited given the current sanitization needs.

Focusing on unsupervised sanitization methods, one of the first approaches is the Scrub system [22], which also relies on detection patterns for specific data types. Detected data is replaced with another term of similar type. Similar schemes that focus on removing sensitive terms from medical records [2], [23] use specific patterns designed according to the HIPAA "Safe Harbor" rules that mentions 18 data elements (e.g., names, dates, medical record numbers) that must be removed from clinical data [8]. A main drawback of these proposals is the fact that data semantics are omitted during the masking of sensitive data, hampering document's utility. Besides, the use of specific patterns strongly limits their applicability.

The authors in [16] present a scheme that detects and removes sensitive elements using a database of entities (persons, products, diseases, etc.) instead of patterns. Each entity is associated with a set of related terms, which represents the entity's context (e.g., the context of a person includes her name, birth date, etc.). The use of ad-hoc knowledge bases related to specific areas of knowledge also limits the applicability of the scheme when sanitization needs become more heterogeneous. This proposal also offers the possibility to set the desired balance between privacy and utility levels, which is formalized with the "K-safety" concept. K-safety requires that the maximum subset of each sensitive entity's context contained in a document is also contained by the contexts of at least K other entities. Authors note, however, that obtaining a K-safety sanitized document is NP-hard and requires from relatively homogeneous sets of documents (i.e., documents cannot be sanitized individually).

Even though [24] does not present any concrete system, this work represents two main advances in comparison to the above schemes. First, it focuses on general unstructured documents. To enable the a general-purpose sanitization, authors propose the use of named entity recognition techniques to identify the sensitive terms. It is worth to mention that this proposal assumes that named entities are always sensitive. The second advance regards the entity protection approach. Instead of entity removal, it discusses, from a *theoretical point of view*, three alternative approaches to hide protected entities: (i), *Entity generalization:* entities can be generalized to achieve some degree of privacy while preserving some of their semantics; (ii), *Entity swapping:* entities of different documents of the same set or within the same document can be swapped depending on the concrete case; and (iii), *Entity noise addition:* an entity can be substituted by another similar one extracted from another repository. Note that the last two methods do not guarantee the semantic coherence of the resulting sanitization.

The authors in [9] present a semiautomatic sanitization tool for Microsoft Word based on trained classifiers. An interesting characteristic of this method is the use of a general knowledge base like WordNet [25] to generalize the protected entities to preserve their semantics. Regarding the detection process, this work focuses on documents directly linked to certain companies. The data to be detected is divided into two main categories: (i) person names, location names, phone numbers, etc.; and (ii) words and phrases that reveal what company the document pertains to. Similarly to [24], these authors use the Stanford Named Entity Recognizer [26] to automatically recognize entities which belong to the former category and a Naive Bayes classifier for the latter. Note that, the use of trained classifiers may hamper the generality of the method, in addition to requiring manual training. An interesting feature is the fact that the sanitization is configurable according to the proposed "K-confusability" model. The authors state that a third party cannot ascertain which company among other $K - 1$ appears on a sanitized document that holds K-confusability. Similarly to the "K-safety"[16], this requires from homogeneous sets of documents to sanitize.

Finally [19] solely focuses on the protection of already detected entities, trying to preserve the utility of sanitized documents by means of generalizing terms. The main contribution of this work a theoretic measure ("t-plausibility") on the quality of sanitized documents from a privacy protection point of view. A generalized text document holds the t-plausibility model if at least t base documents can be generalized to such a sanitized document where a base text refers to one that has not been sanitized in any way. Therefore, formal reasonings on how and why a more general term is chosen to replace sensitive information are provided. To enable a general-purpose solution, WordNet [25] is used to generalize the entities. However, authors admit that an optimal sanitization according to their model is NP-hard. Moreover, they also note that setting the t-plausibility level is not intuitive for end users and that their expected results are hardly predictable, since it depends on variable dimensions like document size, amount of sensitive entities and the number of generalizations.

B. Contribution and Plan of this Paper

In this paper, we present a new document sanitization method that, based on information theory, offers the following contributions:

- It focuses on preserving the utility of the sanitized document. Contrary to methods based on term suppression [2], [16], [23] our method relies on knowledge bases to hide sensitive information while preserving its meaning, retaining more document utility.
- It relies on external general-purpose knowledge bases/corpora instead of problem-specific ad-hoc knowledge bases or trained classifiers [2], [9], [16], [22], [23] that, as reported previously, severely hamper the applicability of sanitization methods. As a result, our method offers a domain independent solution that can be applied to textual documents regardless their contents.
- It is fully automatic, requiring no user supervision during the sanitization process (detection and hiding of sensitive parts). This provides a more scalable solution than manual and semiautomatic methods [9], [22], enabling its application to environments with large sanitization needs.
- It allows the user to configure the level of sanitization applied to the document being more flexible than methods based on a fixed sanitization policies [2], [9], [22]–[24]. Moreover, since it is based on the well-known information theory and on linguistic labels, its configuration is more intuitive and comprehensible than methods based on abstract numerical models [9], [16], [19].

The rest of the paper is organized as follows. Section II presents and formalizes our proposal, covering the detection of sensitive terms and their sanitization. Section III evaluates our method from the perspectives of disclosure risk and document's utility preservation, in comparison with other general-purpose approaches. Section IV reports the conclusions and presents some lines of future research.

II. Proposed Method

Since we propose to design a general purpose and unsupervised text sanitization method, in this section, we first define what we consider *sensitive information* and how this can be automatically detected. Next, we detail our sanitization strategy, aiming at retaining a document's utility while ensuring a user-configurable level of privacy.

A. Detecting Sensitive Information

Sensitive information refers to pieces of text that can either reveal the identity of a private entity or refer to confidential information. To discover sensitive information, problem-specific related works rely on predefined lists of sensitive words [16] or use machine learning methods (like trained classifiers [9] or pattern-matching techniques [23]) aimed at detecting specific types of information. The former can provide accurate results, but lists have to be manually compiled (which is costly and time-consuming) for specific problems (which lacks generality); the latter methods manually train/design classifiers/patterns to detect domain specific sensitive data (like PHIs in the medical context [14], [23] or organizational data [9]), which can be hardly generalized.

On the other hand, general purpose methods [9], [24] usually associate the discovery of sensitive data to the detection of generic Named Entities (NEs). Due to their specificity and the fact that they represent individuals rather than concepts, NEs are likely to reveal private information. NEs can be accurately detected in an automatic manner, either using matching patterns or trained classifiers [26], [27]. However, they are hampered by two main problems. First, as it will discussed later, not all NEs refer to sensitive information and not all sensitive data are represented by means of NEs. Second, most generic NE recognition packages detect a limited amount of NE types, usually *persons*, *locations* and *organizations* [26]. Both of these problems negatively affect the detection recall, which is crucial for text sanitization.

To overcome these problems, we base the text sanitization on a more general notion of *"what sensitive information is?"*. To define this, we rely on the following arguments. First, as also noted in NE-based methods [9], [24], sensitive terms are such that, due to their specificity, provide *more information* than common ones. Second, assuming that a potential attacker has basic information about the environment, an ideal text sanitization would focus on those terms that would *increase the information* that the attacker possesses about the entity to protect [24]. Hence, the key-point to detect sensitive information is to quantify *how much information* each textual term provides, removing/hiding those that provide *more information* than what the attacker is assumed to possess.

How can we quantify the amount of information provided by a textual term? In the context of the information theory, this is given by its *information content* (IC). Hence, by accurately quantifying the IC of terms in an automatic manner, and assuming a baseline information for a potential attacker, we will be able to detect and sanitize those terms that provide more information than desired.

1) Computing the IC of Terms: The Information Content (IC) of a term states the amount of information it provides. General terms (e.g., *sports*, *diseases*) provide less IC than more concrete and specialized ones (e.g., *scuba diving*, *tuberculosis*). Formally, the IC of a term t is computed as the inverse of its probability of appearance in a corpus $(p(t))$:

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

To accurately compute IC, the corpus should be large, heterogeneous and up-to-date so that it models the current distribution of terms at a social scale [28]. This will also help to minimize data sparseness problems when dealing with very concrete terms (e.g., rare diseases) or recently minted/trending terms (e.g., netbook, tablet).

When looking for an appropriate corpus, the Web stands out, as it covers almost any possible up-to-date term. Compared to other corpora, the Web provides maximum recall. Moreover, it has been said that the Web is so large and heterogeneous that it represents the true current distribution of terms at a social scale [29]. This configures an ideal corpus for IC calculus [28] and, hence, to assist our document sanitization method.

To compute term probabilities in the Web, authors [30]–[33] have used the *hit count* provided by web search engines (WSEs) when querying the term. Since this enables an automatic and domain-independent calculus, in this work, we compute the IC of terms in this manner.

Definition 1: The IC of a term t is computed from the *hit count* of a WSE as follows:

$$IC_{WSE}(t) = -\log_2 p_{WSE}(t) = -\log_2 \frac{hit\_count_{WSE}(t)}{|webs\_indexed\_by\_WSE|} \quad (2)$$

where $|webs\_indexed\_by\_WSE|$ quantifies the total amount of web sites indexed by the search engine.

2) Document Parsing: Sensitive terms are those referring to concepts or instances that are concrete enough to reveal identities or confidential information. These are referred in text by means of *nouns*, whose semantics can be refined by adding more nouns or adjectives (e.g., sports → water sports), creating *noun phrases (NPs)*.

To enable a coherent sanitization with regards to document semantics, our method focuses on the detection of NPs as potentially sensitive semantic units. To detect NPs, we rely on several natural language processing tools (OpenNLP[1]) enabling *sentence detection*, *tokenization* (i.e., word detection, including contraction separation), *part-of-speech tagging (POS)* and *syntactic parsing*. As a result of this process, POS-tagged words are put together according to their role, obtaining verbal (VPs), prepositional (PPs) or nominal phrases (NPs). From these, only NPs are considered.

Once NPs are detected, we quantify their *amount of information* by querying them in a WSE and using (2).

As a result, given the input document d, we represent it as a list of IC-valued terms:

$$T = \langle NP_1, IC_{WSE}(NP_1)\rangle, \ldots, \langle NP_n, IC_{WSE}(NP_n)\rangle .$$

3) Sanitization Threshold and Sensitive Entity Detection: The next step consists of selecting which NPs provide *too much information*, and hence, which ones should be sanitized. NE-based works assume that NEs *always* provide too much information [24]. From an information theoretic perspective, this is a rough criteria that may result in unnecessarily sanitizing very general terms (e.g., *"Europe"* results in more than 1.000 million hits in Bing, which results in very low IC); at the same time, very informative terms are omitted because they are not NEs (e.g., the concept "metastatic pancreatic carcinoma" provides around 6.300 hits in Bing, which provides a comparatively much higher IC). Moreover, the reliance on NE detection results in a fixed sanitization criterion that cannot be configured for specific scenarios and sanitization needs, which is usually desirable [9].

As discussed in Section I-A, other related works enabled a more flexible sanitization by relying on a theoretical formulation of the desired privacy level. In the *t-plausibility* [19], *K-safety* [16] and *k-confusability* [9] models, authors consider that a document is safely sanitized if it is indistinguishable from at least $t-1$ or $k-1$ other documents. However, as acknowledge by the authors [16], [19] setting up sanitization parameters is difficult and nonintuitive because their influence in the sanitized text would depend on variables like the amount of available documents, the document size, the granularity of the knowledge base (in approaches relying on them) or the amount of sanitized terms.

Our work also enables a flexible, but more intuitive, sanitization. Assuming a *baseline amount of information* that an attacker possesses, our goal is to sanitize those terms that provide *additional information*. This baseline amount of information configures the desired privacy level, so that if one pursues a *high level of privacy*, a *low amount of baseline information* would be assumed, sanitizing many terms because they could easily reveal *new* information, and vice-versa.

To intuitively configure this baseline value, we compute it according to the *amount of information* provided by *the most concrete feature* (φ) that one would like to *reveal* about a private entity. For example, if we wanted to hide the specific address of a company (supposed to be located in *Silicon Valley, San Francisco, California*) but reveal its approximate location (because it is assumed to be known and retains some document's utility), we could specify that $\varphi = "United\ States"$ or $\varphi = "California"$. Then, to set the *baseline amount of information* (β), we compute the $IC_{WSE}(\varphi)$. If one desires to protect several features of a private entity (e.g., activity, location), one can specify a set of features $\Phi = \varphi_1, \ldots, \varphi_k$. Then, β will correspond to the IC of the *most concrete* feature (i.e., the one that provides *maximum information*):

Definition 2: Given the set of features $\Phi = \varphi_1, \ldots, \varphi_k$ that one would like to reveal about a private entity, the baseline amount of information β is computed as:

$$\beta = \max_{\forall \varphi_i \in \Phi} (IC_{WSE}(\varphi_i)) \quad (3)$$

In our method, β acts as a *sanitization threshold*: any term extracted from d that provides more information than β, should be sanitized to avoid revealing more information than desired.

Definition 3: Given the set T of NPs extracted from d, the set of terms to be sanitized (Ψ) is:

$$\Psi = \{NP_i \in T | IC_{WSE}(NP_i) > \beta\} \quad (4)$$

Note that β can be configured using linguistic terms that are coherent with the document's scope, domain, and sanitization needs, providing a more intuitive way to control the sanitization behavior than the use of abstract numerical parameters. Many authors argued that the use of linguistic terms is the most preferable way to express user preferences [34].

B. Sanitizing Sensitive Information

Detected sensitive terms should be removed/transformed so that the amount of information they provide is annulled or, more desirably, reduced enough. Since semantics are the mean to interpret and extract conclusions from the analysis of textual data, the retention of text semantics is crucial to maintain the utility

[1]http://opennlp.apache.org/

of documents [11], [35]. As shown in Section I-A, most related works (e.g., [2], [16], [23]) and classic sanitization approaches (e.g., [20], [21]) simply remove sensitive text, an action that severely hampers document's semantics and, hence, utility [19]. To tackle this problem, recent methods [9], [19], [24] propose replacing sensitive information by generalized versions (e.g., *"iPhone"* → *"cell phone"*). In this manner, the document still retains a degree of semantics (and hence, a level of utility) while revealing less information. To enable term generalizations, a knowledge base (KB) modeling the taxonomic structure of terms to sanitize is needed.

We also rely on KBs and term generalization as the way to sanitize text, exploiting it from an information theoretic perspective. Our method is general enough to be applied to any KB with a taxonomic backbone (e.g., structured thesaurus, folksonomies, ontologies, etc.). In the following, we detail the KBs that are useful in a general-purpose scenario.

1) Knowledge Bases: An ideal KB for text sanitization should have two desirable features. First, it should provide a high *recall*, so that it covers as many sensitive terms found in the input document as possible. This is because, if a sensitive term is not found in the KB, the only option to sanitize would be to remove it [16], to replace it by a random entity [24] or to substitute it by the most general abstraction of the KB. In all cases, an excessive loss of information will occur. Some authors rely on ad-hoc constructed KBs offering a high recall for the sanitized documents [2], [9], [16], [22], [23], but this is neither feasible nor scalable in environments with large and heterogeneous sanitization needs. Second, the KB should offer a *detailed knowledge representation*, so that fine grained taxonomical trees of generalizations can be obtained for a sensitive term. In this manner, the loss of information resulting from each generalization step will be minimized.

The obvious choice for general-purpose sanitization methods is WordNet [25]. It is a domain-independent knowledge source that describes more than 100,000 concepts. These are linked by means of semantic relationships such as hyponymy/hypernymy, meronymy, etc. WordNet offers a detailed and coherent structure, because it has been manually developed by knowledge engineers [25]. Due to its desirable characteristics, several methods [9], [19] have used it to generalize sensitive data. However, WordNet offers a limited coverage of NEs (e.g., proper nouns, locations, brands, etc.)[27]. Moreover, it models concepts in a general manner, so that it rarely cover complex NPs. At the same time, these terms are the typical focus of sanitization due to their high specificity.

To improve the recall, other repositories can be used. The Open Directory Project (ODP)[2] is the largest, most comprehensive human-edited directory of the Web. The purpose of ODP is to list and categorize web sites. Manually created *categories* are taxonomically structured and associated with related web resources. This structure can be used in the same manner as WordNet to retrieve term generalizations. The advantage is its large size and high recall, with more than 1 million categories covering up-to-date NEs.

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

In this work, we use both WordNet and ODP even though WordNet, given its higher taxonomic coherency, is preferred in cases in which terms are found in both repositories.

2) Optimal Generalization of Sensitive Data: From the data utility perspective, an optimal sanitization is such that, while fulfilling the desired level of privacy, minimizes the loss of information resulting from hiding sensitive data. In our method, the level of privacy is stated by the sanitization threshold β, so that none of the terms appearing in a sanitized document provide more information than β. At the same time, sanitized terms could retain up-to β of their original information (i.e., semantics), so that they are still useful while being general enough to minimize the disclosure risk. To achieve these, we rely on term generalization to reduce the amount of information provided by sensitive terms while retaining a degree of their semantics.

Given a set Ψ of terms to sanitize, we propose replacing them by their *generalizations* that provide the *maximum information while fulfilling* β. By picking up the most informative generalization, the sanitized document retains maximum semantics and, hence, utility. To do so, each NP_i in Ψ is mapped to its conceptual abstraction in the KB. When found, the KB returns a hierarchy of generalizations $H_i = h_{i1} \rightarrow \ldots \rightarrow h_{il}$ to which NP_i belongs. For example, if we look for *"iPhone"* (covered by ODP, but not by WordNet), ODP will return the hierarchy: *"iPhone"* → *"Smartphones"* → *"Handhelds"* → *"Systems"* → *"Computers"*. Then, our method selects the generalization that sanitizes NP_i by looking for the h_{ij} in H_i that provides maximum IC while fulfilling β.

Definition 4: Given a hierarchy of generalizations $H_i = h_{i1} \rightarrow \ldots \rightarrow h_{il}$ corresponding to NP_i, and a sanitization threshold β, the sanitized generalization of $NP_i(\Gamma(NP_i))$ is selected as follows:

$$\Gamma(NP_i) = \underset{\forall h_i \in H_i | IC_{WSE}(h_i) < \beta}{\arg\max} (IC_{WSE}(h_i)) \qquad (5)$$

If NP_i is not found in any KBs, we look for its simpler forms by iteratively removing adjectives/nouns starting from the one most on the left (e.g., *"metastatic pancreatic cancer"* → *"pancreatic cancer"* → *"cancer"*). This improves the recall of specific NPs, which are hardly found in some KBs, while maintaining the core semantics. Only when the simplest form of NP_i is not found in any of the KBs (for example, if it is misspelled), NP_i will be replaced by the most abstract generalization (e.g., "world").

Note that this process provides optimum sanitizations (regarding the fulfillment of the desired privacy level and the maximization of the document's utility) in an efficient manner with regards to β and the background KBs. Computationally, our method scales $O(|\Psi| \cdot g)$, where $|\Psi|$ is the number of terms to sanitize and g is a constant stating the maximum number of generalizations for each term. This is several orders of magnitude lower than other knowledge-based methods (e.g., [16], [19]), for which the optimal generalization ends to be NP-hard.

III. Evaluation

In this section, we evaluate the behavior of our method from two perspectives: (i) the accuracy in detecting sensitive

TABLE I
ENTITIES, FEATURES (φ), AND THRESHOLDS (β) USED IN THE EVALUATION, WITH ASSOCIATED HIT COUNTS (IN MILLIONS, FROM BING) AND IC VALUES

Entity (type)	hit_count	$IC(Entity)$	Summary	φ	hit_count	$\beta = IC(\varphi)$
Wozniak (Anglo-Saxon person)	8,4 M	8,7	American computer engineer and programmer who founded Apple Inc. with Steve Jobs.	Steve Jobs Engineer Apple	55 M 226,7 M 776,3 M	5,99 3,94 2,17
Gaudi (Spanish person)	18 M	7,6	Spanish/Catalan architect of Catalan Modernism. He was born in Reus. His work is concentrated in Barcelona.	Reus Architect Barcelona	45,7 M 168,8 M 409 M	6,26 4,37 3,09
Dreamworks (Anglo-Saxon organization)	11,5 M	8,25	American film studio which develops, produces, and distributes films, video games and television programming.	Shrek Producer Hollywood	52,4 M 284,8 M 392,4 M	6,06 3,62 3,15
PortAventura (Spanish organization)	3 M	10,18	Spanish theme park and a resort in Salou, Catalunya; by the Mediterranean, close to Barcelona.	Salou Catalunya Mediterranean	12,2 M 52,7 M 129,7 M	8,16 6,05 4,75
Yellowstone (Anglo-Saxon location)	60 M	5,86	U.S. national park located primarily in the state of Wyoming although it also extends into Montana and Idaho.	National park Wyoming North America	83,2 M 159 M 302,6 M	5,39 4,46 3,53
Tarragona (Spanish location)	48 M	6,19	Capital of a Spanish province located in the south of Catalunya on the north-east of Spain, by the Mediterranean.	Catalunya Mediterranean Province	52,8 M 113 M 280 M	6,05 4,95 3,64

information, focusing on its influence in the disclosure risk, and (ii) the retention of document semantics (i.e., utility). Our method has been compared with general-purpose unsupervised strategies based on NE detection and/or term suppression.

A. Evaluation Dataset

General-purpose related works usually evaluate their methods by means of synthetically-created datasets that fit with their sanitization models. In K-safety/K-confusability models [9], [16] authors compile a minimum set of document/entities to perform the anonymization. In other cases, the dataset is created from terms appearing in WordNet [19]. This simplifies the detection, because datasets are, in some cases, simple collections of potentially sensitive words, rather than natural language texts [16]. In other cases, since all terms are found in their background KBs, the recall is artificially maximum [19]. These "ideal conditions" suppose a simplification of the sanitization problem and, hence, results could significantly differ from what could be achieved in a real scenario.

To test our method in a more realistic setting, we use *real* raw texts containing highly sensitive information. These correspond to Wikipedia English articles of a set of entities of different domains. Articles have been selected so that they describe *persons*, *organizations* or *locations*. This criterion offers a favorable scenario for NE-based methods. Moreover, for each entity type, we selected two types of articles. The first one refers to Anglo-Saxon worldwide entities, so that NEs appearing in text will typically be expressed in English. This eases the detection for English-trained NE recognizers because it increases their chance of appearing "as is" in training data. The second one refers to Spanish entities that, even though their descriptions are written in English, could include NEs expressed by nontranslatable Spanish words or localisms. In this manner, we can compare the degree of language-dependency of our method against those based on NE detection.

Evaluated entities are listed in Table I, together with the sanitization features (φ) and hit_counts and thresholds (β). A brief *summary* of each entity is also included in Table I to better understand the sense of the features used to guide the sanitization. These have been picked up so that at least the name of the described entity would be hidden. Moreover, they offer different degrees of generality so that we can evaluate the influence of the thresholds in the utility of the sanitized document. To compute the IC of terms, Bing WSE have been used, fixing the total amount of indexed web sites in 3.5 billions.[3]

B. Sensitive Data Detection

The first test evaluates the accuracy of our method in automatically detecting sensitive information, analyzing its influence in the disclosure risk. As discussed in Section II-A, our method relies on the assumption that sensitive data can be detected by computing their IC according to their distribution in the Web. The goal of this first evaluation is to test the practical goodness of this assumption. To do so, we requested two human experts to select and agree on which terms (i.e., words or NPs) could reveal too much information for each entity and sanitization threshold. We refer to the set of terms selected by the human experts as Ω. The detection accuracy is then quantified by means of *precision* and *recall* measures.

Precision ((6)) is computed as the ratio between the number of automatically detected sensitive terms (in Ψ) that have been also selected by the human expert (in Ω), and the total amount of automatically detected terms (i.e., $|\Psi|$). The higher the precision, the lower the amount of unnecessary sanitization would occur in the later stage.

$$Precision = \frac{|\Psi \cap \Omega|}{|\Psi|} \cdot 100 \tag{6}$$

Recall ((7)) indicates how much sensitive terms have been detected. It is computed as the ratio between the number of terms in Ψ that also belong to Ω, and the total amount of terms detected by the human expert (i.e., $|\Omega|$). The higher the recall, the lower the disclosure risk, because a lower amount of potentially identifying terms would remain in the sanitized document.

$$Recall = \frac{|\Psi \cap \Omega|}{|\Omega|} \cdot 100 \tag{7}$$

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

F-measure ((8)) quantifies the harmonic mean of recall and precision:

$$F - measure = \frac{2 \cdot Recall \cdot Precision}{Recall + Precision} \quad (8)$$

Our method has been compared against a strategy based on NE-detection (like [9], [24]), using the state-of-the-art *Stanford Named Entity Recognizer* [26]. The first three columns of Fig. 1 show evaluation results for the different entities, thresholds (i.e. β values reported in Table I, which define the X-axis of each graph) and evaluation measures.

Analyzing *precision* (first column of graphs in Fig. 1), we observe that the NE-based schema tends to provide higher figures than our method. Precision mainly depends on the amount of false positives that, in our case, tend to be higher than for the NE-based one. This is because our method tends to select complex NPs as sensitive information. Complex NPs are those composed by several words and/or those presenting complex syntactic constructions that, when queried "as is" in the Web, tend to produce a relatively low hit count, resulting in high IC values. This behavior is caused by the strict terminological matching applied by Web search engines. The NE-based method, on the contrary, omits NPs not containing proper nouns.

Even though a high precision is desirable to incur in a lower information loss in the subsequent sanitization stage *recall* figures, shown in the second column of graphs in Fig. 1, suggest that the NE-based method suffers from a significantly higher disclosure risk. Low recall implies that a number of terms considered as sensitive will appear in the sanitized document, regardless of what is done in the second stage. Results show that our method produces significantly higher recall values than NE-based methods, achieving, in several cases, perfect (100%) results. This shows that not only NEs appearing in text provide too much information, but also NPs referring to concrete concepts. Differences are higher when dealing with Spanish entities (i.e., Gaudi, Tarragona, PortAventura), for which our method doubles or even triples the recall figures of the NE-based method. This indicates that not only concrete NPs are missed by NE-based methods but also an amount of NEs because, in some cases, they refer to Spanish localisms. For example, for the Tarragona entity, the NE recognizer failed to detect NEs like *"Tarragonés"* (the county in which Tarragona is located) or *"Catalan"* (Tarragona is located in Catalonia), which are highly revealing. The case of PortAventura was more serious, since the entity's name was not detected, resulting in an instant disclosure. This shows the limitations of classifiers based on training data: they base the recognition on the fact that the entity or a similar one has been previously tagged. When aiming at designing a general-purpose method, training data may not be enough when dealing with specific entities, or they may be outdated with regards to recently minted entities. This is however, the most common sanitization scenario. In comparison, our method bases the detection on the fact that few evidences are found in the Web. This is a more desirable behavior because sensitive data is detected when it is very likely to act as an identifier. The reliance on the lack of evidences rather than on the presence of them also avoids being affected by the data sparseness that characterizes manual training/knowledge-based models [28]. Moreover, contrary to tagged corpora, the Web offers up-to-date results and covers almost any possible domain [28].

Regarding threshold influence (β), we observe that recall for the NE-based method usually decreases (and, hence, disclosure risk rises) as thresholds decrease (i.e., sanitization features become are more general), a circumstance that would require a more exhaustive sanitization. Precision values behave inversely. We also observe that, since the NE-based method does not adapt its behavior to the sanitization threshold, its accuracy would depend on the suitability of the threshold and the entities appearing in text. If the threshold is more general than the average generality of NEs found in text, NE-based recall decreases because many too informative terms are omitted. Inversely, when the threshold is more concrete, too many NEs will be tagged as sensitive, resulting in lower precision. On the contrary, our method adapts its behavior to the sanitization needs maintaining recall values stable in the 85–100% range. Precision is more variable; it tends to decrease when sanitization features become more concrete (i.e. lower thresholds) because of the above-mentioned tendency to detect terminologically complex NPs as sensitive data.

As a result of the differences between methods' recalls (i.e., disclosure risk), when comparing their global accuracy (i.e., *F-measure*), our method surpass NE-based ones in all cases (see third column of graphs in Fig. 1). Therefore, even considering precision and recall as equally important, our method produces a more accurate sanitization.

C. Utility-Preserving Sanitization

The second test evaluates the behavior of the utility-preserving sanitization procedure detailed in Section II-B.

As discussed in Section II-B2, the utility of textual data is given by their semantics. To quantify the amount of semantics preserved for a given document d and evaluate our method, we quantify the *amount of information* provided by its sanitization result (document d') in comparison with the original d. This quantification makes sense given that our sanitization method is based on *taxonomically generalizing* terms; that is, terms in d' refer to *concept generalizations* of those referred in d, so that the semantics in d' includes those provided by d.

The utility of d' is measured by the *amount of information* preserved during the sanitization process, as follows.

First, the *amount of information provided by a document* d (i.e., $IC(d)$) is computed as the sum of IC of its NPs.

$$IC(d) = \sum_{\forall NP_i \in d} IC(NP_i) \quad (9)$$

Then, the utility of the sanitized document d' with respect to d is the percentage of preserved information:

$$Utility(d') = \frac{IC(d')}{IC(d)} \cdot 100 \quad (10)$$

In this section, we present the utility figures obtained when sanitizing sensitive terms for the entities and thresholds used in the previous section. To contextualize and compare our utility values, we have also implemented and evaluated the following sanitization strategies used by related works:

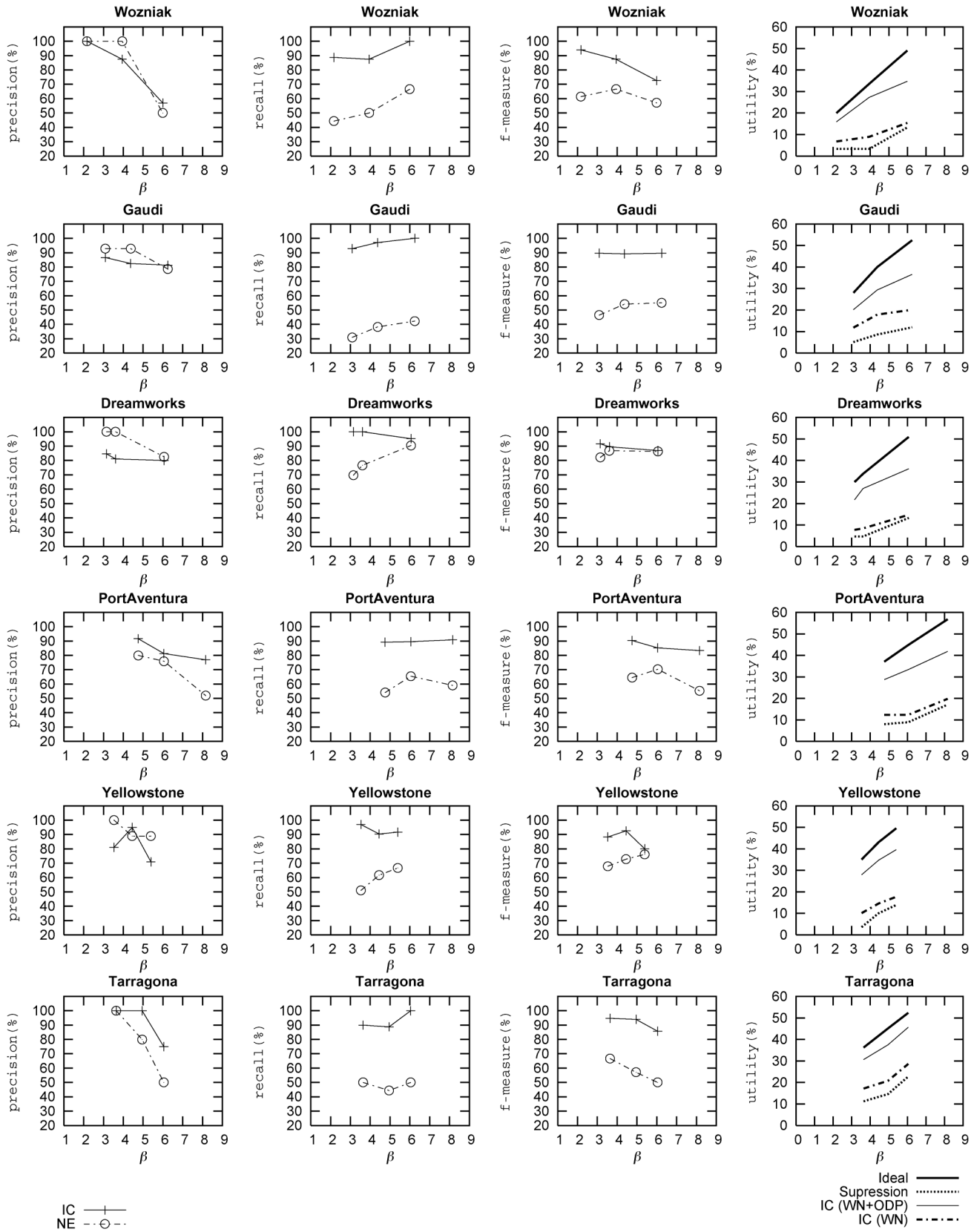

Fig. 1. Evaluation results. For each entity: *precision*, *recall*, *F-measure* (for the sensitive term detection stage), and *utility* (for the term sanitization stage).

- *WordNet-based generalization.* The same sanitization method described in Section II-B, but relying solely on WordNet (rather than WordNet plus ODP) to obtain term generalizations. As discussed in Section II-B1, WordNet offers a limited coverage of concrete terms, such as NEs or specific NPs, which usually correspond to sensitive

data. Even though, some related works [9], [19] only use WordNet to propose generalizations.

- *Suppression.* Sensitive terms are directly removed from text [2], [16], [22], [23].

To contextualize utility values, we have also computed the utility resulting from an *ideal sanitization*. This is such that, for a given set of sensitive terms and a threshold β, each term t is replaced by a generalization $(\Gamma(t))$ that provides the same amount of information as the threshold (i.e., $IC(\Gamma(t)) = \beta$). In this manner, β is fulfilled while resulting in *minimal information loss*. This scenario can be only achieved if the KB provides the ideal $\Gamma(t)$, which is extremely rare. However, by quantifying the degree of utility preservation under ideal circumstances, we can set an upper-bound, contextualize the analysis and evaluate the suitability of the KB, whose taxonomic granularity enables a more or less accurate approximation of the ideal generalization. The ideal utility preservation value is computed by assuming that each sensitive term t has been replaced by a *fictitious* generalization that fulfills $IC(\Gamma(t)) = \beta$.

Data utility for the different entities, thresholds and sanitization strategies are shown in the fourth column in Fig. 1.

Figures are consistent for the different tests and sanitization strategies. The best results are obtained when replacing sensitive terms with the most concrete generalization that fulfills the thresholds, using both WordNet and ODP. Utility values decrease quite linearly as the sanitization threshold is more general (i.e., more exhaustive sanitization). Utility preservation ranges around 20–45%, growing as thresholds increase. Even though these values may seem low, they are close to the ideal upper-bound, representing around a 60–80% of the total amount of information that can be preserved in ideal circumstances. Values are also coherent to the fact that sanitized texts corresponds to biographical sketches and detailed descriptions so that most of the information they contain can identify the described entity. Hence, even with the more relaxed sanitization needs (i.e., most concrete thresholds) much information should be hidden.

When applying the same method but relying solely on WordNet, utility values are significantly worse (i.e., less than half utility than with ODP+WN). This indicates that, in most cases, sensitive terms could not be found in WordNet and, hence, since no generalization trees are available, they are replaced by the most general abstraction (i.e., *"world"*), producing a high information loss. As discussed in Section II-B1, WordNet offers a very limited coverage for NEs and concrete terms, which are the focus of the sanitization process. This shows the importance of using adequate KBs to assist the sanitization, so that the ideal generalization can be more closely approximated. Obviously, the suppression of sensitive terms results in the worst utility, with values that are below a 10% in many cases. This suggests that the quality of the sanitized document has been severely affected.

IV. Concluding Remarks

In this paper, an automatic text sanitization method has been proposed. It relies on the theoretical foundations of the information theory and a corpus as global as the Web to offer a general-purpose solution that can be applied to heterogeneous textual data (and not only NEs [9], [24]). Contrary to methods based on k-anonymity models [9], [16], which deal with groups of k documents with a similar structure/topic in order to swap and replace sensible entities, our method is able to sanitize each document independently. Moreover, it offers a flexible and intuitive way (in comparison with abstract numerical parameters [16], [19]) to configure the sanitization degree, based on domain-specific linguistic features. Finally, special care has been put in the preservation of document's utility, as a function of its semantics. General-purpose knowledge sources have been used to reduce the amount of information given by document terms while maintaining, up to a degree, their semantics. Evaluation results, obtained for entities of different domains, sustained the theoretical premises, showing a high detection recall in comparison with general-purpose approaches based on trained classifiers. Document's utility was also better retained, in comparison with methods based on term suppression, with values close to the ideal sanitization and coherent with sanitization thresholds.

As future work, some aspects can be improved to provide more accurate sanitizations. First, as stated in Section III-B, the term detection *precision* suffers when dealing with complex NPs, whose IC values appear to be higher than what they should. This is caused by the different lexico-syntactical forms of the same concept. As a result, IC values depend on *term variability* (i.e., synonymy). To obtain more accurate IC values, we can use synonyms dictionaries or construct semantically equivalent NPs so that the final IC value can be computed as the *sum* of the IC of each variation.

Moreover, it is important to note that our method, in the same manner as most related works, sanitize terms independently. However, some authors [36] have noted that the analysis of the relationship between terms may increase the disclosure risk. Certainly, the *cooccurrence* of terms (e.g., *medical conditions + treatments*) may provide *more information* than the information obtained when analyzing them independently. To quantify the amount of information provided by term tuples *collocation measures* [30], which also consider the amount of information provided by the cooccurrence of terms, can be used in further research.

References

[1] U.S. Department of Justice, U.S. Freedom of Information Act (FOIA) 2012 [Online]. Available: http://www.foia.gov/

[2] A. Tveit, O. Edsberg, T. B. Rost, A. Faxvaag, O. Nytro, M. T. Nordgard, M. T. Ranang, and A. Grimsmo, "Anonymization of general practioner medical records," in *Proc. Second HelsIT Conf.*, Trondheim, Norway, 2004.

[3] S. Paquette, P. T. Jaeger, and S. C. Wilson, "Identifying the security risks associated with governmental use of cloud computing," *Gov. Inf. Quart.*, vol. 27, no. 3, pp. 245–253, 2010.

[4] S. Marston, Z. Li, S. Bandyopadhyay, A. Ghalsasi, and J. Zhang, "Cloud computing the business perspective," *Decision Support Syst.*, vol. 51, no. 1, pp. 176–189, 2011.

[5] S. K. Dash, R. Mishra, D. P. Mishra, and A. Tripathy, "A privacy preserving repository for securing data across the cloud," in *Proc. 3rd Int. Conf. Electronics Computer Technology*, 2011, vol. 5, pp. 6–10.

[6] D. Chen and H. Zhao, "Data security and privacy protection issues in cloud computing," in *Proc. 2012 Int. Conf. Computer Science and Electronics Engineering*, 2012, pp. 647–651.

[7] S. Pignal, "EU eyes big fines for privacy breaches," *Financial Times* 2011 [Online]. Available: http://www.ft.com/intl/cms/s/2/bf962998-1d01-11e1-a26a-00144feabdc0.html#axzz1fe8ewpqO

[8] Department of Health and Human Services, Office of the Secretary, The Health Insurance Portability and Accountability Act of 1996, Tech. Rep. Federal Register 65 FR 82462, 2000.

[9] C. Cumby and R. Ghan, "A machine learning based system for semi-automatically redacting documents," in *Proc. 23rd Innovative Applications of Artificial Intelligence Conf.*, 2011, pp. 1628–1635.
[10] L. Sweeney, "K-anonymity: A model for protecting privacy," *Int. J. Uncertainty, Fuzziness and Knowledge-Based Syst.*, vol. 10, no. 5, pp. 557–570, 2002.
[11] S. Martínez, D. Sánchez, A. Valls, and M. Batet, "Privacy protection of textual attributes through a semantic-based masking method," *Inf. Fusion*, vol. 13, no. 4, pp. 304–314, 2012.
[12] S. Martínez, D. Sánchez, and A. Valls, "Semantic adaptive microaggregation of categorical microdata," *Comput. Security*, vol. 31, no. 5, pp. 653–672, 2012.
[13] D. Sánchez, J. Castellá-Roca, and A. Viejo, "Knowledge-based scheme to create privacy-preserving but semantically-related queries for web search engines," *Inf. Sci.*, vol. 218, no. 1, pp. 17–30, 2012.
[14] S. M. Meystre, F. J. Friedlin, B. R. South, S. Shen, and M. H. Samore, "Automatic de-identification of textual documents in the electronic health record: A review of recent research," *BMC Med. Res. Methodol.*, vol. 10, pp. 70–86, 2010.
[15] D. A. Dorr, W. F. Phillips, S. Phansalkar, S. A. Sims, and J. F. Hurdle, "Assessing the difficulty and time cost of de-identification in clinical narratives," *Methods Inf. Medicine*, vol. 45, no. 3, pp. 246–252, 2006.
[16] V. T. Chakaravarthy, H. Gupta, P. Roy, and M. Mohania, "Efficient techniques for document sanitization," in *Proc. ACM Conf. Information and Knowledge Management'08*, 2008, pp. 843–852.
[17] U.S. Department of Energy, Department of Energy Researches Use of Advanced Computing for Document Declassification 2012 [Online]. Available: http://www.osti.gov/opennet
[18] DARPA, New Technologies to Support Declassification Request for Information (RFI) Defense Advanced Research Projects Agency. Solicitation Number: DARPA-SN-10-73, 2010.
[19] B. Anandan, C. Clifton, W. Jiang, M. Murugesan, P. Pastrana-Camacho, and L. Si, "t-plausibility: Generalizing words to desensitize text," *Trans. Data Privacy*, vol. 5, pp. 505–534, 2012.
[20] Nat. Security Agency, Redacting With Confidence: How to Safely Publish Sanitized Reports Converted From Word to pdf, Tech. Rep. I333-015R-2005, 2005.
[21] National Security Agency, Redaction of pdf Files Using Adobe Acrobat Professional X 2011 [Online]. Available: http://www.nsa.gov/ia/files/vtechrep/I73 025R_2011.pdf
[22] L. Sweeney, "Replacing personally-identifying information in medical records, the scrub system," in *Proc. 1996 American Medical Informatics Association Ann. Symp.*, 1996, pp. 333–337.
[23] M. M. Douglass, G. D. Cliffford, A. Reisner, W. J. Long, G. B. Moody, and R. G. Mark, "De-identification algorithm for free-text nursing notes," *Proc. Computers in Cardiology'05*, pp. 331–334, 2005.
[24] D. Abril, G. Navarro-Arribas, and V. Torra, "On the declassification of confidential documents," in *Proc. Modeling Decisions for Artificial Intelligence'11*, 2011, pp. 235–246.
[25] C. Fellbaum, *WordNet: An Electronic Lexical Database*. Cambridge, MA, USA: MIT Press, 1998.
[26] J. R. Finkel, T. Grenager, and C. Manning, "Incorporating non-local information into information extraction systems by gibbs sampling," in *Proc. 43rd Ann. Meeting of the Association for Computational Linguistics*, 2005, pp. 363–370.
[27] D. Sánchez, D. Isern, and M. Millan, "Content annotation for the semantic web: An automatic web-based approach," *Knowl. Inf. Syst.*, vol. 27, no. 3, pp. 393–418, 2011.
[28] D. Sánchez, M. Batet, A. Valls, and K. Gibert, "Ontology-driven web-based semantic similarity," *J. Intell. Inf. Syst.*, vol. 35, no. 3, pp. 383–413, 2010.
[29] R. L. Cilibrasi and P. M. B. Vitanyi, "The Google similarity distance," *IEEE Trans. Knowl. Data Eng.*, vol. 19, no. 3, pp. 370–383, Mar. 2007.
[30] P. D. Turney, "Mining the web for synonyms: Pmi-ir versus lsa on toefl," in *Proc. 12th Eur. Conf. Machine Learning*, 2001, pp. 491–502.
[31] D. Sánchez, "A methodology to learn ontological attributes from the web," *Data Knowl. Eng.*, vol. 69, no. 9, pp. 573–59, 2010.
[32] D. Sánchez and D. Isern, "Automatic extraction of acronym definitions from the web," *Appl. Intell.*, vol. 34, no. 2, pp. 311–327, 2011.
[33] A. Viejo, D. Sánchez, and J. Castellá-Roca, "Preventing automatic user profiling in web 2.0 applications," *Knowl.-Based Syst.*, vol. 36, pp. 191–205, 2012.
[34] L. Marin, D. Isern, A. Moreno, and A. Valls, "On-line dynamic adaptation of fuzzy preferences," *Inf. Sci.*, vol. 220, pp. 5–21, 2013.
[35] V. Torra, "Towards knowledge intensive data privacy," in *Proc. 5th Int. Workshop (DPM 2010), and 3rd Int. Workshop (SETOP 2010)*, 2011, pp. 1–7.
[36] B. Anandan and C. Clifton, "Significance of term relationships on anonymization," in *Proc. Web Intelligence/IAT Workshops'11*, Lyon, France, 2011, pp. 253–256.

David Sánchez is a tenure-track lecturer at Universitat Rovira i Virgili, Tarragona, Spain. He received the Ph.D. degree in computer science from the Technical University of Catalonia, Barcelona, Spain, in 2008.

His research interests include data semantics, knowledge acquisition, and privacy preservation of textual data. He has participated in several National and European funded research projects and authored several papers and conference contributions.

Montserrat Batet is a postdoctoral researcher at Universitat Rovira i Virgili, Tarragona, Spain. She received the Ph.D. degree in computer science from the Universitat Rovira i Virgili in 2011.

Her research interests include the assessment of semantic similarity, semantic data clustering, and privacy preservation of textual data. She has participated in several National and European funded research projects and authored several papers and conference contributions.

Alexandre Viejo is a tenure-track lecturer at Universitat Rovira i Virgili, Tarragona, Spain. He received the Ph.D. degree in computer science from the Universitat Rovira i Virgili in 2008.

In 2009, he was a researcher at Humboldt-Universität zu Berlin, Berlin, Germany. He has authored several papers and conference contributions. His fields of activity are data privacy, data security, and cryptographic protocols.