Detecting sensitive information from textual documents: an information-theoretic approach

David Sánchez, Montserrat Batet, Alexandre Viejo

Departament d'Enginyeria Informàtica i Matemàtiques,
UNESCO Chair in Data Privacy, Universitat Rovira i Virgili
Av. Països Catalans 26, E-43007 Tarragona, Spain
{david.sanchez,montserrat.batet,alexandre.viejo}@urv.cat

Abstract. Whenever a document containing sensitive information needs to be made public, privacy-preserving measures should be implemented. Document sanitization aims at detecting sensitive pieces of information in text, which are removed or hidden prior publication. Even though methods detecting sensitive structured information like e-mails, dates or social security numbers, or domain specific data like disease names have been developed, the sanitization of raw textual data has been scarcely addressed. In this paper, we present a general-purpose method to automatically detect sensitive information from textual documents in a domain-independent way. Relying on the Information Theory and a corpus as large as the Web, it assess the degree of sensitiveness of terms according to the amount of information they provide. Preliminary results show that our method significantly improves the detection recall in comparison with approaches based on trained classifiers.

Keywords: Privacy, Document sanitization, Information Theory.

1 Introduction

In the context of the Information Society, many documents are needed to be made public every day [1]. Since some of these documents may contain confidential information about private entities, measures should be taken prior their publication to avoid revealing sensitive data or disclosing individuals' identities.

Document sanitization precisely pursuits the removal of sensitive information from text (which can yield to revealing private information/identities of the entities referred in the document) so that it may be distributed to a broader audience.

In the past, sanitization has been usually tackled manually by governments and companies. Standard guidelines [2] detailing the correct procedures to ensure irreversible suppression or distortion of sensitive parts in physical and electronic documents have been proposed. In the medical context, the *Health Insurance Portability and Accountability Act (HIPAA)* [3] states safe harbor rules about the kind of personally identifiable information which should be removed in medical documents prior allowing their publication.

However, manual sanitization is expensive, time-consuming [4], prone to disclosure risks [5] and does not scale as the volume of data increases [6]. Considering the amount of digital textual information made available daily (*e.g.,* the US Department of Energy's OpenNet initiative [7] requires sanitizing millions of documents yearly), one can realize of the need of automatic text sanitization methods. This need is manifested in initiatives like the DARPA's request for new technologies to support the declassification of confidential documents [8] or the creation of the Consortium for Healthcare Informatics Research (CHIR) [9], which aims at building new methods and tools for de-identification of medical data in order to utilize them for research and operational purposes.

To tackle this problem, semi-automatic applications assisting the sanitization process have been developed, focusing on structured sensitive data like email addresses, dates, telephone numbers or credit card/social security numbers. Commercial applications like Adobe Acrobat Professional [10] incorporate patterns that are able to recognize this kind of data thanks to its regular structure. However, they leave the detection of sensitive textual data (like names, locations or descriptive assertions) to a human expert. In fact, the sanitization of this kind of free text data (which is the most usually available one) is specially challenging due to its unbounded and unstructured nature [11].

In this paper, we tackle the problem of automatic detection of sensitive text for sanitization purposes. Relying on the foundations of the Information Theory, we mathematically formulate what we consider *sensitive information* and how it can be applied to detect potentially sensitive textual entities. Our method has been compared to other general-purpose approaches relying on trained classifiers, showing that it is able to improve the recall detection while offering a more general and less constrained solution.

The rest of the paper is organized as follows. Section 2 describes related works focusing on detecting sentitive terms in textual documents. Section 3 presents our method, discussing its theoretical premises and formalizing its design. Section 4 details preliminary experiments carried out with highly identifying biographical sketches, showing promising results regarding the detection recall. The final section depicts the conclusions and presents some lines of future research.

2 Related work

Among the unsupervised sanitization methods available, one of the first approaches that can be found is the Scrub system [12]. It finds and replaces patterns of identifying information such as Social Security number, medical terms, age, date, etc. Similar schemes that focus on removing sensitive terms from medical records [13, 14] use very specific patterns designed according to the HIPAA "Safe Harbor" rules that mention 18 data elements that must be removed from clinical data in order to anonymize it [3]. Examples of those sensitive elements are: names, dates, medical record numbers, biometric identifiers, full face photographs, etc.

The authors in [6] present a scheme that detects sensitive elements using a database of entities (persons, products, diseases, etc.) instead of patterns. Each entity in this database is associated with a set of terms related to the the entity; this set is the context of the entity that should be hidden (*e.g.,* the context of a person entity could include her name, birth date, etc).

The method proposed in [11] focuses on domain-independent unstructured documents. Authors propose the use of named entity recognition techniques to identify the entities of the documents that require protection. It is worth to mention that this proposal assumes that named entities (such as person and organization names and locations) are always sensitive data and, hence, they should be sanitized.

The authors in [5] present a semi-automatic tool build into Microsoft Word that suggests to the user the entities that should be anonymized. Regarding the entity detection process, this work focuses on documents directly linked to certain companies (*i.e.,* documents to be sanitized describe certain companies/organizations or their activities). The data to be detected is divided into two categories: (i) *Client Identifying Information:* this information includes any words and phrases that reveal what company the document pertains to; and (ii) *Personally Identifying Information:* this includes any person names, location names, phone numbers, etc. Similarly to [11], authors uses the Stanford Named Entity Recognizer [15] to automatically recognize people, organizations and locations. Additionally, specific patterns are used to detect social security numbers or telephone numbers. Regarding the Client Identifying Information, a Naive Bayes classifier is implemented to recognize it.

3 A general purpose method to detect sensitive terms in textual documents

Our method pursuits to automatically detect sensitive pieces of text in a general and unconstrained way, so that it can be applied to heterogeneous documents (both regarding its structure and knowledge domain), and to any kind of textual term (instead of predefined types or lists). To do so, we first discuss the notion of sensitive information and how it can be detected.

Sensitive information regards 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 [6] or use machine learning methods (like trained classifiers [5] or pattern-matching techniques [14]) 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, 9] or organizational data [5]), which can be hardly generalized.

On the other hand, general purpose methods [11] 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 patterns [16, 17] or trained classifiers [15]. However, they are hampered by several problems. First, some detected NEs could refer to very general entities (*e.g.,* continents), which are not needed to be sanitized and whose removal would result in unnecessary information loss. On the other hand, some words or combinations of words, which may be omitted since they are not NEs, could refer to very concrete concepts (*e.g.,* rare diseases, concrete employments), which are likely reveal confidential or identifiable information. Moreover, most generic NE recognition packages only detect a limited amount of NE types, usually *persons*, *locations* and *organizations* [11, 15]. Finally, they are language-dependent, since the NE recognition accuracy depends on the availability of training data, which is expressed in a concrete language. These problems negatively affect the detection recall, which is crucial to avoid disclosure risk.

To overcome these problems, we base the text sanitization on a more general notion of *sensitive information.* In our approach sensitive terms are those that, due to their specificity, provide *more information* than common terms. Hence, the key-point to detect them is to quantify *how much information* each textual term provides, sanitizing those that provide *too much information* (according to a sanitization criteria).

To quantify the amount of information provided by a textual term, we rely on the information theory and the notion of Information Content (IC).

3.1 Information Content estimation

The Information Content (IC) of a term measures the amount of information provided by the given term when appearing in a context (*e.g.,* a document). Specific terms (*e.g., pancreatic cancer*) provide more IC than those more general ones (*e.g., disease*). Formally, the IC of a term t is computed as the inverse of the probability of encountering t in a corpus ($p(t)$). In this way, infrequent concepts obtain a higher IC than more common ones.

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

Classical methods [18] used tagged textual data as corpora, so that term frequencies can be computed unambiguously. The use of this kind of corpus provided accurate results in the past, when applied to general terms [18] at the cost of manually compiling and tagging it. However, the limited coverage and relative small size of used corpora resulted in data sparseness problems (*i.e.,* the fact that not enough data is available to extract reliable conclusions from their analysis) when computing the IC of concrete terms (*e.g.,* rare diseases), NEs (*e.g.,* names) or recently minted/trending terms (*e.g.,* netbook, tablet) [19, 20]. Considering that document sanitization focuses precisely on concrete (*i.e.,* highly informative) terms, a wider corpus covering them would be desirable to obtain robust IC values.

When looking for a general-purpose corpus covering as much terms as possible, the Web stands out. Its main advantages are its free and direct access and its wide coverage of almost any possible up-to-date term. In fact, it has been argued that the Web is so large and heterogeneous that represents the true current distribution of terms at a social scale [21]. Since IC calculus relies on term distribution to compute probabilities, the characteristics of the Web makes it specially convenient [19].

The main problem of computing term appearances in the Web is that the analysis of such an enormous repository is impracticable. However, the availability of Web Information Retrieval tools (IRs) like Web Search Engines (WSEs) can help in this purpose. WSEs directly provide web-scale page counts (stating term appearances) for a given query. Many authors [22, 19, 23] have used these page counts to compute term probabilities in the Web. Hence, by estimating term probabilities at a social/Web scale one can compute, in an unsupervised and domain-independent manner, their IC.

Taking into consideration the Web size, its high coverage for any kind of terms (including concrete ones and NEs) and the possibility of obtaining web-scale term distribution measures in an immediate way, in this work, we quantify the IC of a potentially sensitive term t found in a document d to be sanitized, as follows:

$$IC_{web}(t) = -\log_2 p_{web}(t) = -\log_2 \frac{page\_counts(t)}{total\_webs} \qquad (2)$$

where $page\_counts(t)$ is the number provided by a WSE when querying t and $total\_webs$ quantifies the total amount of web sites indexed by the search engine. (*e.g.,* around 3.5 billions in Bing[1]).

To avoid the need of on-line querying that, in addition to overhead the process, may disclose sensitive words, one can use databases of some WSEs that can be stored and queried off-line [24, 25].

3.2 Extracting terms from textual documents

In this section, we detail how sensitive terms t are extracted from a document d to be sanitized. Given an input text like the one shown in Figure 1, sensitive data is such corresponding to concrete concepts (*e.g.,* pancreatic cancer) or individual names (*e.g.,* Peter Greenow) that reveal too much information. These are referred in text by means of *nouns* or, more generally, *noun phrases (NPs).* Hence, the detection of sensitive terms focuses on NPs found in the input document.

To detect NPs, we rely on several natural language processing tools [26], which perform (i) *sentence detection*, (ii) *tokenization* (*i.e.,* word detection, including contraction separation), (iii) *part-of-speech tagging (POS)* of individual tokens and (iv) *syntactic parsing* of POS tagged tokens, so that they are put together according to their role, obtaining verbal (VPs), prepositional (PPs) or nominal phrases (NPs). From these, NPs are considered (see an example of the

[1] http://www.worldwidewebsize.com/ [last accessed: May 8th, 2012]

Peter Greenow, from Syracuse, United States, suffers from pancreatic cancer. He was given treatment in the Community General Hospital for his condition by an oncologist.

Fig. 1. Sample text of a document to sanitize

output of this analysis in Figure 2). As discussed in the previous section, the *amount of information* NPs provide (IC) will be quantified by querying them in a WSE and using eq. 2.

[*NP* Peter Greenow] , from [*NP* Syracuse], [*NP* United States], suffers from [*NP* pancreatic cancer]. [*NP* He] was given [*NP* treatment] in [*NP* the Community General Hospital] for [*NP* his condition] by [*NP* an oncologist].

Fig. 2. Noun Phrases (NP) detected in sample text

In order to focus the IC-based analysis on the information provided by the conceptualization to which each NP refers, we also remove *stop words.* Stop words configure a finite list of domain independent terms like determinants, prepositions or adverbs which can be removed from NPs without altering their conceptualizations (*e.g.,* an oncologist → oncologist). The motivation of removing stop words is to avoid their influence in the computation of IC values by means of web queries. For example, in a WSE like Bing[2], the query *"an oncologist"* results in a page count (654.000) an order of magnitude lower than the query *"oncologist"* (5.870.000), even though both refer to the same concept and, hence, both should provide the same *amount of information.*

Note that, even though both natural language processing tools and stop words are language-dependent, both are available for many languages, including English, Spanish, Portuguese, German or Danish [26].

3.3 Detecting sensitive terms

The final step consists on assessing which of the NPs provide *too much information* according to their computed IC; these will be considered as *sensitive.*

As discussed at the beginning of the section, NE-based methods assume that NEs *always* provide too much information. From an information theoretic perspective, this is a rough criteria that may result in unnecessarily sanitizing very general terms (*e.g., "United States"* results in 1.300 million pages in Bing, obtaining a very low IC); at the same time, more informative terms are omitted because they are not NEs (*e.g.,* the concept "pancreatic cancer" results in around 6,5 million page counts in Bing, which provides a comparatively much higher IC).

[2] http://www.bing.com/ [accessed: May 8th, 2012]

Relying on the notion of IC, our proposal enables a more comprehensive and adaptable sanitization, which considers as *sensitive* those NPs whose IC (computed using eq. 2) is higher or equal than a given value β, which acts as the *detection threshold*. This value, which is also expressed in terms of IC, represents the *degree of informativeness* above which terms are considered to reveal *too much information.* β can be defined in an intuitive way by associating it to the IC of the most general feature that should remain hidden in the sanitized document.

Formally, any NP_i in d (*i.e.,* the document to be sanitized) whose IC is higher or equal than β will be considered as sensitive:

$$Sensitive\_NPs = \{NP_i \in d | IC_{web}(NP_i) >= \beta\} \tag{3}$$

For example, if we would like to sanitize the text shown in Figure 1 so that a potential attacker cannot discover that *Peter Greenow* has *cancer* (and any other more concrete information, like his name and detailed census data), we can specify the detection threshold as $\beta = IC_{web}(cancer)$. In this manner, any reference to *cancer* or any other more concrete (*i.e.,* more informative) term like *pancreatic cancer* or *Community General Hospital* will be considered as sensitive. Table 1 shows the detection results, presenting sensitive terms (according to the specified threshold) in **bold**. One can realize that some concrete concepts that are not NEs (*i.e.,* oncologist, pancreatic cancer) have been appropriately tagged as sensitive, whereas very general NEs (*i.e.,* United States) have not. Compared to NE-based methods [5, 11], the former case minimizes the disclosure risk, whereas the later case contributes to retain the sanitized document's utility.

Table 1. Detected Noun phrases (NP) with their corresponding *page_counts* (from Bing) and IC_{web}. Words in (brackets) are stop words that are not considered in the IC calculus. **Bold** rows correspond to sensitive terms according to the detection threshold (*i.e.,* $IC_{web}(cancer)$ = 2.7, given that *page_counts*(*cancer*) = 536.000.000). The last column states which ones are Named Entities (NE)

NP	*page_counts*	IC_{web}	NE?
Peter Greenow	**21**	**27.3**	**Yes**
Syracuse	**68.000.000**	**5.7**	**Yes**
United States	1.300.000.000	1.4	Yes
pancreatic cancer	**6.550.000**	**9.1**	**No**
(He)	Not Considered	N/C	No
treatment	616.000.000	2.5	No
(the) Community General Hospital	**146.000**	**14.5**	**Yes**
(his) condition	702.000.000	2.3	No
(an) oncologist	**7.200.000**	**8.9**	**No**

4 Experiments

In this section, some preliminary results are presented, showing the accuracy of the detection when applying our method to highly sensitive textual documents. Since most general-purpose related works [5, 11] rely on the detection of NEs to sanitize text, our method has been compared against the state-of the art *Stanford Named Entity Recognizer* [15], which is able to detect and classify NEs as *persons*, *locations* or *organizations*. Both approaches have been evaluated against the criterion of two human experts stating which pieces of text could reveal too information about the described entity.

To test our method in a realistic setting, we use *real* raw texts containing highly sensitive information. In particular, we used biographical sketches describing *actors/actresses* taken from English Wikipedia articles. Wikipedia descriptions of concrete entities usually contain an high amount of potentially identifiable information, which makes the detection of sensitive information a challenging task. Two types of actors have been selected: three American actors (*Sylvester Stallone*, *Arnold Schwarzenegger* and *Audrey Hepburn*), so that most terms and NEs appearing in text would be expressed in English (easing the detection for English-trained NE recognizers), and three Spanish (but well-known) actors (*Antonio Banderas*, *Javier Bardem* and *Jordi Mollà*) for which, even though their descriptions are written in English, could include NEs expressed with non-translatable Spanish words or localisms. In this manner, we can also compare the degree of language-dependency of our method against NE recognizers based on English-trained classifiers.

To evaluate the results obtained by both methods, we requested two human experts to select and agree on which terms (*i.e.,* words or NPs, including NEs) reveal too much information, considering that it is desired to hide the fact that the described entities are *actors*. Hereinafter, we will refer to the set of sensitive terms selected by the human experts as $Human\_Sensitive\_NPs$. Coherently with our method's design, we set $\beta = IC_{web}(actor)$, so that any term providing more information than the term *actor* will be detected as *sensitive*. To compute the IC of terms Bing Web Search Engine have been used, fixing the total amount of indexed web sites in 3.5 billions [3]. The detection performance is quantified by means of *precision*, *recall* and *F-measure*.

Precision (eq. 4) is calculated as the ratio between the number of automatically detected sensitive terms ($Sensitive\_NPs$) that have been also selected by the human experts ($Human\_Sensitive\_NPs$), and the total amount of automatically detected terms (*i.e.,* $|Sensitive\_NPs|$). The higher the precision, the lower the amount of incorrectly detected sensitive terms.

$$Precision = \frac{|Sensitive\_NPs \cap Human\_Sensitive\_NPs|}{|Sensitive\_NPs|} \cdot 100 \qquad (4)$$

Recall (eq. 5) is calculated as the ratio between the number of terms in $Sensitive\_NPs$ that also belong to $Human\_Sensitive\_NPs$, and the total amount

[3] http://www.worldwidewebsize.com/ [last accessed: May 8th, 2012]

of terms in $Human\_Sensitive\_NPs$. Recall indicates the number of detected sensitive terms. The higher the recall, the less the disclosure risk, because a lower amount of non-detected sensitive terms would remain in the text.

$$Recall = \frac{|Sensitive\_NPs \cap Human\_Sensitive\_NPs|}{|Human\_Sensitive\_NPs|} \cdot 100 \tag{5}$$

Finally, the *F-measure* (eq. 6) quantifies the harmonic mean of recall and precision, summarizing the accuracy of the detection stage:

$$F\text{-}measure = \frac{2 \cdot Recall \cdot Precision}{Recall + Precision} \tag{6}$$

The obtained values for *precision*, *recall* and *F-measure* for our method (*IC*) and for the method based on NE detection (*NE*) are listed in Table 2.

Table 2. Precision, Recall and F-measure for evaluated entities and methods

		Sylvester Stallone	Arnold Schwarz.	Audrey Hepburn	Antonio Banderas	Javier Bardem	Jordi Mollà
Precision	NE	100%	100%	87.10%	100%	100%	75%
	IC	82.35%	72%	81.13%	94.44%	100%	83.33%
Recall	NE	46.67%	47.37%	58.69%	52.94%	33.33%	57.14%
	IC	93.33%	94.74%	93.48%	100%	100%	95.24%
F-measure	NE	63.64%	64.28%	70.13%	69.23%	50%	64.86%
	IC	87.5%	81.82%	86.87%	97.14%	100%	88.89%

Analyzing *precision*, we realize that, in most cases, the NE-detection method provided better results than ours. Since precision mainly depends on the number of false positives, this states that our method tends to select too much terms as sensitive. A reason for this is the fact that our method detected, in some cases, syntactically complex NPs as sensitive terms, even though they may refer to general (non-revealing) concepts. Complex NPs are those composed by several words and using complex syntactic constructions that, when queried in a Web Search Engine, tend to provide a relatively low page count, giving the impression of a high IC. The fact that the page count depends on the lexico-syntactical construction of queried terms is caused by the strict terminological matching implemented by Web search engines in which our method relies. On the other hand, since the NE-based method obtained a perfect precision in most cases, this suggest that most (but no all) NEs are sensitive. A worth-noting case is *Jordi Mollà*, in which the NE-based method provided a lower precision than ours. In this case, the NE-detection package tagged general entities like *United States* or *Spain* (since they represent a *location*) which were not considered as sensitive by human experts due to their generality. Our method, on the contrary, relying on the low IC these term provide, behave inversely, achieving a higher precision and retaining more information.

Recall represents a more important dimension in the context of document sanitization, since a low recall implies that a number of terms considered as sensitive will appear in the sanitized document. In this case, recall figures for NE-based methods are significantly lower than ours. In fact, when our method was able to stay in the 95-100% range in most cases, the NE-based method resulted in recall values around 50%. On the one hand, this shows that not only NEs appearing in text are sensitive, but also NPs (*e.g. film and fashion icon of the twentieth century*, referred to *Audrey Hepburn*) referring to concrete concepts. On the other hand, NE-based methods are limited by the scope of the trained classifiers. The fact that only certain NE types (*locations*, *persons* and *organizations*, in this case) are detected, resulted in the omission of an amount sensitive NEs like *movie titles.* Moreover, the worst results were obtained for an Spanish actor (*Javier Bardem*) due to the presence of Spanish localisms and Spanish movie titles, which are difficult to detect for an English-trained classifier. It is worth mentioning that several of these omissions were highly revealing, resulting in an instant disclosure (*e.g. Rocky* for *Sylvester Stallone* or *Governator* for *Arnold Schwarzenegger*). 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 be not 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 [19]. Moreover, on the contrary to tagged corpora, the Web offers up-to-date results and covers almost any possible domain [19].

As a result of the significant differences between methods' recalls (*i.e.,* disclosure risk), when comparing them according to their global accuracy (*i.e., F-measure*), our method surpass NE-based ones in all cases.

5 Conclusions and future work

In this paper, an automatic method to detect sensitive information in text documents is presented. The method's generality is given by the theoretical foundations of the Information Theory and a corpus as general/global as the Web. As a result, it can be applied to heterogeneous textual data (and not only NEs [5, 11]) in a domain-independent fashion.

As future work, we plan to tackle the limitations observed in the IC calculus regarding the too strict query matching applied by Web search engines. In this case, different lexico-syntactical forms of the same terms can be queried and page count results can be aggregated to obtain a more general (and accurate) estimation of their informativeness.

Moreover, it is worth noting that even though most sanitization models propose removing those terms that are detected as potentially sensitive [6, 14, 13], this is not the most desirable strategy. Since the purpose of document sanitization is to provide a privacy-preserved but still useful version of the input document to the audience, a systematic removal of sensitive terms may hamper the document's utility. In fact, 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 of documents [27, 28]. To tackle this problem, recent methods [5, 11] propose replacing sensitive information by generalized versions (*e.g., "iPhone"* → *"cell phone"*) instead of removing it. 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 taxonomical structure of sanitized terms is needed. We plan to use general-purpose KBs to provide more accurate sanitizations, exploiting them from an information theoretic perspective.

Disclaimer and acknowledgments

Authors are solely responsible for the views expressed in this paper, which do not necessarily reflect the position of UNESCO nor commit that organization. This work was partly supported by the 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 and Audit Transparency Voting Process IPT-430000-2010-31), by the Spanish Ministry of Industry, Commerce and Tourism (through projects eVerification2 TSI-020100-2011-39 and SeCloud TSI-020302-2010-153) and by the Government of Catalonia (under grant 2009 SGR 1135).

References

1. U.S. Department of Justice: U.S. freedom of information act (FOIA) (2012)
2. Nat. Security Agency: Redacting with confidence: How to safely publish sanitized reports converted from word to pdf. Technical Report I333-015R-2005 (2005)
3. Department of Health and Human Services, Office of the Secretary: The health insurance portability and accountability act of 1996. Technical Report Federal Register 65 FR 82462 (2000)
4. Dorr, D.A., Phillips, W.F., Phansalkar, S., Sims, S.A., Hurdle, J.F.: Assessing the difficulty and time cost of de-identification in clinical narratives. Methods of Information in Medicine **45**(3) (2006) 246–252
5. Cumby, C., Ghan, R.: A machine learning based system for semi-automatically redacting documents. In: Proceedings of the 23rd Innovative Applications of Artificial Intelligence Conference. (2011) 1628–1635
6. Chakaravarthy, V.T., Gupta, H., Roy, P., Mohania, M.: Efficient techniques for document sanitization. In: Proceedings of the ACM Conference on Information and Knowledge Management – CIKM'08. (2008) 843–852

7. U.S. Department of Energy: Department of energy researches use of advanced computing for document declassification (2012)
8. DARPA: New technologies to support declassification. Request for Information (RFI) Defense Advanced Research Projects Agency. DARPA-SN-10-73 (2010)
9. Meystre, S.M., Friedlin, F.J., South, B.R., Shen, S., Samore, M.H.: Automatic de-identification of textual documents in the electronic health record: a review of recent research. BMC Medical Research Methodology **10**(70) (2010)
10. National Security Agency: Redaction of pdf files using Adobe Acrobat Professional X (2011)
11. Abril, D., Navarro-Arribas, G., Torra, V.: On the declassification of confidential documents. In: Proceedings of Modeling Decisions for Artificial Intelligence – MDAI'11. (2011) 235–246
12. Sweeney, L.: Replacing personally-identifying information in medical records, the scrub system. In: Proceedings of the 1996 American Medical Informatics Association Annual Symposium. (1996) 333–337
13. Tveit, A., Edsberg, O., Rost, T.B., Faxvaag, A., Nytro, O., Nordgard, M.T., Ranang, M.T., Grimsmo, A.: Anonymization of general practioner medical records. In: Proceedings of the second HelsIT Conference. (2004)
14. Douglass, M.M., Cliffford, G.D., Reisner, A., Long, W.J., Moody, G.B., Mark, R.G.: De-identification algorithm for free-text nursing notes. In: Proceedings of Computers in Cardiology'05. (2005) 331–334
15. Finkel, J.R., Grenager, T., Manning, C.: Incorporating non-local information into information extraction systems by gibbs sampling. In: Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics. (2005) 363–370
16. Sánchez, D., Isern, D.: Automatic extraction of acronym definitions from the web. Applied Intelligence **34**(2) (2011) 311–327
17. Sánchez, D., Isern, D., Millan, M.: Content annotation for the semantic web: an automatic web-based approach. Knowledge and Information Systems **27**(3) (2011) 393–418
18. Resnik, P.: Using information content to evalutate semantic similarity in a taxonomy. In: Proceedings of 14th International Joint Conference on Artificial Intelligence. (1995) 448–453
19. Sánchez, D., Batet, M., Valls, A., Gibert, K.: Ontology-driven web-based semantic similarity. Journal of Intelligent Information Systems **35**(3) (2010) 383–413
20. Sánchez, D., Batet, M., Isern, D.: Ontology-based information content computation. Knowledge-based Systems **24**(2) (2011) 297–303
21. Cilibrasi, R.L., Vitanyi, P.M.B.: The google similarity distance. IEEE Transactions on Knowledge and Data Engineering **19**(3) (2006) 370–383
22. Turney, P.D.: Mining the web for synonyms: Pmi-ir versus lsa on toefl. In: Proceedings of the 12th European Conference on Machine Learning. (2001) 491–502
23. Sánchez, D.: A methodology to learn ontological attributes from the web. Data and Knowledge Engineering **69**(6) (2010) 573–597
24. M. J. Cafarella, O.E.: A search engine for natural language applications. In: Proceedings of the 14th international conference on WWW. (2005) 442–452
25. Open Directory Project: ODP (2012)
26. Apache Software Foundation: OpenNLP (2012)
27. Martínez, S., Sánchez, D., Valls, A., Batet, M.: Privacy protection of textual attributes through a semantic-based masking method. Information Fusion **13**(4) (2012) 304–314
28. Martínez, S., Sánchez, D., Valls, A.: Semantic adaptive microaggregation of categorical microdata. Computers and Security **31**(5) (2012) 653–672