Document Sanitization: Measuring Search Engine Information Loss and Risk of Disclosure for the Wikileaks cables

David F. Nettleton[1,2], Daniel Abril[1,3]

[1]IIIA-CSIC Artificial Intelligence Research Institute - Spanish National Research Council,
[2]Universitat Pompeu Fabra, [3]Universitat Autonòma de Barcelona
{dnettleton, dabril}@iiia.csic.es

Abstract. In this paper we evaluate the effect of a document sanitization process on a set of information retrieval metrics, in order to measure information loss and risk of disclosure. As an example document set, we use a subset of the Wikileaks Cables, made up of documents relating to five key news items which were revealed by the cables. In order to sanitize the documents we have developed a semi-automatic anonymization process following the guidelines of Executive Order 13526 (2009) of the US Administration, by (i) identifying and anonymizing specific person names and data, and (ii) concept generalization based on WordNet categories, in order to identify words categorized as classified. Finally, we manually revise the text from a contextual point of view to eliminate complete sentences, paragraphs and sections, where necessary. We show that a significant sanitization can be applied, while maintaining the relevance of the documents to the queries corresponding to the five key news items.

Keywords: document sanitization, privacy, information retrieval, search engine, queries, information loss, disclosure risk, Wikileaks cables.

1 Introduction

The recent case of the publishing of more than 250,000 US Embassy Cables by Wikileaks has caused a great debate between those who uphold the freedom of information and those who defend the right to withhold information. Key documents which relate to national and international events are withheld from the public domain because they are designed as "classified" by official security criteria. In the United States, the three main classifications are: Top Secret, Secret and Confidential. Classification categories are assigned by evaluating the presence of information in a document whose unauthorized disclosure could reasonably be expected to cause identifiable or describable damage to the national security [1]. This type of information includes military plans, weapons systems, operations, intelligence activities, cryptology, foreign relations, storage of nuclear materials, weapons of mass destruction. On the other

hand, some of this information is often directly related to national and international events which affect millions of people in the world, who in a democracy may wish to know the decision making processes of their elected representatives, ensuring a transparent and open government. One problem with Wikileaks' publishing of the US Embassy Cables [2] is that they were published in a "raw" state, without any sanitization. That means that they included information (emails, telephone numbers, names of individuals and certain topics) whose absence may not have significantly impaired the informative value of the documents with respect to what are now considered the most important revelations of the Cables.

The main goal of this research is to find new mechanisms to evaluate the information loss and the disclosure risk of a set of sanitized documents. To do so, we have implemented a semi-automatic method to sanitize the Wikileaks documents and then we have evaluated them.

The structure of the paper is as follows: in Section 2 we briefly review the state of the art and related work; in Section 3 we describe the documents and queries used and the sanitization process; in Section 4 we describe the information loss metrics used and the search engine which we programmed ourselves in Java; Section 5 details the empirical results for information loss and risk of disclosure; finally, Section 6 concludes the paper.

2 Related Work

Document sanitization is a field which does not have such an extensive literature as that of the anonymization of structured and semi-structured data in general. However, it is a field of crucial importance with respect to online content publishing. Recent works include [3, 4, 5, 6, 7, 8, 9, 10, 11], which we will now briefly comment.

Chakaravarthy et al. in [3] present the ERASE (Efficient RedAction for Securing Entities) system for the automatic sanitization of unstructured text documents. The system prevents disclosure of protected entities by removing certain terms from the document, which are selected in such a way that no protected entity can be inferred as being mentioned in the document by matching the remaining terms with the entity database. Each entity in the database is associated with a set of terms related to the entity; this set is termed the context of the entity.

Cumby et al. in [4] present a privacy framework for protecting sensitive information in text data, while preserving known utility information. The authors consider the detection of a sensitive concept as a multiclass classification problem, inspired in feature selection techniques, and present several algorithms that allow varying levels of sanitization. They define a set D of documents, where each $d \in D$ can be associated with a sensitive category $s \in$

S, and with a finite subset of non-sensitive utility categories $U_d \subset U$. They define a privacy level similar to k-anonymity [5], called k-confusability, in terms of the document classes.

Hong et al. in [6] present a heuristic data sanitization approach based on 'term frequency' and 'inverse document frequency' (commonly used in the text mining field to evaluate how relevant a word in a corpus is to a document). In [7], Samelin et al. present an RSS (redactable signature scheme) for ordered linear documents which allows for the separate redaction of content and structure. Chow et al., in [8] present a patent for a document sanitization method, which determines the privacy risk for a term by determining a confidence measure $c_s(t_1)$ for a term t_1 in the modified version of the document relative to sensitive topics s. In the context of the sanitization of textual health data, [9] presents an automated de-identification system for free-text medical records, such as nursing notes, discharge summaries, X-ray reports, and so on.

Privacy preserving text mining: In [10], Abril et al. consider the problem of protecting classified documents by substituting keywords by more general ontological terms. We observe that the original "document" and the protected "document" consist of lists of extracted keywords, and not the complete text itself. In [11], the protection of complete documents is considered (not just lists of keywords). The anonymization process works by recognizing specific entities (names of persons, places, and organizations) and substituting them with generalizations, swapping them or adding noise.

In the present work, the named entity recognition step is similar to [11], however, we add a second step of classified word detection and at the end of the process a human has to recognize clusters of detected entities and s/he must decide whether or not the sentence or paragraph will be deleted.

3 Documents/Queries used and Sanitization Process

In this Section we explain how we have selected the document set used, the queries and the sanitization process.

3.1 Documents and Queries - Information Loss and Risk of Disclosure

We have used the online Wikileaks Cable repository [2] as the source for the informational and risk documents. To obtain a set of documents, we selected five queries derived from the top ten revelations published by Yahoo! News [12], as is shown in Table 1. Then we searched using these queries as keywords on *www.cablegatesearch.net* [2] to find the corresponding cables, thus obtaining a set of documents for each query. We observe that a sixth document set, *i6*, was randomly chosen from [2] for benchmarking purposes. The same five queries (Table 1) were used to test information loss (utility) in the

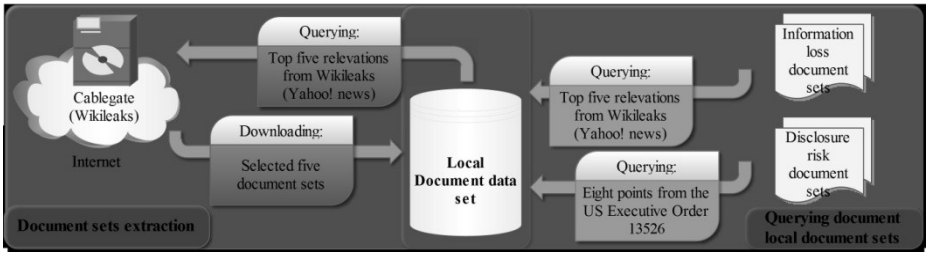

Fig. 1. Scheme for document extraction and querying

empirical results section. In Fig. 1 we see a schematic representation of the process.

With respect to the risk, we extracted 30 seed terms from the eight risk points defined in Section 1.4 of the US Executive Order 13526 [1], as is shown in Table 2. Hence, we defined eight different queries, one for each risk point, which are designated as $rq_1 \rightarrow rq_8$, with corresponding to document sets r1 $\rightarrow$ r8. We defined a ninth query, rq_9, composed of all the terms from queries 1 to 8, whose corresponding document set is r9.

3.2 Sanitization Process

We have implemented a simple supervised sanitization method based on entity recognition and pattern-matching techniques in order to detect entities and sensitive words in the text, which is summarized in Fig. 2.

Table 1. Queries and documents used to test Information Loss

Id. Query	Keywords (utility queries)	TC, CH[1]	ID[2]	Top five news item revelations (Yahoo!)[12]
uq_1	{ saudi, qatar, jordan, UAE, concern, iran, nuclear, program }	35, 10	il1	"Middle Eastern nations are more concerned about Iran's nuclear program than they've publicly admitted".
uq_2	{ china, korea, reunify, business, united, states}	3,3	il2	"U.S. ambassador to Seoul said that the right business deals might get China to acquiesce to a reunified Korea, if the newly unified power were allied with the United States".
uq_3	{ guantanamo, incentives, countries, detainees }	12,10	il3	"The Obama administration offered incentives to try to get other countries to take Guantanamo detainees, as part of its plan to progressively close down the prison".
uq_4	{diplomats, information, foreign, counterparts }	6,6	il4	"Secretary of State Hillary Clinton ordered diplomats to assemble information on their foreign counterparts".
$uq_{5\text{-}1}$	{ putin, berlusconi, relations }	97,10	il5	"Russian Premier Vladimir Putin and Italian Premier Silvio Berlusconi have more intimate relations than was previously known".
$uq_{5\text{-}2}$	{ russia, italy, relations }			
-	-	10,10	il6[3]	-

1Total Cables, Cables chosen; 2 Informational document sets; 3 represents a set of randomly chosen documents to be used as a benchmark

Table 2. Queries used to test Risk of Disclosure

Id. Query	Keywords (risk queries)	ID[1]	Classification categories, a→h, see [1]
rq_1	{military, plan, weapon, systems}	r1	(a)
rq_2	{intelligence, covert, action, sources}	r2	(b)
rq_3	{cryptology, cryptogram, encrypt}	r3	(c)
rq_4	{sources, confidential, foreign, relations, activity}	r4	(d)
rq_5	{science, scientific, technology, economy, national, security}	r5	(e)
rq_6	{safeguard, nuclear, material, facility}	r6	(f)
rq_7	{protection, service, national, security}	r7	(g)
rq_8	{develop, production, use, weapon, mass, destruction}	r8	(h)
rq_9	All terms from rq_1 to rq_8.	r9	-

1 disclosure risk document set

This process consists of two steps: (i) the anonymization of names and personal information of individuals and (ii) the elimination of blocks of "risk text", following the guidelines of [1].

(i) Anonymization of names and personal information of individuals We have used the 'Pingar' online application [13] and 'api' to process the text, which anonymizes the following: people, organizations, addresses, emails, ages, phone numbers, URLs, dates, times, money and amounts. This process simply substitutes the information with {Pers1, Pers2, …}, {Loc1, Loc2, …}, {Date1, Date2, …} and so on. We also observe that the names of countries (Iran, United States, Russia, Italy, …) and places (London, Abu Dhabi, Guantanamo, …) are unchanged in this process.

(ii) Elimination of blocks of "risk text". With reference to Table 2, risk text blocks are identified by the presence of one or more of the concepts defined in points (a) to (h) of [1]. The concepts are represented by an initial list of 30 "risk" keywords.

For each of these keywords, we then used WordNet ontology database [14] to find the corresponding synonyms and hyponyms taking into account the specific or closer sense to the original term. We note that this word sense disambiguation was performed manually. By hyponym we mean the lower part of the ontology tree starting from the given keyword. For example, "weapon" would give the following: "knife, sling, bow, arrow, rock, stick, missile, cannon, gun, bomb, gas, nuclear, biological, …". This produced a list with a total of 655 terms (original + synonyms + hyponyms).

Then we processed the documents generating an output file in which all the keywords are signaled thus "****Keyword****", and which also indicates the

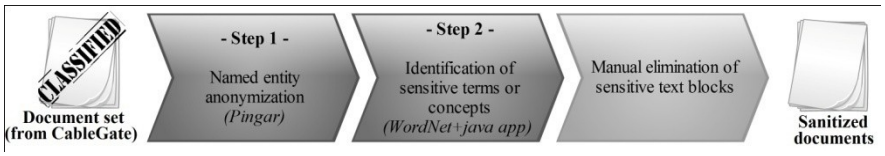

Fig. 2. Scheme for document sanitization

relative distance of each "risk" keyword found from the start of the file. We cluster these distances for each file and use the information to signal documents with text areas which have a high density of risk keywords, which would be candidates to be eliminated from the file. We note that we applied a *stemming process* (using the Porter Stemming algorithm version 3 [15], implemented in Java) to the keyword list and the words in the documents in order to match as many possible variants as possible of the root term. Finally, we manually revised the labeled files, using the clustered distance information for support, and deleted the paragraphs identified as having the highest clustering of "risk terms".

4 Search Engine, Information Loss and Risk Metrics

In this Section we describe the information loss and risk metrics, and the Vectorial model search engine. We note that the same metrics are used to measure information loss and disclosure risk. However, as previously mentioned, these two metrics require different sets of queries (utility and risk queries) to perform the evaluation and give a different interpretation. The utility queries consist of terms about the general topic of each document set (see Table 1) and the risk queries consist of terms that define sensitive concepts (see Table 2).

4.1 Information Loss and Risk of Disclosure Metrics

We have used as a starting point a set of typical information retrieval metrics, which are listed in Table 3. Precision is considered as the percentage of retrieved documents above the relevance threshold that are relevant to the informational query. Recall, on the other hand, is considered as the percentage of retrieved documents above the relevance threshold that are defined as truly relevant. The formulas are defined in terms of the following sets of documents: $q(RT_D)$, 'true_relevant_documents', is the set for a given query. For information loss, this will be the document set retrieved from *Cablegate-search*[2]; for risk of disclosure, it will be the unchanged document set retrieved by the corresponding risk query by the Vectorial search engine. On the other hand, $q(RV_D)$, 'retrieved_documents' is the set returned by the search engine in reply to a given query which are above the relevance threshold;

Table 3. Information Retrieval Metrics

Metric	Formula	
Precision	$P=\frac{\lvert\{relevant\_docs\}\cap\{retrieved\_docs\}\rvert}{\lvert\{retrieved\_docs\}\rvert}$	(1)
Recall	$R=\frac{\lvert\{relevant\_docs\}\cap\{retrieved\_docs\}\rvert}{\lvert\{true\_relevant\_docs\}\rvert}$	(2)
F-measure	$F=2\cdot\frac{precision\cdot recall}{precision+recall}$	(3)
Coverage	$C=\frac{\lvert\{true\_relevant\_docs\_returned\}\rvert}{\lvert\{true\_relevant\_docs\}\rvert}$	(4)
Novelty	$N=\frac{\lvert\{false\_relevant\_docs\}\rvert}{\lvert\{total\_relevant\_docs\}\rvert+\lvert\{false\_relevant\_docs\}\rvert}$	(5)

*See [16] for more details of these metrics.

and $q(RE_D)$, 'relevant_documents', are the documents above the relevance threshold which are members of $q(RT_D)$.

The *F-measure* (or balanced F-score) combines precision and recall and mathematically represents the harmonic mean of the two values. For the novelty metric and coverage metrics, we define the following sets of documents: $q(RR_D)$, 'true_relevant_docs_returned' are the documents in 'true_relevant_docs' which are returned by the search engine in any position (above or below the threshold); Finally, $q(RF_D)$, 'false_relevant_docs', are documents not members of 'true_relevant_docs' but which are returned above the relevance threshold. For our selected document corpus, we interpret novelty as undesirable with respect to the quality of the results, because we assume that we have correctly identified the set of all true relevant documents.

As well as the four metrics listed in Table 3, we also consider four other measures: (i) average relevance of the documents whose relevance is above the relevance threshold; (ii) the total number of documents returned by the query whose relevance is greater than zero; (iii) the number of random documents which are members of the set of relevant documents for a given query; (iv) NMI (Normalized Mutual Information), we use an NMI type metric [17] for counting document assignments to query document sets before and after sanitization. That is, we compare the results of the document assignments to query sets by identifying the documents in each query document set before sanitization, and the documents which are in the same corresponding query document set after sanitization.

Quantification of information loss and risk: in order to obtain a single resulting value, we have studied all the parameters presented and defined a formula in terms of the factors which showed the highest correlation between the original and sanitized document metrics: F = F-measure, C = coverage, N = novelty, TR = total number of documents returned, PR = percentage of ran-

dom documents in the relevant document set, and the NMI value. Hence IL, the information loss is calculated as:

$$\mathrm{IL} = \frac{(2 \times \mathrm{F}) + \mathrm{C} - \mathrm{N} + \mathrm{TR} - \mathrm{PR} - (2 \times \mathrm{NMI})}{8} \tag{6}$$

We observe that of the six terms in the formula, F and NMI are given a relative weight of 25%, and the other four terms are given a relative weight of 12.5%. The weighting was assigned by evaluating the relative correlations of the values before and after document sanitization for each factor, for information loss and risk of disclosure. For the risk of disclosure, RD, we use the same formula and terms, however the interpretation is different: for IL a negative result represents a reduction in information, and for RD a negative result represents a reduction in risk.

Relevance cut-off value for informational document sets. In order to apply the same criteria to all the search results, after studying the distributions in general of the relevance of the different queries, we chose a relevance of 0.0422 as the cut-off. That is, we define an inflexion point between the relevant documents (relevance greater or equal to 0.0422) and non-relevant documents (relevance less than 0.0422). See Annex Table 8 and Fig. 3 for a graphic example.

Relevance cut-off value for risk document sets. After studying the distributions of the relevance for each risk document set returned by the search engine, we assigned the relevance threshold of 0.010 for all the results sets, with the exception of result sets *r9*, *r1* and *r2* which were assigned a threshold of 0.020. The metric calculations then followed the same process as for the informational document sets.

4.2 Search Engine

We have implemented our own search engine in Java, with the following main characteristics: an inverted index to store the relation between terms and documents and a hash-table to efficiently store the terms (vocabulary); elimination of stop-words and stemming; calculation of term frequency, inverted document frequency, root of the sum of weights for of the terms in each document; implementation of the Vectorial Model formula to calculate the similarity of a set of terms (query) with respect to the corpus of documents. Refer to [16] for a complete description of the Vectorial model and the formula used. We observe that the queries are by default ‘OR’. That is, if we formulate the query "term1 term2 term3", as search engines do by default, an OR is made of the terms and the documents are returned which contain at least one of the three given terms, complying with "term1 OR term2 OR term3".

Table 4. Information Loss: percentage (%) differences of NMI metric for original and sanitized document corpuses (steps 1+2)

	uq_1	uq_2	uq_3	uq_4	$uq_{5\text{-}1}$	$uq_{5\text{-}2}$
Step 1	0.00	0.00	0.00	0.00	100.00	0.00
Step 2	11.00	0.00	14.00	50.00	100.00	0.00

Table 5. Information Loss: percentage (%) differences of statistics for original (Annex Table 9) and sanitized (Annex Table 11) document corpuses (steps 1+2)

	P	R	F	C	N	AR	TR	PR	IL
uq_1	-1.56	-12.50	-0.08	0.00	0.00	-38.15	-15.38	0.00	-6.625
uq_2	-40.00	0.00	-0.25	0.00	40.00	-0.38	-4.76	20.00	-14.37
uq_3	0.00	-14.29	-0.09	0.00	0.00	3.77	-12.50	0.00	-7.375
uq_4	-62.50	-75.00	-0.70	0.00	33.33	9.80	-10.81	25.00	-38.62
$uq_{5\text{-}1}$	-100.00	-100.00	-1.00	-100.00	-100.00	-100.00	-4.55	0.00	-75.62
$uq_{5\text{-}2}$	-11.11	0.00	-0.05	0.00	38.46	-5.03	0.00	0.00	-13.75

Legend: P=precision, R=recall, F=F measure, C=coverage, N=novelty, AR=Average relevance for documents above threshold, TR= total docs. returned, PR=percentage of random docs in relevant doc set, IL=percentage information loss calculated using formula 6

5 Empirical Results

In this section we present the results for information loss and risk of disclosure, comparing the metrics calculated for the original documents with those calculated for the sanitized documents (step 1 + step 2, see Section 3.2).

5.1 Information Loss

In Table 4 we see the NMI metric applied to the original and sanitized document query sets. We see only a small reduction in correspondence for the majority of query document sets, except for uq_4 and $uq_{5\text{-}1}$, however, the latter is due to the loss of the named query terms in the documents (Putin and Berlusconi were masked as named entities in step 1 of the sanitization process).

In the case of uq_4, a value of 50% for step 2 means that 50% of the relevant documents from the original document set returned by the search engine, are to be found in the relevant documents from the sanitized document set returned by the search engine.

Table 5 shows the percentage change for each metric value and informational document set, of the original documents (see Annex Table 9) and the sanitized documents processed by steps 1 and 2 (See Annex Table 11). We observe that the indicators used in the information loss formula (6) are highlighted in grey. The information loss calculated using *formula 6* is shown in

Table 6. Risk of Disclosure: percentage (%) differences of NMI metric for original and sanitized document corpuses (steps1+2)

rq_1	rq_2	rq_3	rq_4	rq_5	rq_6	rq_7	rq_8	rq_9
60.00	67.00	-	36.00	25.00	56.00	63.00	70.00	58.00

Table 7. Risk of Disclosure: percentage (%) differences of statistics for original (Annex Table 12) and sanitized (Annex Table 13) document corpuses (steps 1+2)

	P	R	F	C	N	AR	TR	PR	RD
rq_1	-66.67	-60.00	-0.64	-16.67	40.00	-26.94	-44.44	30.0	-47.37
rq_2	-66.67	-66.67	-0.67	-33.33	40.00	27.07	-48.39	16.7	-50.75
rq_3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0	-
rq_4	-18.18	-35.71	-0.28	-7.14	15.38	17.80	-4.17	1.96	-19.5
rq_5	-57.14	-25.00	-0.45	-12.50	50.00	11.74	-18.60	8.90	-28.87
rq_6	-60.00	-55.56	-0.58	-22.22	40.00	8.07	-55.26	17.8	-45.37
rq_7	-71.43	-50.00	-0.64	-12.50	55.56	-0.49	-33.33	35.7	-49.00
rq_8	-50.00	-70.00	-0.63	-50.00	23.08	-39.31	-29.41	23.3	-48.87
rq_9	-54.55	-58.33	-0.57	0.00	35.29	-14.29	-10.20	9.9	-35.62

Legend: P=precision, R=recall, F=F measure, C=coverage, N=novelty, AR=Average relevance for documents above threshold, TR= total docs. returned, %PR=percentage of random docs in relevant doc set, RD=percentage risk decrease calculated using formula 6

the rightmost column (IL), the average giving a value of 26.1% including query $uq_{5\text{-}1}$, and a value of 16.1% excluding query $uq_{5\text{-}1}$.

With reference to query $uq_{5\text{-}1}$, the names of two persons, "berlusconi" and "putin", were substituted. As they were essential for the successful retrieval by this query of the corresponding documents, this resulted in a total loss of retrieval. In Table 5 we also observe that the F measure (which is a ratio of precision and recall) has reduced for uq_2 and uq_4, and the novelty (N) and percentage of random documents (PR) have increased. Novelty is considered a negative aspect, given that we interpret it as the entry of irrelevant documents into the set of relevant documents (above the threshold).

In conclusion, step 1 (*anonymization of names and personal information of individuals*) has little or no effect on the success of the informational queries, except those which contain specific names of people. However, this is an important required process because it is necessary to preserve the confidentiality of the individuals who appear in these documents. On the other hand, step 2 (*elimination of 'risk text'*) inevitably had a higher impact, given that we are eliminating blocks of text from the documents. Moreover, from the results of Table 5, we observe that the information loss is query dependent, the F and TR indicators being the most consistent. By manual inspection of the docu-

ments, we can conclude in general that a worse value is due to the loss of key textual information relevant to the query.

5.2 Risk of Disclosure

In Table 6 we see the NMI metric applied to the original and sanitized document query sets. We see a significant reduction in the correspondence, which contrasts with the results for the same metric applied to the information loss query document sets. Table 7 shows the percentage change for each of the metrics we described in Section 3.1, for each of the nine 'risk' queries, for the original documents (Annex Table 12) and the sanitized documents of step 2 (Annex Table 13). In general, we observe a significantly greater percentage change in comparison to the information loss results of Table 5.

We observe that query rq_3 did not retrieve any documents, although we included it in the results as it corresponds to point (c) of [1]. The risk decrease calculated using *formula 6* is shown in the rightmost column (RD), the average value being -47.26%.

However, the calculated risk of disclosure and information loss (formula 6) is considered as a guide rather than an absolute value. For example, with reference to Table 7, the user could visually inspect the most highly ranked documents of the group (rq_4) showing the least reduction in RD (19%), and those showing the highest information loss (uq_1 and $uq_{5\text{-}1}$) of Table 5).

By observing the relative ranking of the documents returned by the queries, we saw that some documents with 'risk' terms actually went up the ranking. After inspecting the corresponding documents, we found that this was due to the presence of terms such as 'nuclear', but in a peaceful (energy) context, and 'war' with reference to minor conflicts such as the Balkans, which had no relation to US national security. However, we re-checked our editing of the documents corresponding to query rq_5, given the increased presence of these documents in the highest ranked positions. We confirmed that the sanitization was consistent with the other document groups.

6 Conclusions

In this paper we have proposed a novel approach and methodology for the evaluation of information loss and disclosure risk for a data set of sanitized documents. In order to evaluate these two values we have implemented a vectorial model search engine and we also have defined a formula to evaluate the information loss and disclosure risk by means of querying both document sets.

The results show a relatively low information loss (16% excluding query $uq_{5\text{-}1}$) for the utility queries (uq_1 to uq_5), whereas an average reduction of 47% was found for the risk queries (ur_1 to ur_9). As future work, we propose a

greater automation of step 2 by using a program to demarcate the "risk" text and via a user interface, asking the user if s/he wishes to eliminate it, or not. Also we could use an optimization process to learn the weighting for each of the terms in formula 6, and benchmark different sanitization methods.

ACKNOWLEDGEMENTS

This research is partially supported by the Spanish MEC projects CONSOLIDER INGENIO 2010 CSD2007-00004 and eAEGIS TSI2007-65406-C03-02. The work contributed by the second author was carried out as part of the Computer Science Ph.D. program of the Universitat Autònoma de Barcelona (UAB).

References

1. Executive Order 13526 (2009) of the US Administration - Classified National Security Information, Section 1.4, points (a) to (h). *http://www.whitehouse.gov/the-press-office/executive-order-classified-national-security-information*
2. Wikileaks Cable repository, www.cablegatesearch.net
3. Chakaravarthy, V.T., Gupta, H., Roy, P., Mohania, M.K. Efficient Techniques for Document Sanitization. CIKM'08, October 26–30, 2008, Napa Valley, California, USA. 0
4. Cumby, C., Ghani, R. A Machine Learning Based System for Semi-Automatically Redacting Documents. Proc. IAAI 2011.
5. Sweeney, L. k-anonymity: a model for protecting privacy. 2002. *Int. Journal of Uncertainty, Fuzziness and Knowledge-Based Systems (IJUFKS)*. Vol. 10, Issue: 5, pp. 557-570.
6. Hong, T.P., Lin C.W., Yang, K.T., Wang, S.L. A Heuristic Data-Sanitization Approach Based on TF-IDF. LNCS 2011, Volume 6703/2011, 156-164.
7. Samelin, K, Pöhls, H.C., Bilzhause,A., Posegga,J., de Meer, H. Redactable Signatures for Independent Removal of Structure and Content. Proc. ISPEC 2012, LNCS 7232, pp.17–33.
8. Chow, R., Staddon, J.N., Oberst, I.S. Method and apparatus for facilitating document sanitization. US Patent Application Pub. No. US 2011/0107205 A1, date May 5, 2011.
9. Neamatullah, I., Douglass, M.M., Lehman L.H., Reisner, A., Villarroel, M., Long, W.J., Szolovits, P., Moody, G.B., Mark, R.G., Clifford, G.D., Automated de-identification of free-text medical records. BMC Medical Informatics and Decision Making 2008, 8:32
10. Abril, D., Navarro-Arribas, G., Torra, V. *Towards Semantic Microaggregation of Categorical Data for Confidential Documents.* MDAI 2010, LNAI Vol. 6408, p.266-276.
11. Abril, D., Navarro-Arribas, G., Torra, V. *On the declassification of confidential documents.* MDAI 2011, LNAI, Volume 6820, p.235-246.
12. Yahoo! News. Top 10 revelations from Wiki Leaks cables. *http://news.yahoo.com/blogs/lookout/top-10-revelations-wikileaks-cables.html*
13. Pingar – Entity Extraction Software. *http://www.pingar.com*
14. Miller, G.A., Beckwith, R., Fellbaum, C.D., Gross, D., Miller, K. 1990. WordNet: An online lexical database. Int. J. Lexicograph. 3, 4, pp. 235–244.
15. Porter, M.F. 1980, An algorithm for suffix stripping, Program, Vol. 14, no. 3, pp 130-137.
16. Baeza-Yates, R., Ribeiro-Neto, B. 2011. *Modern Information Retrieval: The Concepts and Technology behind Search*, 2nd Edition, ACM Press Books. ISBN: 0321416910.
17. Manning, C.D., Raghavan, P. and Schütze, H. 2008. Introduction to Information Retrieval, Cambridge University Press. 2008. ISBN: 0521865719.

Annexes

Table 8. Example search results

VECTOR MODEL SEARCH ENGINE		
Search terms: query $uq_{5\text{-}1}$		
Query "putin berlusconi relations"		
Rank	Doc id	Relevance
1	u5.6	0.262488
2	u5.1	0.210500
3	u5.2	0.107093
4	u5.3	0.098520
5	u5.4	0.087844
6	u3.7	0.076260
7	u5.8	0.052028
8	u5.10	0.022432
…	….	……
44	ur.9	0.000034

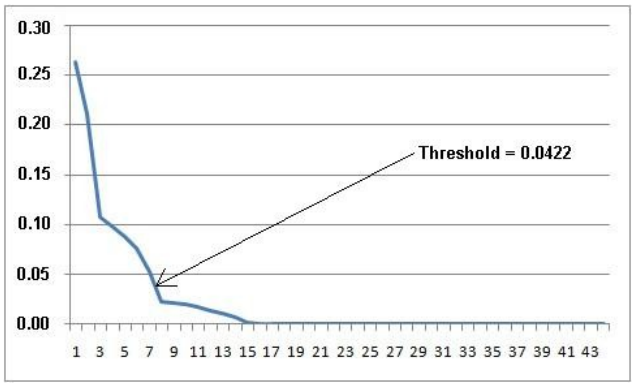

Fig. 3. Example distribution of relevance (x-axis) of ranked documents (y-axis) corresponding to the query of Table 8.

With reference to Table 8 and Fig. 3, the inflexion point of 0.0422 defines that documents ranked 1 to 7 are relevant and 8 to 43 are not relevant. For this example, and with reference to the definitions given in Table 3, the information loss metrics are calculated as follows: *(i) precision* = 6 / 7 = 0.8571. That is, there were 6 known relevant documents from a total of 7 above the relevance threshold; *(ii) recall* = 6 / 10 = 0.6. That is, six of the 10 known relevant documents were returned above the relevance threshold; *(iii) F-measure* = 2 × ((0.8571 × 0.6) / (0.8571 + 0.6)) = 0.7058, where the precision is 0.8571 and the recall is 0.6; *(iv) coverage* = 10 / 10 = 1.0, because all 10 known relevant documents were returned among the 44 results of the search engine; *(v) novelty* = 1 / (10 + 1) = 0.0909, where there are 10 known documents relevant to the query (Table 1) and in the list of relevant documents (relevance ≥ 0.0422), one of the documents (u3.7, ranked sixth) is not in the set of 10 known documents.

Table 9. Information Loss. Values of IR metrics for original file corpus

	P	R	F	C	N	AR	TR	PR
uq_1	0.8888	0.80	0.8421	1.00	0.0909	0.1768	39	0.0
uq_2	1.0000	1.00	1.0000	1.00	0.0000	0.1479	42	0.0
uq_3	1.0000	0.70	0.8235	1.00	0.0000	0.0770	32	0.0
uq_4	0.6667	0.67	0.6666	1.00	0.2500	0.0759	37	0.0
$uq_{5\text{-}1}$	0.8571	0.60	0.7058	1.00	0.0909	0.1278	44	0.0
$uq_{5\text{-}2}$	0.7500	0.60	0.6667	1.00	0.0833	0.2009	45	0.0

Legend: P=precision, R=recall, F=F measure, C=coverage, N=novelty, AR=Average relevance for documents above threshold, TR= total docs. returned, PR=percentage of random docs in relevant doc set

Table 10. Information Loss. Values of IR metrics for sanitized document corpus (step 1)

	P	R	F	C	N	AR	TR	PR
uq_1	0.9000	0.90	0.9000	1.00	0.1667	0.1409	39	0.0
uq_2	0.7500	1.00	0.8571	1.00	0.2500	0.1234	42	0.0
uq_3	1.0000	0.70	0.8235	1.00	0.0000	0.0826	32	0.0
uq_4	0.6667	0.67	0.6666	1.00	0.2500	0.0778	37	0.0
$uq_{5\text{-}1}$	0.0000	0.00	0.0000	1.00	0.0000	0.1380	44	0.0
$uq_{5\text{-}2}$	0.7500	0.60	0.6667	1.00	0.1667	0.2251	45	0.0

Legend: P=precision, R=recall, F=F measure, C=coverage, N=novelty, AR=Average relevance for documents above threshold, TR= total docs. returned, PR=percentage of random docs in relevant doc set

Table 11. Information Loss. Values of IR metrics for sanitized document corpus (step 2)

	P	R	F	C	N	AR	TR	PR
uq_1	0.8750	0.70	0.7777	1.00	0.0909	0.1093	33	0.0
uq_2	0.6000	1.00	0.7500	1.00	0.4000	0.1473	40	20.0
uq_3	1.0000	0.60	0.7500	1.00	0.0000	0.0799	28	0.0
uq_4	0.2500	0.17	0.2000	1.00	0.3333	0.0834	33	25.0
$uq_{5\text{-}1}$	0.0000	0.00	0.0000	0.00	0.0000	0.0000	42	0.0
$uq_{5\text{-}2}$	0.6667	0.60	0.6315	1.00	0.2307	0.1908	45	0.0

Legend: P=precision, R=recall, F=F measure, C=coverage, N=novelty, AR=Average relevance for documents above threshold, TR= total docs. returned, PR=percentage of random docs in relevant doc set

Table 12. Risk of Disclosure. Values of IR metrics for original file corpus

	P	R	F	C	N	AR	TR	PR
rq_1	1.00	1.00	1.00	1.00	0.00	0.0443	36	20.0
rq_2	1.00	1.00	1.00	1.00	0.00	0.0257	31	0.00
rq_3	0.00	0.00	0.00	0.00	0.00	0.0000	0	0.00
rq_4	1.00	1.00	1.00	1.00	0.00	0.0198	48	7.14
rq_5	1.00	1.00	1.00	1.00	0.00	0.0223	43	12.5
rq_6	1.00	1.00	1.00	1.00	0.00	0.0201	38	22.2
rq_7	1.00	1.00	1.00	1.00	0.00	0.0206	45	0.00
rq_8	1.00	1.00	1.00	1.00	0.00	0.0336	17	10.0
rq_9	1.00	1.00	1.00	1.00	0.00	0.0324	49	8.30

Legend: P=precision, R=recall, F=F measure, C=coverage, N=novelty, AR=Average relevance for documents above threshold, TR= total docs. returned, PR=percentage of random docs in relevant doc set

Table 13. Risk of Disclosure. Values of IR metrics for sanitized document corpus (steps 1+2)

	P	R	F	C	N	AR	TR	PR
rq_1	0.33	0.40	0.3636	0.8333	0.4000	0.0324	20	50.0
rq_2	0.33	0.33	0.3333	0.6667	0.4000	0.0327	16	16.7
rq_3	0.00	0.00	0.0000	0.0000	0.0000	0.0000	0	0.0
rq_4	0.82	0.64	0.7200	0.9286	0.3803	0.0233	46	9.1
rq_5	0.43	0.75	0.5454	0.8750	0.5490	0.0249	35	21.4
rq_6	0.40	0.44	0.4210	0.7778	0.4000	0.0217	17	40.0
rq_7	0.29	0.50	0.3636	0.8750	0.5556	0.0205	30	35.7
rq_8	0.50	0.30	0.3750	0.5000	0.2308	0.0204	12	33.3
rq_9	0.45	0.42	0.4347	1.0000	0.3529	0.0278	44	18.2

Legend: P=precision, R=recall, F=F measure, C=coverage, N=novelty, AR=Average relevance for documents above threshold, TR= total docs. returned, PR=percentage of random docs in relevant doc set