An Information Retrieval Approach to Document Sanitization

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

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

Abstract. In this paper we use information retrieval metrics to evaluate the effect of a document sanitization process, measuring information loss and risk of disclosure. 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. It embodies two main steps: (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. For empirical tests, we use a subset of the Wikileaks Cables, made up of documents relating to five key news items which were revealed by the cables.

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

1 Introduction

The 28th of November 2010 marks the largest release of classified data, when WikiLeaks, a non-profit organization, published more than 250,000 United States diplomatic cables that had been sent to U.S. international relations department between December 1966 and February 2010, by 274 of its consulates, embassies, and diplomatic missions around the world. From this large set of published documents there were over 115,000 labeled as “confidential” or “secret” and the remaining ones are unclassified by the official security criteria. According to the United States government the documents are classified at 4 levels: “Top secret”, “Secret”, “Classified” and “Unclassified”. These 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, and 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. Therefore, releasing such amount of confidential data caused a great debate between those who uphold the freedom of information and those who defend the right to withhold information.

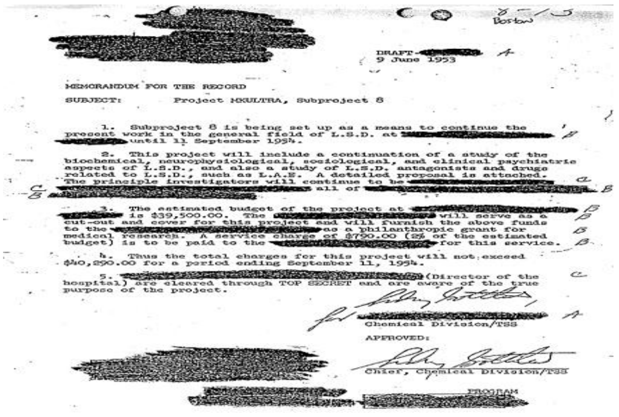
In the summer of 2010, WikiLeaks reached an agreement with some media partners from Europe and the United States to publish a set of cables in an edited form, removing the names of sources and other sensitive data. However, later on all the US Embassy cables [2] were published on the Internet fully unedited, in a "raw" state. That means that they included all kinds of confidential information such as 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 goal of this research is twofold. On the one hand, we have focused on new ways that could help to automate the concealment of confidential data. To do so, we have implemented a semi-automatic method to sanitize confidential unstructured documents, such as the released WikiLeaks documents. On the other hand, this research has also focused on finding new mechanisms to evaluate the information loss and the disclosure risk of a set of sanitized documents. We have proposed a technique relying on traditional information retrieval metrics which evaluates both the information loss and the risk of disclosure of a sanitized data set, by means of query comparisons.

This paper is organized as follows: the section 'Related Work' briefly reviews the state of the art and related work which is followed by the section 'Sanitization Method' which presents the sanitization method. Then, in the 'Information Loss and Risk Evaluation' section we describe the information loss and disclosure risk evaluation process. This is followed by the 'Experimental analysis' section which details the empirical results for information loss and risk of disclosure. Finally, in 'Conclusions' we summarize the paper and detail future lines of work.

2 Related Work

Document sanitization is the process of declassification or reduction of a documents classification level, by means of removing the sensitive information from a document. Figure 1 is an example of a US government document that has been manually sanitized prior to release. In recent years there have been many efforts to automate or help people to perform the anonymization process by saving time and getting more accurate results.



[redacted]

8-5-/3
Boston

DRAFT-[redacted] A
9 June 1953

MEMORANDUM FOR THE RECORD

SUBJECT: Project MKULTRA, Subproject 8

1. Subproject 8 is being set up as a means to continue the present work in the general field of L.S.D. at [redacted] B [redacted] until 11 September 1954.

2. This project will include a continuation of a study of the biochemical, neurophysiological, sociological, and clinical psychiatric aspects of L.S.D., and also a study of L.S.D. antagonists and drugs related to L.S.D., such as L.A.E. A detailed proposal is attached. The principle investigators will continue to be [redacted] a C [redacted] all of [redacted] B B [redacted]

3. The estimated budget of the project at [redacted] B [redacted] is $39,500.00. The [redacted] will serve as a B cut-out and cover for this project and will furnish the above funds to the [redacted] as a philanthropic grant for B medical research. A service charge of $790.00 (2% of the estimated budget) is to be paid to the [redacted] for this service. B

4. Thus the total charges for this project will not exceed $40,290.00 for a period ending September 11, 1954.

5. [redacted] (Director of the C hospital) are cleared through TOP SECRET and are aware of the true purpose of the project.

for [redacted] A
Chemical Division/TSS

APPROVED:

Chief, Chemical Division/TSS

[redacted]

PROGRAM

[redacted]

Fig. 1. Sanitization example (source: Wikipedia).

Document sanitization consists of two main tasks. The first one is the detection of sensitive data within the text and once the sensitive information is spotted the second task is performed, that consists in hiding the previously detected information, with the aim of minimizing the disclosure risk, while causing the least distortion to the document content. The first task is usually solved by Named Entity Recognition and Classification systems, which are a set of techniques developed by a subfield of Information Retrieval that intends to identify and classify atomic elements and entities which appear within a text. The second task has been studied and carried out in several ways; below we briefly describe some of them.

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 defined as the context of the entity.

Saygin et al. [4] propose a sanitization approach that first automatically detects sensitive named entities, such as person and organization names, dates, credit card numbers, etc. and then those named entities are perturbed and generalized to hide the sensitive information, i.e., enforcing k-anonymity [5] at individual term level.

Cumby et al. in [6] 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 ∈ D can be associated with a sensitive category s ∈ S, and with a finite subset of non-sensitive utility categories Ud ⊂ U. They define a privacy level similar to k-anonymity [5], called k- confusability, in terms of the document classes.

Hong et al. in [7] 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 [8], 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 [9] present a patent for a document sanitization method, which determines the privacy risk for a term by determining a confidence measure cs(*t1*) for a term *t1* in the modified version of the document relative to sensitive topics *s*. In the context of the sanitization of textual health data, [10] presents an automated de-identification system for free-text medical records, such as nursing notes, discharge summaries, X-ray reports, and so on.

Finally, Anandan et al. [11] focus on the protection of detected named entities by generalizing the sensitive words. This generalization relies on WordNet [12], an ontology that provides complete semantic relationship taxonomy between words. As this perturbation method relies on the semantic meaning of words it ensures less information loss in the sanitization process. Moreover, the authors present a measure, *t*-plausibility, to evaluate the quality of the sanitized documents from a privacy protection point of view. A generalized document holds the *t*-plausibility if at least *t* base documents can be generalized to a given sanitized document where a base document refers to one that has not been sanitized in any way.

3 Sanitization Method

In this section a simple supervised sanitization method based on entity recognition and pattern-matching techniques is presented. The purpose of this method is to identify and delete all entities and sensitive terms within classified documents that could disclose confidential information. As shown in Figure 2 we have divided the sanitization method in two steps. The first one performs the identification and anonymization of sensitive names or other personal information, while the second one performs the identification of text blocks which containing “risk” concepts, which later will be manually reviewed and eliminated. Both steps are described in detail below.

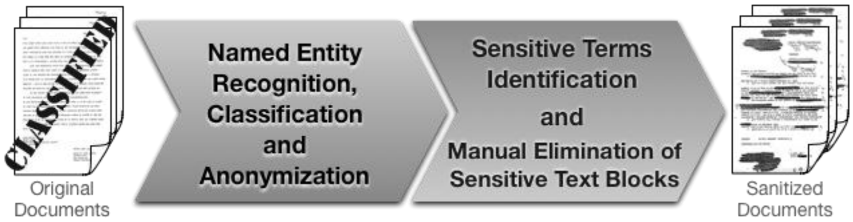

CLASSIFIED
Original Documents
Named Entity Recognition, Classification and Anonymization
Sensitive Terms Identification and Manual Elimination of Sensitive Text Blocks
Sanitized Documents

Fig. 2. Scheme for document sanitization.

3.1 Anonymization of names and personal information of individuals

To perform the first step we have used Pingar [13], an entity extraction software. This software identifies, classifies and anonymizes all named entities. It is able to detect the following named entities: people, organizations, addresses, emails, age, phone numbers, URLs, dates, times, money and amounts. The anonymization process is carried out replacing the identified sensitive information by its category plus an identification number. That is, {Pers1, Pers2, …}, {Loc1, Loc2, …}, {Date1, Date2, …} and so on. We also observe that the names of countries (Iran, United States, Russia, Italy, etc.) and places (London, Abu Dhabi, Guantanamo, etc) are unchanged in this process.

3.2 Elimination of text blocks of "risk text"

This step is also divided in to tasks; the identification of "risk" text blocks, which are those which contain the "risk" concepts, and the manual elimination of them. The risk concepts are represented by 30 keywords extracted from Section 1.4 of Executive Order 15326 [1]. In this section are stated eight points (a) to (h) defining the topics which the US government considers of risk in terms of national security. In Table 1 there is the list of the first 30 initial risk terms. As a list of 30 concepts are not enough to figure out if a text makes reference to any of the stated points we have used the WordNet ontology database [14] to extend it. So, for each of these initial concepts we have extracted a set of new words related with its sense, i.e., synonyms and hyponyms. By hyponym we mean the lower part of the ontology tree starting from the given keyword, that is, more specific words. For example, "weapon" would give the following: "knife, sling, bow, arrow, rock, stick, missile, cannon, gun, bomb, gas, nuclear, biological, …". Finally we have obtained a list with a total of 655 risk terms (original + synonyms + hyponyms). We note that in this extraction process the word sense disambiguation was performed manually.

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

In this section we present the method to evaluate the information loss and disclosure risk from a set of sanitized documents. This is performed by means of the results comparison when querying the original and the sanitized data set. In the 'Search Engine' sub-section we describe the characteristics of the vectorial model search engine implemented and in 'Metrics' we define the information loss and risk metrics. We note that the same metrics are used to measure information loss and disclosure risk. However, 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 and the risk queries consist of terms that define sensitive concepts.

4.1 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 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".

4.2 Information Loss and Risk of Disclosure Metrics

We have used as a starting point a set of well-known information retrieval metrics, which are listed in Table 1 and briefly described below. The formulas are defined in terms of the following sets of documents: *true_relevant_documents* is the unchanged, non-sanitized, document set retrieved by the corresponding query by the Vectorial search engine. *Retrieved_documents* is the set returned by the search engine in reply to a given query that is above the relevance threshold, *relevant_documents*, are the documents above the relevance threshold which are members of the *true_relevant_documents* set. *True_relevant_docs_returned* are the documents in *true_relevant_documents* that are returned by the search engine in any position (above or below the threshold) and finally, *false_relevant_docs* are the documents not members of *true_relevant_documents* but which are returned above the relevance threshold.

- *The Precision* is considered as the percentage of retrieved documents above the relevance threshold that are relevant to the informational query.
- The *Recall*, on the other hand, is considered as the percentage of retrieved documents above the relevance threshold that are defined as truly relevant.
- The *F-measure* (or balanced F-score) combines precision and recall and mathematically represents the harmonic mean of the two values.

- The *Novelty* is the proportion of documents retrieved and considered relevant which previously were not relevant for that query. That is, it measures the new information introduced for a given query. 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.
- The *Coverage* is the proportion of relevant documents retrieved out of the total true relevant documents, documents known previously as being the correct document set for a given search.

Table 1. Information Retrieval Metrics

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

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

As well as the four metrics listed in Table 1, we also consider four other measures:

- The average relevance of the documents whose relevance is above the relevance threshold.
- The total number of documents returned by the query whose relevance is greater than zero.
- The number of random documents which are members of the set of relevant documents for a given query.
- NMI (Normalized Mutual Information), we use an NMI type metric [17] for counting document's 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 random 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 Table 2 as an example for the search results of a given query.

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.

Table 2. 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

5 Experimental Analysis

In this section we describe the documents set used and how we have obtained a set of classified documents. Then, we present the results for information loss and risk of disclosure, comparing query results between the original and the sanitized data set by means of the presented metrics.

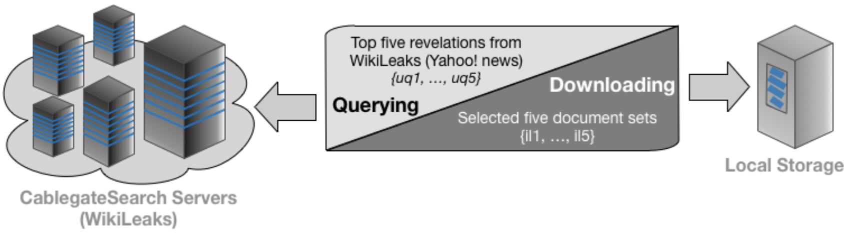


Fig. 3. Scheme for document extraction and querying.

5.1 Document Extraction

In order to test the proposed sanitization and evaluation techniques we have extracted a set of documents from the online Wikileaks Cable repository [2]. As in this online repository there are lots of documents related with different subjects, we selected the first five topics from the top ten revelations published by Yahoo! News [18]. We derived five queries corresponding to these five selected topics, as shown in Table 3. 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 3) were used to test information loss (utility) in the empirical results section. Figure 3 shows a schematic representation of the process.

As was mentioned in the Section ‘Sanitization Method’, we extracted 30 seed terms from the eight risk points defined in Section 1.4 of the US Executive Order 13526 [1], which are shown in Table 4. Hence, we defined eight different queries, one for each risk point, which are designated as {rq_1, …, rq_8}, corresponding to document sets {r1, .., r8}. These terms were used in our sanitization processing to detect ‘risk’ text blocks, and were also employed to define eight different queries which are used to evaluate the risk. We also defined a ninth query, rq_9, composed of all the terms from queries rq_1 to rq_8, whose corresponding document set is r9.

Table 3. 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 4. Queries used to test Risk of Disclosure

Id. Query	Keywords (risk queries)	ID1	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 rq1 to rq8.	r9	-

5.2 Information Loss

In Table 5 we see the results of applying the NMI metric to the original and sanitized document query sets. For the majority of query document sets, in general we see a relatively small information loss. In the case of query $uq_{5\text{-}1}$, the high information loss was due to the elimination of the named query terms , 'Putin' and 'Berlusconi', in the documents.

Table 6 shows the percentage change for each metric value and informational document set, of the original documents and the sanitized documents processed by steps 1 and 2. The indicators used in the information loss formula (6) are highlighted in grey. The information loss calculated using *formula 6* is shown in the rightmost column (IL), giving an average value of 26.1%.

Table 5. 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_{5-1}	uq_{5-2}
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 6. Information Loss: percentage (%) differences of statistics for original and sanitized 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_{5-1}	-100.00	-100.00	-1.00	-100.00	-100.00	-100.00	-4.55	0.00	-75.62
uq_{5-2}	-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

To summarize, 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. This step preserves the confidentiality of the personal data of individuals who appear in the documents. Step 2 (*elimination of 'risk text'*) inevitably had a higher impact, given that blocks of text are eliminated from the documents. From the results of Table 6, we see that the information loss is query dependent, the F and TR indicators being the most consistent. By manual inspection of the documents, we can conclude in general that a worse value is due to the loss of key textual information relevant to the query.

5.3 Disclosure Risk

We recall that the NMI metric measures the degree of correspondence between different groups. In Table 7 this metric is applied to the original and sanitized document query sets. A significant reduction can be seen in the correspondence, which contrasts with the results for the same metric applied to the information loss query document sets. Table 8 shows the percentage change for each of the metrics we described in Section 4.2, for each of the nine 'risk' queries, for the original documents and the sanitized documents of processing step 2. In general, we see a significantly greater percentage change in comparison to the information loss results of Table 6. The risk decrease calculated using *formula 6* is shown in the rightmost column (RD), the average value being -47.26%.

Table 7. 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 8. Risk of Disclosure: percentage (%) differences of statistics for original and sanitized 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

6 Conclusions and Future Work

In this paper we have used information retrieval metrics to evaluate information loss and disclosure risk for a set of sanitized documents. In order to evaluate these two values we implemented a vectorial model search engine and also 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_{5-1}) 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 semi-supervised learning methods applied to tagged examples. Also we could use a learning process to find the best overall descriptive formula for information loss and disclosure risk.

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, of the US Administration - Classified National Security Information, Section 1.4, points (a) to (h) (2009), http://www.whitehouse.gov/the-press-office/ executive-order-classified-national-security-information
2. Wikileaks Cable repository, http://www.cablegatesearch.net
3. Chakaravarthy, V.T., Gupta, H., Roy, P., Mohania, M.K.: Efficient Techniques for Document Sanitization. In: CIKM 2008, Napa Valley, California, USA, October 26–30 (2008).
4. Saygin, Y., Hakkani-Tr, D., Tr, G. (2009) Sanitization and Anonymization of Document Repositories.
5. Sweeney, L.: k-anonymity: a model for protecting privacy. Int. Journal of Uncertainty, Fuzziness and Knowledge-Based Systems (IJUFKS) 10(5), 557–570 (2002).
6. Cumby, C., Ghani, R.: A Machine Learning Based System for Semi-Automatically Redacting Documents. In: Proc. IAAI 2011 (2011).
7. Hong, T.-P., Lin, C.-W., Yang, K.-T., Wang, S.-L.: A Heuristic Data-Sanitization Approach Based on TF-IDF. In: Mehrotra, K.G., Mohan, C.K., Oh, J.C., Varshney, P.K., Ali, M. (eds.) IEA/AIE 2011, Part I. LNCS, vol. 6703, pp. 156–164. Springer, Heidelberg (2011)
8. Samelin, K., Pöhls, H.C., Bilzhause, A., Posegga, J., de Meer, H.: Redactable Signatures for Independent Removal of Structure and Content. In: Ryan, M.D., Smyth, B., Wang, G. (eds.) ISPEC 2012. LNCS, vol. 7232, pp. 17–33. Springer, Heidelberg (2012)
9. Chow, R., Staddon, J.N., Oberst, I.S.: Method and apparatus for facilitating document sanitization. US Patent Application Pub. No. US 2011/0107205 A1, May 5 (2011)
10. 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 8, 32 (2008).
11. Anandan, B., Clifton, C., Jiang, W., Murugesan, M., Pastrana-Camacho, P., and Si, L. (2012). t-Plausibility: Generalizing Words to Desensitize Text. Trans. Data Privacy 5, 3. pp. 505-534.
12. Miller, G.A., Beckwith, R., Fellbaum, C.D., Gross, D., Miller, K.: WordNet: An online lexical database. Int. J. Lexicograph 3(4), 235–244 (1990)
13. Pingar – Entity Extraction Software, http://www.pingar.com
14. Miller, G.A., Beckwith, R., Fellbaum, C.D., Gross, D., Miller, K.: WordNet: An online lexical database. Int. J. Lexicograph 3(4), 235–244 (1990)
15. Porter, M.F.: An algorithm for suffix stripping. Program 14(3), 130–137 (1980)
16. Baeza-Yates, R., Ribeiro-Neto, B.: Modern Information Retrieval: The Concepts and Technology behind Search, 2nd edn. ACM Press Books (2011)
17. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to Information Retrieval. Cambridge University Press (2008)
18. Yahoo! News. Top 10 revelations from Wiki Leaks cables, http://news.yahoo.com/blogs/lookout/ top-10-revelations-wikileaks-cables.html