A Review on Text Sanitization

Veena Vasudevan
PG Scholar, CSE
T.K.M college of engineering
Kerala,India

Ansamma john
Associate professor, CSE
T.K.M college of engineering
Kerala,India

ABSTRACT
Information is essential for all purpose of activities such as research, business decision making, etc. In this internet technology age there is no scarcity of information also. But if the information reveals the identity of a person or if it discloses confidential matters, then such information is a serious threat to privacy. So before publishing or sharing documents, the sensitive information should be removed or masked. This is the major goal of Text sanitization. Several semi-automatic and automatic methods are used for identifying sensitive information and thereby sanitizing the document by removing such terms. This broadens the users using the document due to their lowered classification level and also privacy is preserved.

General Terms
Text Mining, Privacy Preserving Data Publishing, Text Redaction

Keywords
Document Declassification, Data Publishing, Term correlation Privacy, Information Theory, Named Entity Recognition

1. INTRODUCTION
In a recent study it has been found out that the internet contains about five hundred billion gigabytes of digital information content. This information might come handy in various purposes such as scientific research, business decision making, etc. Nowadays in the cloud computing industry information needs to be shared [1].The data owners can allow the data access to one party, and in turn the party can further share the data to another party without the permission of the data owners. Therefore the data owners need to consider whether the third party continues to maintain the original protection measures and usage restrictions. If the information used in such cases reveals the identity of a person or is confidential information, then it poses a fatal threat to privacy. Such information is called sensitive information, which should be taken care of prior to the publishing or sharing of data. Official regulations have been developed at this respect. According to a recent EU bill [2] if any company moves the confidential data of their clients to the Cloud, then they are liable to pay a fine which may amount to a substantial part of its revenue. In medical field, according to Health Information portability and Accountability Act (HIPAA) of 1996 [3] confidential information in patient records should be removed prior to publishing.

Document sanitization is the process of identifying sensitive information and removing or hiding such information from a document. So we can say that it is a two step task. In the first phase it identifies the sensitive terms and the second phase consist of removal of identified sensitive terms. In earlier days more attention is given to the sanitization of structured documents. But organizations such as intelligence bureau, government enterprises and various other large organizations also want to remove sensitive information from semi-structured and unstructured documents before publishing them outside their organization. In a structured document, identifiers, quasi-identifiers and sensitive attributes are defined explicitly. Identifiers are the personally identifying information like name, social security number, etc. Quasi-identifier attributes are attributes which when taken together can identify an individual such as zip code, date of birth etc. Sensitive attributes are not known to the intruders and are considered as sensitive such as disease, salary, etc. Whereas in an unrestricted text, that has no strict rule governing its content and structure, sensitive information is relatively difficult to recognize. By removing the sensitive information the classification level of the document is lowered and it will be available to much more people. Documents need to be modified in such a way that utility of the document is maximized and distortion of data is minimized.

Sanitization of text document is traditionally done manually by qualified reviewers. Manual sanitization becomes difficult as the volume of data increases. So semi-automatic and automatic methods are proposed. The developed methods remove the scalability and time problems. The rest of the paper is organized as follows. Section 2 discusses various techniques used to identify sensitive terms. Section 3 discusses the methods used to mask the identified sensitive terms. Section 4 concludes the paper.

2. VARIOUS SCHEMES TO IDENTIFY SENSITIVE TERMS IN A DOCUMENT

2.1 Using HIPAA rules as base for identifying sensitive terms:
According to HIPAA "safe harbor" prescript, eighteen data terms, such as name, location, medical record number, etc, must be removed from clinical data prior to sharing it to a third party. Latanya Sweeney [4] proposed a system named scrub that uses detection algorithms, competing in parallel, to determine patterns of identifying information such as name, location, social security number, age, etc. Each detection algorithm recognizes a specific entity. The detectors identify sensitive information utilizing regular expression type templates and knowledge sources. At run-time the user sets the threshold and use of special detectors. The data determined as sensitive is replaced with some pseudo-value. For each detection algorithm there is a corresponding replacement algorithm which is responsible for producing the replacement text. The format of replacement text matches the template that was recognized. This improves the results of the straight search-and-replace technique. But it does not detect most references to family members, nick names, extra phone

numbers, etc. On the other hand, it discovers 99-100% of all personal information.

A group of researchers from MIT [5] proposed the use of lexical look-up tables, simple heuristics and regular expressions to identify personal health information in free text nursing notes with an overall sensitivity of 0.92. It uses regular expressions for determining whether numeric tokens should be separated as telephone numbers, dates or other types of identifying numbers. Non-numeric tokens are sorted by applying simple heuristics and by using lexical matching. But the look-up tables have some shortcomings such as the scarceness of common abbreviations and names of drugs. Also this method suffers from having high false positive rates, so an additional manual reviewing of selected terms are essential.

In [6], the authors proposed a six step anonymization scheme that detects and replace patient names with pseudonyms .It is a semi-automatic corpus-driven approach which utilizes local dictionaries in addition to external dictionaries and simple NLP. But the method ignores the context in which a word appears in the document. The main limitations of the above proposed systems are that they neglect the data semantics during the masking of sensitive terms. Hence this affects the utility of the document. Also applicability is reduced due to the use of specific patterns.

2.2 Using database of entities to identify sensitive terms

The authors in [7] proposed a system named ERASE (Efficient RedAction for Securing Entities), an automatic sensitive term detection framework that model public knowledge as a database of entities. The database can be structured or unstructured. There is no need of manual compilation of the database. The framework sanitize a document in such a way that users can access only what they are authorized to know i.e same document can be viewed differently by different users. In an unstructured free-text domain, ERASE is the first proposal that provides a principled solution. Within the database, each entity (such as person, product, location, etc) is tagged with a set of related terms called context of the entity. Some entities are marked as protected, disclosure of which reveals the identity of entity. By erasing certain terms from the document ERASE prevents disclosure of entities marked as protected. The algorithm removes only a minimum number of terms. Only after considering the terms that co-occur with a term, the particular term is removed. It tries to make a globally optimal solution. The sanitized document obtained is least distorted, thereby increasing the utility of document. The privacy and utility levels of the document are set to a desired level by using the "k-safety" concept. K-safety demands that the maximum subset of each sensitive entity's context incorporated in a document is also included by the contexts of atleast K other entities .But when the sanitization needs differ, the use of specific purpose knowledge bases hinders the applicability of the technique.

2.3 Using trained classifiers to identify sensitive terms:

In large enterprises there is a need to redact documents before publicly releasing them or before sharing with internal fellow workers. So the sensitive information such as client identifying information and personally identifying information must be removed from the documents. The authors in scheme [8], proposed a semi-automatic sensitive information removing tool. Detection of a sensitive concept is treated as a multiclass classification problem. A Naive Bayes classifier is trained using machine learning techniques to identify the client identifying information and then perturbation is applied to the text. It tries to optimally perturb the document so that classification error for the sensitive class within this set is maximal. Only those terms that are previously tagged to the classifier are identified, so it halters the generality of the method. Sanitization can be set up according to the suggested "K-confusability" model. The authors say that a third party cannot find out which company among other appears on a sanitized document that holds K-confusability.

Daniel.et.al [9] proposed a system that uses trained classifiers to detect the sensitive entities. It consists of two tasks: private entity recognition and Private entity protection. A private entity in an unstructured document is an entity which reveals information about the correspondents (organization or individual) directly or indirectly associated with the document. Private entities are identified to be coincident with named entities. It uses named entity recognition techniques to identify the entities of the document that need to be protected. In an unstructured text typical named entity are proper locations, names or organizations. In order to identify named entities, Named Entity Recognition and classification systems rely on different features associated to the text at different levels, normally: word, list, and document. The proposed entity protection method includes generalization of entities, swapping of entities, and adding noise to entities. The main limitations are that not all sensitive terms can be represented using named entity and all named entities may not be sensitive. With reference to their specificity and the fact that they represent individuals rather than concepts, NEs are likely to reveal private information. Also a limited number of Named Entity types are detected using most generic Named Entity Recognition packages.

2.4 Using information theory to evaluate the degree of sensitiveness of terms:

Sensitive terms are those that provide more information than other terms in a document. On the basis of information theory, the amount of information provided by a term can be represented by using information content (IC). The IC of a term t is computed from the hit count of a web search engine (WSE) using equation I and II:

$$IC_{WSE}(t) = -log_2 P_{WSE}(t) \quad \text{(I)}$$

$$-log_2 P_{WSE}(t) = -log_2 \frac{hit_{count_{WSE}}(t)}{|WEBS\ INEXED\ BY\ WSE|} \quad \text{(II)}$$

Higher the information content value, higher the sensitivity of the term. The schemes [10],[11], are based on this concept and use web as the corpus. In [10], the author use information content value to identify individual sensitive terms. The proposed method focuses on the detection of noun phrases (NP) as potentially sensitive semantic units. Once NPs are detected, then the amounts of information provided by them are measured by querying them in a web search engine. The NPs having information content greater than or equal to a minimum threshold are considered as sensitive. Then replace the identified sensitive terms using generalization strategy. The main focus is on preserving the utility of the sanitized

document. The schemes [5], [6], [7], uses term suppression but [10] relies on knowledge bases, such as WordNet [12], to hide sensitive information while preserving its meaning, retaining more document utility. Instead of using problem-specific ad-hoc knowledge bases or trained classifiers as in [4],[5],[6],[7],[8],that severely hamper the applicability of sanitization methods,[8] relies on external general-purpose knowledge bases/corpora such as web. As a result, [10] offers a domain independent solution that can be applied to textual documents regardless of their contents. It also allows the user to set up the level of sanitization applied to the document .It is also more flexible than methods based on a fixed sanitization policies such as [4],[5],[6],[7],[8]. Moreover, since it is based on the well-known information theory and on linguistic labels, its configuration is more intuitive and comprehensible than methods based on abstract numerical models as in [5],[6], [14]. But semantically correlated terms are also present in the document, using which the sensitive terms can be inferred. So in [11], the authors proposed a method to find the correlation between terms, to identify the terms that represent a feasible disclosure risk. It also consists of two tasks: measuring the correlation between the sanitized and non-sanitized term and removing the non sanitized terms for which the degree of correlation is high. The correlated terms are identified using the concept of Point-wise Mutual Information (PMI), which quantifies the difference between the probability of their co-occurrence given their join distribution and their marginal distributions [13]. If two terms co-occur in a text by chance then their PMI value will be zero. But if there is a correlation between the terms then their PMI vale will be positive. Therefore negative or zero values for PMI is used to infer that there will be no disclosure risk, however positive values will infer an increasing disclosure risk. This method can be used as a second step in all the above algorithms which analyze and detect sensitive terms independently. This method minimizes the disclosure risk and also reasonable level of utility is preserved.

3. VARIOUS TECHNIQUES USED TO HANDLE IDENTIFIED SENSITIVE TERMS

Detected Sensitive terms should be removed or masked before publishing the document. To maintain its utility text semantics should be retained. In earlier search-and-replace technique the sensitive terms are replaced with pseudo-values or synonyms. Whereas some proposed systems [11] completely removes the identified sensitive terms inorder to avoid de-identification. So it provides zero information. But complete removal of the details is not always essential because it affects the readability and usability of the text. So various other techniques are also used.

3.1 Generalization

Generalization is the process of replacing more specific terms with less specific or generalized terms (eg: iphone is replaced by Smartphone). By this method semantics of the text can be retained even though some amount of information is lost. Knowledge bases that provide high recall and offer detailed knowledge representation can be used in generalization process. If the detected sensitive term is not present in knowledge base then remove the term or replace with some random entity or replace it by the most general abstraction. In all these cases so much information is lost. By having detailed Knowledge representation, the information lost during each generalization step can be reduced. WordNet is an obvious choice for general-purpose sanitization methods. According to [14], a word can be generalized to other word if both of them belong to a same hypernym tree. The work proposes a theoretic measure, called "t-plausibility", which evaluates the quality of sanitized documents from a privacy protection point of view. A generalized text document holds the t-plausibility model if it can be associated with atleast t-base documents and any one of them could be the original document. A limitation of this method is that it does not consider any term relationships.

3.2 Swapping

Swapping process swaps private entities between documents of the same set, or within the same document depending on the concrete case. Relatively similar private entities can be interchanged. To perform swapping, the private entities can be ranked using a distance metric.

3.3 Noise Addition

To provide some degree of anonymity some semantic noise can be added to the private entities. Instead of swapping similar PEs, substitute a PE by another similar PE, which is not present in the document. If higher protection is required, random substitution at higher levels can be performed.

4 CONCLUSION

In this information age document publishing is one of the important tasks. But preserving the privacy of the document is also very much important. For preserving privacy the sensitive information need to be removed, but its utility should not be affected. . In this paper, various techniques used for identifying and removing sensitive terms from a document are discussed. These techniques are very essential in this fast growing information age where millions of documents are published and shared. Thus text sanitization is a fast growing research area where privacy and utility measures plays an important role.

5 REFERENCES

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

[2] S. Pignal, "EU eyes big fines for privacy breaches," FinancialTimes2011[Online].Available:http://www. ft.com/intl/cms/s/2/bf962998-1d01-11e1-a26a-00144feabdc0.html#axzz1fe8ewpqO

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

[4] L. Sweeney, "Replacing personally-identifying information in medical records, the scrub system," in Proc. 1996 American Medical Informatics Association Ann. Symp., 1996, pp. 333–337.

[5] M. M. Douglass, G. D. Cliffford, A. Reisner, W. J. Long, G. B.Moody, and R. G. Mark, "De-identification algorithm for free-text nursing notes," Proc. Computers in Cardiology'05, pp. 331–334, 2005.

[6] 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)

[7] V. T. Chakaravarthy, H. Gupta, P. Roy, and M. Mohania, "Efficient techniques for document sanitization," in Proc. ACM Conf. Information and Knowledge Management'08, 2008, pp. 843–852

[8] C. Cumby and R. Ghan, "A machine learning based system for semiautomatically redacting documents," in Proc. 23rd Innovative Applications of Artificial Intelligence Conf., 2011, pp. 1628–1635.

[9] D. Abril, G. Navarro-Arribas, and V. Torra, "On the declassification of confidential documents," in Proc. Modeling Decisions for Artificial Intelligence'11, 2011, pp. 235–246.

[10] Sánchez, D., Batet, M., and Viejo, A, "Automatic General-Purpose Sanitization of Textual Documents", IEEE Transactions on Information Forensics and Security, VOL. 8, NO. 6, JUNE 2013,pp. 853-862

[11] Sánchez, D., Batet, M., and Viejo, A," Minimizing the disclosure risk of semantic correlations in document sanitization", Information Sciences,2013,pp. 110-123

[12] C. Fellbaum, WordNet: An Electronic Lexical Database. Cambridge,MA, USA: MIT Press, 1998

[13] Church, K. W., and Hanks, P. (1990). Word association norms, mutual information, and lexicography. Computational Linguistics, 16 (1), 22-29.

[14] B. Anandan, C. Clifton, W. Jiang, M. Murugesan, P. Pastrana-Camacho, and L. Si, "t-plausibility: Generalizing words to desensitize text," Trans. Data Privacy, vol. 5, pp. 505–534, 2012