BIBLIOGRAPHY

(A REVIEW ON RESEARCH PAPERS)

Paper 1

Title of the paper: Information leakage through Document Redaction: Attacks and Countermeasures.

Paper reference: Daniel Lopresti and A. Lawrence Spitz. "Information leakage through Document Redaction: Attacks and Countermeasures", Computer Science and Engineering, Lehigh University, Bethlehem, PA 18015, USA.

Abstract:

Redaction has been known to mask sensitive data and present it to the user, but sometimes the redacted data could be broken i.e. the original data can be recovered using image processing techniques and processes involved relating the natural language techniques. The paper mainly focusses on the process of redaction and what could be done to avoid circumstances that result in leakage of sensitive information.

Evaluation of work:

This paper provides details on redaction and the measures that are needed to be taken to avoid leakage of important data. Though the document is redacted, the original document could be recovered using "NLP techniques" and manual processing. The author of the paper did provide some instances where in the most important data relating a country's security was recovered by attributes relating the font style. Hence the author suggested that there is a need for automatic analysis of data to maintain the integrity of data. The information leakage from a redacted document could happen if the text on the document is not properly redacted due to the toner or ink or may due to the font style and attributes. The redacted text's width could be one of the main reasons for the sensitive information's leakage as natural language techniques combined with text width is capable of breaking the data.

Experiments were carried out involving four groups of words commonly called as Lexica and conclusions were drawn that few font styles are more prone to attacks than others. Several measures were proposed to combat the risks involved. One of the best countermeasures discussed in the paper is the use of special font where each characters width is randomly varied and also a non-redacted font style could be placed around the redacted text to avoid the estimation of the obscured regions width. Recognition techniques related to optical styles could be used to keep the width of the obscured regions hidden. The author has proposed a semi -automated system which takes the redacted document as the input and gives an output whether or not the document be released to the public.

Paper 2

Title of the paper: Replacing Personally-Identifying Information in medical records, the Scrub System.

Paper reference: Latanya Sweeney." Replacing Personally-Identifying Information in medical records, the Scrub System", Clinical Decision making group, Laboratory for Computer Science, Massachusetts Institute of Technology, Cambridge, Massachusetts.

Abstract:

The main goal of any medical institute is to protect the privacy of patient's data, an automated approach where in the personal sensitive information of the patient is located and automatically replaced using certain algorithms. Using these algorithms can maintain the privacy of patient, Scrub system is one such approach which advocates algorithms that detect sensitive data based on templates and special information.

Evaluation of work:

The research paper focusses on the idea of "Scrubbing" which basically aims at removing sensitive identifying data in a patient's medical record as well as maintaining the identity of the patient at the same time. There are different detection algorithms for each entity concerning the patient, various other knowledge sources are as well used to identify whether or not the data is medical term. Humans were also used as an experimental tools and were asked to identify sensitive information pertaining the patient's records. Unlike human approach, in computer approach special template models and designs were used to identify personal information. Along with those templates, worldly facts are also required in training the system to identify data pertaining the patient. Detection algorithms for each entity of the patient are executed in sequential manner, these detection algorithms also use a list of stored items corresponding to each entity along with templates that are used in training the database. Once the sensitive information is identified, the information is replaced with pseudo and fictitious names. The author did discuss about reverse scrubbing technique at the end of the paper.

Paper 3

Title of the paper: Detecting Sensitive information from textual documents: an information-theoretic approach.

Paper reference: David Sanchez, Montserrat Batet, Alexandre Viejo, "Detecting Sensitive information from textual documents: an information-theoretic approach", Department d'Enginyeria Informatica I Matematiques, UNESCO Chair in Data privacy, Universitat Rovira I Virgili.

Abstract:

If an important document is to be released to the public, the sensitive information in the document has to be hidden to protect the privacy. A number of methods are available to detect fields like social security number or email id or any other important information, but there is still a need for methods which automatically identify raw data in textual format.

Evaluation of work:

The author of the paper has very well addressed the need for methods that prevent reverse sanitization of redacted data in order to protect the confidentiality of data, some standard guidelines have to be followed to achieve this. The sanitization done by manual approach could be little expensive and time consuming too, it is very difficult for government agencies to do the process manually as they should process a number of documents. Hence there is a need for semi -automated system that automatically detects sensitive information and sanitizes it. The author has mentioned that unlike scrub systems that use specific patterns to detect the sensitive information, a method has been developed that use data sets pertaining to the entities are used to help identify sensitive information. A tool has been developed which can be used along with Microsoft word to anonymize the data set entities.

The sensitive information in any document in general corresponds to number of sensitive terms, the amount of information contained in these sensitive terms tells us about the degree of confidentiality associated with that term, hence approaches are needed to first identify the amount of information, to address the above issue, the author did mention about "Information Content". Mathematical formulas were proposed to calculate the information content using terms like page count and the number of websites. Formulas for calculating noun phrases also have been computed. Theoretical derivations have been made to incorporate automated and semi -automated methods for detecting sensitive information in the document.

Paper 4

Title of the paper: A Machine Learning Based System for Semi-Automatically Redacting Documents.

Paper reference: Chad Cumby, Rayid Ghani, "A Machine Learning Based System for Semi-Automatically Redacting Documents", Accenture Technology Labs, 161N. Clark St, Chicago,IIlinois 60601.

Abstract:

One of the earliest methods of redacting documents was a manual one. There is a need for semi-automated system to make sure that confidentiality of the document is maintained. Tools built on Microsoft word allow users to sanitize documents before it is shared by masking sensitive information of the client. Experiments were done using the available datasets, the results showed that the text's utility was preserved while still reducing the risk associated with the sensitivity.

Evaluation of Work:

The author of the paper has addressed the need for sanitization of documents in a government agency or commercial organization before it is released to the general public because there is a possibility of confidentiality of the document being broken because of the insiders. The author of the paper has mentioned that there could be two pieces of Information that system could redact and they are "Client Identifying Information" and "Personally Identifying Information". The work referring to both the concepts has been clearly mentioned in the paper. The author did mention about a number of working algorithms and the experimental results were very well presented. The main idea behind the approach was preserving or enhancing the utility of the content while still meeting the privacy and confidentiality needs. At the end of the paper, the author has concluded that he would like to extend his work by experimenting with a number of feature oriented representations, which is a very great idea as it addresses a variety of problems.

Paper 5

Title: Efficient Technique for Document Sanitization.

Paper Reference: V. Chakaravarthy, Himanshu Gupta, Prasan Roy and Mukesh K. Mohania; "Efficient Technique for Document Sanitization"; IBM India Research Lab.

Abstract:

Erasing the sensitive information from a document for reducing its level of classification is referred to as sanitization. This process may yield a document which is unclassified. Various government departments, hospitals, companies etc. must first sanitize the information before it is made available for the public. "Efficient Redaction for Security Entities (ERASE)" is a procedure that is proposed for the automatic sanitization of information in this paper.

Evaluation:

Traditional approach of information sanitization involves qualified persons who perform sanitization manually but this approach is not scalable for large amount of data. In this paper, a nontrivial strategies are proposed to identify the critical information and reduce the document.

According to ERASE model, the public knowledge is modeled as different entities which are associated with certain set of terms of the structured or unstructured database. Context of an entity and the cluster of entities that contain the given term are identified by the database. For instance, consider a hospital database in which certain diseases like AIDS can be considered as the data that must be protected. The data that should be masked from an adversary is considered to be the protected data.

ERASE hides the protected entities of a document by removing hiding them. The basic approach that is described in this article is that critical/sensitive data are identified and deleted before presenting it to the public.

Paper 6

Title: Efficient Signature Schemes supporting redaction, pseudonymization and data de-identification.

Paper Reference: Stuart Haber, Yasuo Hatano, William Horne, Yoshinori Honda, Kunihiko Miyazaki, Tomas Sander, Satoru Tezokuy and Danfeng Yao; **"**Efficient Signature Schemes supporting redaction, pseudonymization and data de-identification". HP Laboratories Palo Alto.

Abstract:

A new Signature algorithm is designed for providing control access to the signed data. Several operations like subdocuments removal, de-identification of hierarchical structured data and pseudonymization can be applied in practical applications. If we directly apply this to redaction, it reduces the cryptographic information overhead that is stored along with the actual information.

Evaluation:

This paper mainly focuses on the procedure pseudonymization, redaction and de-identification of data. Consider the disclosure of government documents on the request of the individual. Here the data is hidden prior to the release and this is called redaction. In some cases data cannot be redacted but can be replaced by pseudonyms in the whole document by maintaining consistency. In other cases where data is related to health records, employee information etc. must be de-identified.

A legal statement that proves that the document has been prepared based on the law would prove to be an example of a signer that appears on the document for all its lifetime. This model includes three types of players- "singers", "redactors" and "users". Preparing, authenticating and producing digital signature of a document is done by signers, data modification based on the list of operations allowed by the signers is done by redactors and finally a user will be able to verify the correctness of data that is modified with the help of extended signature.

Paper 7

Title: An Evaluation of Technologies for the Pseudonymization of Medical Data.

Paper Reference: Thomas Neubauer and Mathias Kolb; **"**An Evaluation of Technologies for the Pseudonymization of Medical Data**".**

Abstract:

The most important fundamental issue in health care is privacy. This paper mainly focuses on privacy protection of patients and implements the approaches of pseudonymization in technical and legal aspects. On the whole, this paper provides support to the decision makers in deciding the privacy of the systems and even for the researchers to identify the limitations of current procedures in privacy protection for future aspects.

Evaluation:

The most important topic discussed in this paper is Electronic Health Records (EHR) the helps in improving communication between the providers of health care, documentation and data access which results in good quality of service. Diagnostic tests and images are digitized by EHR to confirm massive savings. Personal information and sensitive data over the internet are highly protected.

The approaches of Pseudonymization include certain requirements to be related to the privacy laws in United States and European Union. "User Authentication", "Data Ownership", "Limited Access", "Protection against Unauthorized and Authorized Access", "Notice about uses of patients data" and "Access and Copy own data" are a few requirements that are extracted from legal acts.

To provide a good level of security pseudonymization approaches also include "Fallback Mechanism", "Un-observability", "Secondary Use", "Emergency Access", "Insider Abuse"and "Modification/Physical Compromise of the database". Finally, this paper gives an overview of the directions that are used currently in e-health for the privacy protection.

Paper 8

Title: Sanitization's Slippery Slope: The Design and Study of a Text Revision Assistant.

Paper Reference: Richard Chow, Ian Oberst and Jessica Staddon; **"**Sanitization's Slippery Slope: The Design and Study of a Text Revision Assistant**";** Oregon State University.

Abstract:

Revision of sensitive information prior its release mostly involves redaction which means "blacking out" the sensitive phrases and terms but redaction has a drawback of reducing access to the content and making it no longer useful. Government and many other departments have started using revision as an alternative to redaction that conserves the data utility and this is called sanitization. Pseudonyms replace names and hypernyms replace sensitive attributes in a sanitized document. In this paper, a new tool is designed that assists the user to sanitize the sensitive information but is not fully automated. The reason for appointing sanitization assistant is to identify all the sensitive terms that cannot be done by sanitization alone.

Evaluation:

A new prototype for text sanitization using the sanitization assistant is designed and developed in this paper. This tool estimates the Privacy risk using the web based data mining, clear general phrasing is provided by online directories that decreases the privacy risk and finally allowing the user to use scoring mechanism to reduce as much data as possible.

The study in this paper proves to save the user's work and improves the facility of maintaining and increasing the privacy of sanitized document based on the familiarity of the user with the subject. By this facility the privacy of a document may be increased to the highest which might be a problem. On the whole, this tool focuses on sanitizing the document based on the user familiarity with the information.

Paper 9

Title: Process of Electronic Document Redaction.

Paper Reference: David George Kelly and Brent Russell Foster; "Process of Electronic Document Redaction"; Onstream Systems Limited.

Abstract:

In this paper, a redaction process that enables a user in redacting an electronic document is described. It includes various sequence of steps in retrieving an electronic image document from an original native format document. Few sections of an electronic image file are redacted by the user and are saved in an electronic file format.

Evaluation:

The electronic documents that are created by various types of aforementioned applications contain metadata. Now, when this document will be redacted the data that is redacted might still be preserved in the electronic file. This makes it easy for the skilled person to recover the data lost or changed.

The redaction process designed and developed in this paper first convert the electronic file into a native electronic image file where "text" and "text location data" are extracted for the process of printing. This electronic image file serves as a format of tagged image file which includes changing the original pixel values to the redaction marking values. These values cannot be modified by the user anymore. And finally, an audit copy of redacted electronic image file is saved by the system which can only be read for the changes made to the actual electronic document.

Broadly according to this invention a redaction process is provided for the user to redact an original document to an electronic image file and then saving it electronically. Another invention includes a computer that is programmed to convert a copy of the original document into a redacted image file that is stored in electronic format.

Paper 10

Title of the paper: Achieving k-Anonymity Privacy Protection Using Generalization and Suppression.

Reference: Latanya Sweeney, "Achieving k-Anonymity Privacy Protection Using Generalization and Suppression", School of Computer Science, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA.

Abstract:

In commercial organization or government agency, it is required that the sensitive information associated with the subject of interest is hidden. The documents are subjected to k-anonymity before release so that the sensitive information in patient record becomes hidden. This approach provides security and privacy to the fullest because the each of the record that is released is related with k individuals. It addresses the use of MinnGen algorithms which helps in achieving K-anonymity.

Evaluation :

The author of the paper has addressed several methods to achieve k-anonymity in order to protect the privacy of the sensitive data and the methods involved include "suppression" and "minimal generalization", he did even mention about MinGen algorithm that uses the above mention methods namely suppression and generalization to produce tables which adhere to the concept of K-anonymity which in term maintain the utility of the data by achieving minimal distortion. Several algorithms and systems were mentioned to achieve the concept. Though the paper has a number of algorithms to achieve the concept, it has got several loop holes and further work has to be done to address them. Overall the idea of the paper was good as it focused on the concept of maintaining confidentiality and privacy.