Intelligent Data Redaction

Sailaja Pedaprolu, Srujana Kondamadugula, Divya Sree Vintha, Lakshmi Bhargavi Kadali
UMKC, School of Computing & Engineering

***Abstract*---The simplest measure to provide confidentiality to a document is ensuring that it does not contain any sensitive data before making it available for public use. This technique of obscuring the private information in a piece of information is termed as Data Redaction. There are several practices to achieve redaction. Based on the application requirements, suitable techniques are employed.**

The domains that Data Redaction operates on could be classified majorly into two groups: structured and non-structured. It is challenging to perform redaction on the latter domain. There is a need to develop a scheme that performs automatic redaction unlike conventional techniques which mostly focus on structured data. The best method to achieve masking is by incorporating machine learning concepts like classifiers or SVMs (Support Vector Machines). We have made an effort to address both the domains in this paper.

***Index Terms*---Data Redaction, structured data, non-structured data, automatic redaction, classifiers.**

I. Introduction

Be it a commercial organization or a government agency, there would be a large number of employees working under a common roof. Each of the employees would have different privilege levels to access a specific information. For privacy reasons and to maintain confidentiality of such information it is optimal if sensitive data is cleared out or made unavailable to the unauthorized viewer. This is termed as Data Redaction. Data sanitization and data masking also refer to the same idea of removing the sensitive data. The process involves two steps: identifying the parts of information to be redacted and choosing a format to mask/ alter the data. The format here implies, if the data is scrambled or replaced with a string or made prone to physical damage. The main purpose here would be to make the chosen private data unreadable or invisible.

Each redaction technique varies with process selected to identify the confidential parts of data and the format of redaction. Determining the data if it is sensitive or if it requires to be redacted could be static or dynamic. Static redaction deals with the information that follows some pattern or structure. So, if structured data is recognized, that is redacted. On the other hand Dynamic redaction deals with non-structured data and thus requires automatic detection. The automatic detection/ recognition of parts of sensitive data in a document requires intelligence in the algorithm. Machine learning brings that intelligence. Key concepts of supervised learning such as various classifiers, support vector machines could be employed to predict if the data falls into the category of private information. Thus, dynamic redaction is achieved.

Redaction could also be general or specific. This classification is based on the access level of the viewer. If a document is open to all the users, each user would belong to a certain access level. General redaction masks the data and makes it unavailable to all the users. On the other hand, specific redaction enables masking of only that data which the user is not supposed to view.

Oracle and IBM provide data redaction policies which allow the users to customize the functionality to redact a document. These are not totally dynamic, but allow general and specific redaction.

The paper showcases the existing work, then gives the details of the problem domain, followed by experiment description that summarizes the assumption made in the design phase. The experiment is divided into two phases. The algorithms implemented in the two phases have been detailed in experiment procedure which is followed by the results and conclusions.

II. Related work

There are several approaches to redaction, one of the conventional and earliest methods of redaction used by government agencies for redacting confidential paper documents is physical redaction where in a copy of the original document is made while still preserving the original document and masking is done on the copy of the original before releasing it to the public, the techniques used were obscuring the confidential information or data with a thick ink or by using special tapes which keep the information hidden.

In some instances, knives and some sharp utensils were also used to damage sensitive part of the document so as to protect it from misuse. But the above method did not prove to be very efficient because the obscured data could have been easily recovered with manual approach combined with certain natural language processing techniques or sometimes if the sensitive data is not properly blackened, the confidentiality of data could be broken.

Font style and spacing between the words in a document also played a major role in the process of document redaction and hence researches even paid closed attention in discovering specialized font styles which could prevent reverse redaction. The researches even focused on placing non redacted text around the missing or redacted part so that the width of redacted word becomes unknown. The special non redacted words that were used where called as jiggles. The next approaches in Data redaction focused on the concepts of "Scrubbing", Scrubbing was basically done on patient's record in a medical organization, and a new approach was developed

in identifying the personal and sensitive information in a patient's record and then replacing it with other unrelated words. Experiments were done on machine scrubbed and hand scrubbed samples.

Numerous detection algorithms were used in the scrubbed system to make the redaction feasible, the detection algorithms were executed on sequential fashion. One of the techniques which posed threat to existing scrubbing technique is the reverse scrubbing techniques wherein the original data was recovered from the scrubbed system results. Insiders were main reason for the reverse scrubbing as they were the ones who posed a serious threat.

The other methods that were used to sanitize documents incorporated the concepts of information theory, this information theory method mainly dealt with sanitizing raw textual data. Information content played a major role in the identification of sensitive data in a document, several formulas were proposed to calculate the content in an information. Page counts and web content information were used as inputs to make theoretical derivations. Though the information content theory proved to be somewhat efficient, it did not meet the needs of Sanitization. There were certain limitations with Information Content concept, the IC calculus concepts had some problems.

Another technique was develop which was a polished version of sanitization, the main idea was to preserve the utility of the content. A tool was developed in this approach which automatically identifies the sensitive information in the documents and aims at revising the content so that the utility of the document or the redacted portion is not last. In this approach, the user is also capable of knowing the privacy risk associated with the content. The sanitized tool developed has two views which are very distinct from each other, the primary view basically aims at inferring phrases related to the sensitive parts in the document and it also enables the user to edit the document. The second word allows us to view replacements for a given word. The tool also enables the user to switch between both the views. This procedure was not very successful as some times it may be problematic to the user because their own sanitization design might require more of background knowledge. This procedure should have also been made available to large groups of study.

The other approach for sanitization is the concept of "ERASE". The main idea behind this concept is that it aims at modeling the entities of database. A set of terms are related to each entity and this idea is briefed as the "context" in an entity. The context in an entity relate to all the details associated with the entity. It also aims at preventing the disclosure of the protected entities as certain terms are removed from the document. The concept of ERASE is somewhat better than others because it does not distort the original document and one of the advantages with the concept of ERASE is that it is capable of identifying the least number of terms in the document so that the document becomes sanitized.

A lot of important information is contained in the legal documents and it is sometimes important that we don't make any changes to the original document as it is abided by law and the originality of the report should be maintained throughout the life time of the document. It does not make any sense in redacting the whole document because redaction of certain parts of the data is just sufficient. Hence there is a need for cryptographic techniques as they prevent further redaction of data. The redaction can be done using keys and the modification of data could be made feasible with hash functions. Redact-able signature algorithms have also been proposed that are useful in redacting the original document.

III. Problem domain

The aim was to design an application that addresses all the different kinds of redaction mentioned earlier. A suitable scenario has been assumed to depict the needs for static, dynamic, general and specific redactions. An organization controlled by an administrator is considered which deals with different projects, each project led by a manager with a group of employees. There exist many documents which could be commonly dealt with all of these entities (administrator, manager and employee) but privileges to see the parts of this document could be variable. So, there should be a *specific redaction* that masks particular data depending upon the designation of the user viewing it. *General redaction* should be employed to alter the data that is not shown universally, meaning irrespective of the viewer's designation that particular data is redacted. As part of general redaction comes the *static redaction* which is redaction based on structured formats like date, email, address follow a structure which could be regarded as a key to identify that this data has to be redacted. There should also be dynamic redaction where data is checked if it belongs to a category, if yes it is redacted.

IV. Experiment Description

The experiment has been performed in two phases.

Phase 1: Two excel spreadsheet have been taken. Each of the excel sheets consist of the details of the managers and employees respectively. A few assumptions have been made which are as follows:

- There is an administrator who can view all the contents of the excel sheet.
- The manager of a project team could only view the details of employees working in the same project except for few details.
- An employee could view the details of all the managers except for few details.

In the course of redaction of excel spreadsheets, static, specific and general redactions are performed.

Phase 2: A general contract agreement text document has been taken and data that follows structured format such as date and address have been redacted. Additionally, dynamic data redaction has been employed to find client name and redact it

when found. The only assumption here is that the requirement is to redact address, date, fee per hour and client name.

IV. Experiment Procedure

Phase 1: The data from the excel sheet is imported, and based on the project number that the viewer belongs to those rows of information are selected first (specific). Now fields like SSN (columns) are redacted irrespective of the viewer (general). After redaction is performed, the SSN number is no longer seen, it is replaced with a string. This is called full redaction where the data is removed completely. Partial redaction has also been implemented where in parts of the field are still visible to the viewer such as in our prototype, last four digits of account number is made visible and the rest of the number is replaced with a string.

Phase 2: The text document is imported. Then fields like date, address are redacted by identifying their structure. The data with structure xx/xx/xxxx tells that it is a date and has to be redacted. Similarly to recognize the address, all the state names are stored in an array and when text such as xxxxxxxxxxxxxxx, XX (state name abbreviation) appears, it is identified as the address and then redacted. Dynamic redaction has been employed using Naive Bayesian classifier. This classifier is used to identify a phrase of words as a valediction. If the phrase falls under the valediction class, it is totally redacted. A dataset consisting of possible valedictions which are copied multiple times to produce a huge dataset is fed as an arraylist input to the classifier and it is trained. Now, every classifier has an optimal critical value. The classifier is tested over several critical values and an optimal critical value is selected where the validation efficiency is the highest. All this is done to make sure the classifier predicts if the given phrase is a valediction. The entire algorithm focusses on identifying the phrase in the text document which is valediction. Correct results were obtained at a critical value of 8.34. A huge dataset is considered so as to satisfy the principle that a Classifier trains best on huge data set.

V. Experiment Results

The requirements and problem domain discussed in previous sections have been addressed completely. The results have been interpreted for a sample of data.

Phase 1: Table 1 and 2 show the two entities. Table 3 is the view generated with redacted data when manager 1 is viewing. It can be observed that since manager of project 1 is viewing, only the employee data who works under project 1 are being displayed. On the other hand, from the table 4 it can be observed that all the managers' details are being displayed with only specific fields getting redacted. Also partial redaction has been performed on account number showing last 4 digits. The full redaction has been performed on SSN.

TABLE I
Manager Entity

Name	Project Number	SSN	Contact Number
John	1	795124589	9493215896
Bryan	2	784128951	9069478251

TABLE II
Employee Entity

Name	Project Number	Account Number	SSN	Contact Number
Bob	1	623598745623	987456321	9845632145
Michael	2	984756114532	984562314	9856892563
Lisa	1	325987412253	745963214	9641062583

TABLE III
View for manager 1

Name	Project Number	Account Number	SSN	Contact Number
Bob	1	********5623	*********	9845632145
Lisa	1	********2253	*********	9641062583

TABLE IV
View for any employee

Name	Project Number	SSN	Contact Number
John	1	*********	9493215896
Bryan	2	*********	9069478251

Phase 2: The imported text document is searched for structured formats such as dates and address first and they have been redacted as shown in the figures below. Also, the valedictions are identified and the phrases following them are redacted.

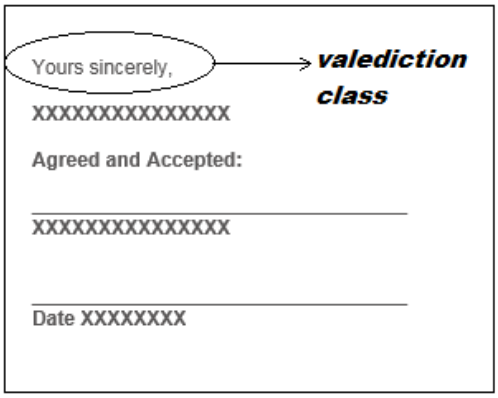

Figure 1. Redaction of date and client name

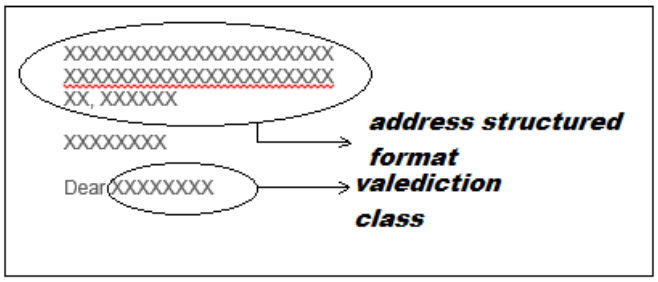

Figure 2. Redaction of address and client name

In this way, a semi-automatic approach of data redaction has been achieved.

VI. Conclusions and Future Scope

The attempt of performing intelligent data redaction was accomplished. In this paper the dynamic quality of redaction is achieved using the Bayesian classifier but has been implemented only for valedictions. A classifier like this could be incorporated for several other classes and their automatic recognition. Machine learning though generates good and amazing results, is hectic and time consuming task to employ. The scope of this application is not universal where as it requires customization at every step such as training, searching for critical value, etc. Artificial intelligence and supervised learning provide with innumerous techniques to induce the quality of automatic detection. Once the detection is achieved, that class of data can be redacted. Thus, complete automatic data redaction given a text document would really be a challenge.

VII. References

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

[2] 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.

[3] 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.

[4] Chad Cumby, Rayid Ghani, "A Machine Learning Based System for Semi-Automatically Redacting Documents", Accenture Technology Labs, 161N. Clark St, Chicago,Illinois 60601.

[5] V. Chakaravarthy, Himanshu Gupta, Prasan Roy and Mukesh K. Mohania; "Efficient Technique for Document Sanitization"; IBM India Research Lab.

[6] 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"

[7] Thomas Neubauer and Mathias Kolb; "An Evaluation of Technologies for the Pseudonymization of Medical Data".

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

[9] David George Kelly and Brent Russell Foster; "Process of Electronic Document Redaction"; Onstream Systems Limited.

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

[11] Rennie, J. D. M.; Shih, L.; Teevan, J.; and Karger, D. R. 2003. Tackling the poor assumptions of naive bayes text classifiers. In Proceedings of ICML-2003.

[12] Rennie, J. D. M.; Shih, L.; Teevan, J.; and Karger, D. R.2003. Tackling the poor assumptions of naive bayes text classifiers. In Proceedings of ICML-2003.

[13] C. Karat, J. Karat, C. Brodie and J. Feng. Evaluating interfaces for privacy policy rule authoring. CHI 2006.

[14] C. Johnson, III. Memorandum M-07-16, "Safeguarding against and responding to the breach of personally identifiable information". FAQ. May 22,2007.

[15] RapidRedact. http://www.rapidredact.com/

[16] L. Sweeney, Information Explosion. Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies, L. Zayatz, P. Doyle, J. Theeuwes and J. Lane (eds), Urban Institute, Washington, DC, 2001.