RC25055 (W1007-006) July 6, 2010
Computer Science

A Text Mining Approach to Confidential Document Detection for Data Loss Prevention

Youngja Park
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 704
Yorktown Heights, NY 10598

A Text Mining Approach to Confidential Document Detection for Data Loss Prevention

Youngja Park
IBM T. J. Watson Research Center
P.O. Box 704, Yorktown Heights, NY 10598, USA
young_park@us.ibm.com

Abstract

Data loss prevention (DLP) systems aim to automatically detect and protect confidential or sensitive information in an organization, for example when it is accidentally leaked by email. Current state-of-the-art DLP systems employ rudimentary content analysis techniques such as regular expression matching to detect sensitive content. The detection accuracy of these current approaches remains very limited for unstructured text, due to the high level of ambiguity and idiosyncracy in human languages. In this paper, we propose problem-specific text mining techniques to assess the sensitivity of documents. Our case study on a corpus of more than 900 confidential documents shows that a lightweight classier with problem-specific features outperform existing methods by at least 10 percentage points.

1 Introduction

Recently, the computer security community has acknowledged the importance of data-centric security with the emphasis to protecting data directly. Data Loss Prevention (DLP) refers to systems that identify, monitor, and protect confidential or sensitive information through deep content inspection [1, 2]. DLP systems are divided based on the status of data into "endpoint DLP" for data at rest (e.g., stored in data storage) and in use (e.g., copy and print actions), and "network DLP" for data in motion (e.g., email or internet transmission). Due to growing concerns on data breaches and increasing requirements from regulatory compliances such as HIPAA (Health Insurance Portability and Accountability Act) [3] and PCI DSS (Payment Card Industry Data Security Standard) [4], DLP has drawn increasing attention from both businesses and academic community.

DLP systems typically apply content inspection to identify confidential or sensitive data, and contextual analysis of transactions (e.g., information about the sender and the recipient, transmission method and time, etc) to prevent unauthorized use or transmission of sensitive information. Since most of data in an organization is published as unstructured free text, accurate analysis and detection of sensitive information in unstructured text is crucial for DLP. However, automatic understanding of unstructured text is notoriously difficult, and computers are far from reaching human-level understanding. The low accuracy is caused mainly by the high level of ambiguity in human languages (e.g., "Cookie" and "classified"), and many different representations of a same concept (e.g., "IBM", "International Business Machines", "Big Blue", etc).

Most current DLP systems employ pattern matching using dictionaries of predetermined terms or regular expressions. More advanced DLP systems apply machine learning-based classification technologies based on document concepts. While both approaches are quite successful for certain applications such as personally identifiable information (PII) detection and document topic categorization, they show limited success for more complicated security applications such as confidential document detection.

In this paper, we show that advanced text mining technologies designed specifically for the given problem achieve higher accuracy. We conduct a case study for identifying confidential documents in an enterprise, and demonstrate that a problem-specific set of features greatly improves the accuracy. In addition, our system generates the confidence of the decision (i.e., the probability of the document being confidential) so that a DLP system can apply more flexible security actions based not only on the binary decision but also on the confidence level of the classification.

2 Challenges in Confidential Document Detection

Automatic identification of confidential documents and enforcement of appropriate protections on these documents are critical for enterprises to protect their intellectual properties. An important task in building an automated system for confidential document detection is to understand what constitute confidential documents. A document can be designated as confidential for a variety of reasons including:

- Contains sensitive entities such as personal, financial, or medical information of employees or customers
- Contains sensitive topics such as human resource and legal documents
- Contains new product development plans
- Contains future acquisition plans
- Contains un-announced financial reports
- Contains proprietary source code

As we can see, confidential documents contain a variety of topics, and, thus, standard information extraction or topic-based classification approaches do not work well for this task.

We note that most government organizations and companies use a predefined set of words and phrases to explicitly specify confidential documents such as "Top Secret", "Classified" and "Confidential", to name a few. When a document contains such a label, it should be treated as confidential unless the document is very old that the confidentiality has expired in accordance with the company's policy. Therefore, documents containing such labels can be a good starting point to identify confidential labels.

Detecting confidential documents using a precompiled list of confidential labels seems trivial at first hand. Some may argue that one can simply search confidential labels in certain locations in a document, such as the title page, footer or header; and if a document contains a label, the document is considered as confidential. This method may work well for documents such as powerpoint slides and source code. Authors of these documents tend to put confidential labels in the header, footer or the beginning of a document.

However, this straightforward method is not satisfactory for two reasons. Firstly, the labels are not fixed, but are very diverse both lexically and syntactically as demonstrated by the example labels shown in Table 1. Furthermore, the labels often used in narrative sentences in many documents. They are not written in stand-alone phrases but included in sentences as in "This is an IBM Confidential document and should not be shared.". It is not feasible to precompile and write patterns to recognize all possible expressions.

Secondly, identifying the likely locations in documents is not a straightforward problem. Document structure information is not consistent across different document formats, and, thus, rules relying on the location information can be very vulnerable to small changes.

Furthermore, this method is incomplete because many documents are often duplicated entirely or partially. In many cases, the duplicated version may not contain the confidentiality label any more, and thus can not be detected by a system relying on a predefined confidential labels. If the original document was correctly identified as a confidential document, one can infer the duplicated version is also confidential based on its content similarity to the original document.

IBM Confidential
An IBM Confidential study
IBM and *aPartnerName* Confidential
Private & Confidential
IBM Proprietary and Confidential
Internal Use Only
IBM Use Only
This document is IBM and IBM Business partner use only
Do Not Distribute (Reproduce/Forward/Disclose)
Please do not disclose this information to non-IBM employee
Should not be disseminate (shared outside/disclosed)
Not for External Distribution
This document is not intended for customer distribution or use with customers
The information contained in this document is confidential or protected by law
For educational purposes only; do NOT share with Customers or any Business Partner not covered by an NDA

Table 1: Examples of phrases and sentences used to specify IBM confidential documents. *aPartnerName* indicates a business partner of IBM and is used to anonymize the identity.

More advanced DLP systems apply a machine learning-based classification based on the concepts or topics represented in the documents. Most systems in this category represent documents as vectors of the words that appear in the training document set (a.k.a., 'bag of words' representation). This approach has been very successful for topic categorization of texts such as news articles or web pages [5, 6, 7]. However, this approach also has a limited value for DLP scenarios. For instance, a document containing personally identifiable information is not sensitive from a DLP perspective, if the information belongs to a publicly well-known person. Similarly, a patent document is not confidential if the patent has been filed.

3 Problem-specific Text Mining Approach

Based on the observations described in Section 2, we propose a customized system for confidential document detection.

3.1 System Overview

Our approach comprises two steps for confidential document detection as depicted in Figure 1.

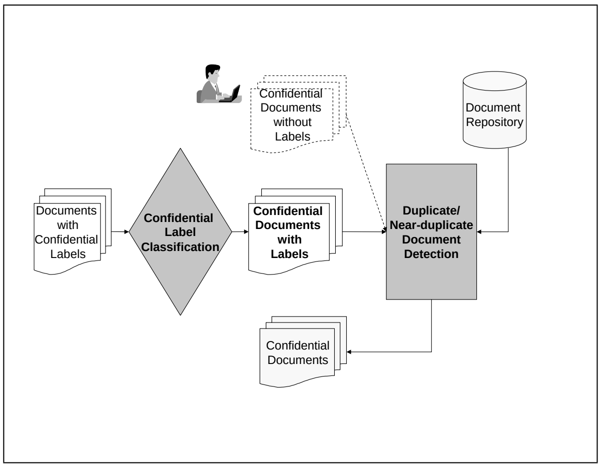
Confidential Documents without Labels
Document Repository
Documents with Confidential Labels
Confidential Label Classification
Confidential Documents with Labels
Duplicate/ Near-duplicate Document Detection
Confidential Documents

Figure 1: High-level Steps for Confidential Document Detection

At the first step, we collect documents containing one or more confidential labels, and apply a machine learning-based algorithm to determine if the labels indeed specify the confidentiality of the given document or are used for different meanings. Please note that people often use very diverse and complex phrases as shown in Table 1, making the detection of such labels non-trivial. In addition, these words and phrases are often used for different contexts as in 'When should you use "IBM Confidential"?', and thus simple detection of such labels is not enough. The identification and classification of confidential labels in unstructured text is the focus of the paper, and we will describe our approach in much details in Section 4.

The second step is to find duplicate/very-similar documents to the confidential documents using duplicate document detection technologies. A number of duplicate document detection schemes have been studied in the web search community [8, 9, 10, 11, 12, 13]. Duplicate and near-duplicate document detection technologies can be categorized into two different groups. One approach is based on lexical similarity between two documents, i.e., measuring how many words a given pair of documents use in common. This method can determine the similarity with high accuracy, but is not feasible for a large volume of document collection, especially. for a real-time application like a network DLP scenario. The other approach uses signatures or document fingerprints instead of words to compute the similarity. This method is much faster than lexical similarity-based method, but very brittle to small changes. We are currently investigating a more robust fingerprinting technology for textual data, which can produce a same fingerprint for lexically and semantically very similar words.

4 Confidential Label Classification

The proposed method consists of the following three steps.

1. Identify canonical forms of confidential labels

2. Classify each label instance in a document to decide if the instance specifies that the document is confidential

3. Based on the classification results at Step 2, classify the document into 'Confidential' or 'Not-Confidential'.

'classified', 'restricted', 'secret', 'top secret', 'confidential', 'ibm confidential', 'internal use only', 'do not distribute', 'do not share', 'do not forward', 'do not disseminate', 'do not reproduce', 'do not disclose', 'not be distributed','not be shared', 'not be forwarded', 'not be disseminated', 'not be reproduced', 'not be disclosed', 'ibm confidential restricted', 'registered ibm confidential', 'highly confidential', 'strictly confidential'

Table 2: Confidential cue phrases used in this study. The cue phrases are widely used in IBM and government organizations.

4.1 Confidential Label Identification

In this work, we compiled a set of canonical cue phrases widely used confidential labels at IBM and at the government. For instance, cue phrase "IBM Confidential" is used in the first three labels. The full list of the confidentiality cue phrases used in this work are listed in Table 2. The first step toward confidential document detection is to recognize all the instances of the cue phrases in a document. In this work, we apply pattern-based matching for this step. Please note that it is very easy to modify the list, and changes in the list do not cause performance change because the label itself is not used in classification as described in the next section.

4.2 Individual Label Classification

For each confidential label instance, we determine if the instance specifies that the document is confidential, or it is used in a different context. Unlike the document concept-based approach described in Section 2, we base the decision on local context and extrinsic information about the file.

Based on the analysis of the sample documents, we selected the following 13 features to predict the confidentiality of individual label. The features represent lexical, contextual and extrinsic information about the label and the document. As we can see, the features are very general and domain independent, and therefore the system can easily be applied to a new domain.

- Label Type: We categorize all confidential labels into three groups. The first group contains one word labels such as 'classified', 'confidential', 'secret'. The second label group contains labels with a negation word such as 'do not distribute', and 'not for external distribution'. All other labels belong to the third group. Use of label type instead of the labels themselves makes the system more robust and domain-independent. For instance, if a user creates a slightly variation of an existing label, a classifier relying on the label will fail to classify the document correctly.

- Capitalized: The feature indicates if the label is written in all-capitalized form (e.g., CONFIDENTIAL), initial capitalized form (e.g., Confidential), or all lower case (e.g., confidential).

- Negated. A boolean value denoting if a label is negated in the context. We analyze 5 words on the left and 3 words on the right from the label to decide if the label is negated.

- Location in the document; "Top", "Middle", or "Bottom". When the label appears 10th percentile from the beginning of the document or from the bottom of the document, the feature value is "Top" and "Bottom" respectively. Otherwise, the value is "Middle".

- Location in the sentence; "LEFT", "CENTER" or "RIGHT". When a sentence starts with a label, the location of the label is "LEFT". When a sentence ends with a label, the location of the label is "RIGHT". When a label appears in other locations, the location value is "CENTER".

- Part of speech of 6 surrounding words: part of speech information for three words on the left and three words on the right. These information provides additional knowledge about the use of the label in the sentence.

- Document format: The file format of the document such as "doc", "html" and "pdf". Software code files such as ".c", ".java" are consolidated as one format.

- Document Age: The year duration of the document since it was created.

In this work, we apply support vector machines (SVMs) for classification [1]. The main idea of SVMs is to find a hyperplane which splits the positive examples from negative examples with the largest distance in between the two example sets [14]. SVMs have been successfully applied to many classification and regression tasks such as text categorization [5] and pattern recognition [15]. In this work, we use C-support vector classification (C-SVC) with a radial basis function (RBF) kernel [14]. C-SVC solves the following problem:

$$\min_{\mathbf{w},b,\xi} \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{l}\xi_i$$

$$\textit{subject to} \;\; y_i(\mathbf{w}^T\phi(\mathbf{x}_i)+b) \geq 1-\xi_i,$$

$$\xi_i \geq 0, i = 1, \ldots, l$$

given training vectors $\mathbf{x}_i \in R^n, i = 1, \ldots, l$ and an answer vector $\mathbf{y} \in R^l$.

The SVM algorithm can also produce the probabilities of an object belonging to each class in the target class set. For instance, the SVM's output for label "Internal Use Only" may be 'Confidential (0.97)' and 'NotConfidential (0.23)'. We take the class with the highest probability (i.e., confidence) as the classification result.

4.3 Confidential Document Classification

After making decisions on all confidential label instances in a document, we apply a voting method to classify the document. The final decision is made as follows. If the document contains only one label, the decision for the document is same as the classification

[1] We also conducted experiments with Logistic Regression, but SVM performed better

result for the label. When there are multiple labels, we discard all labels with the classification probability lower than 0.6 (i.e., the classification result with low confidence), and then take the majority class as the final class for the document.

5 Experimental Results and Performance Evaluation

In this section, we describe the experimental results and compare the performance of our system with two state-of-the-art approaches.

5.1 Experimental Data

We collected 957 documents of various formats including HTML, PDF, DOC, PPT and XLS, which contain at least one confidential cue phrase. The HTML documents are downloaded from an IBM internal wiki. Table 3 shows the counts of documents by document format.

Document Format	Count
html, txt	677
ppt	152
doc, lwp	49
pdf	47
source code	19
xls	13

Table 3: Experimental data by file formats. 'lwp' is a document format from Lotus. The source code column includes .c, .h, .css, .java, .sh and .sql and .tag files.

We then manually inspected the documents and assigned each document to 'Confidential' or 'NotConfidential' class based on the content. 565 documents (i.e., 59%) were assigned to 'Confidential', and 392 documents were assigned to 'NotConfidential'. The hand-labeled data set enables us to quantitatively measure and compare the accuracy of automatic systems.

5.2 Performance Evaluation

We conducted experiments with the 957 documents and performed 10-fold cross validation [2] using stratified sampling on document format. Therefore, the experimental data set is partitioned into the training set and the test data containing 90% and 10% of the experimental data respectively, and the constituents of the training and test sets are representative of the entire data population.

We also conducted the same experiment using the two state-of-the-art methods introduced in Section 2. For the pattern matching-based approach, we built a pattern matching-based classifier which classifies a document as confidential if a confidential label appears within the first 15 sentences of the document (i.e., the window can cover the title page, the header, the footer sections).

For document concept-based classification, we used a commercially available text classification system. The classifier uses Naive Bayes classification algorithm and extracts topical words from the documents as features. The system has been successfully used for automatic email message routing and document organization.

Table 4 compares the performance of different approaches. As we can notice from the results, document concept-based approach shows higher accuracy than the pattern matching approach or the majority class approach. Our approach significantly outperforms the other approaches both in overall accuracy and precision and recall for identifying confidential documents.

6 Conclusions and Future Work

This paper argues that a problem-specific text mining technique greatly improves the accuracy of DLP systems. Specifically, we described our case

[2] K-fold cross-validation is a technique to estimate how a predictive model will perform in practice. The process splits the data set into K subsets; each subset is held out in turn as the evaluation set, and the remaining (K-1) subsets are used for training. The average accuracy over the K rounds is used to predict the accuracy of the system.

	Accuracy	Precision	Recall
TM	86.4%	90.6%	86%
Concept	77.78%	81.96%	81.2%
Pattern	57.16%	57.16%	100%
Majority	59%	n/a	n/a

Table 4: Comparison of the performance of the three approaches. **TM** indicates our approach, **Concept** denotes the document concept-based system, and **Pattern** indicates the pattern matching-based system. **Majority** is an artificial classifier which assigns all documents to the majority class, i.e., "Confidential" for the experimental data set.

study on detecting confidential documents in various formats, and demonstrated that existing methods based on pattern matching and topic classification are not sufficient even for this seemingly straight forward problem. Our experimental results showed that a text mining technique designed for a given problem can perform the task with a high level of accuracy. The presented system can be used to build a repository of "known" confidential documents.

Future research topics are in the following.

- Text mining technologies for network DLP

Content inspection and analysis for network DLP scenarios pose two challenges. Firstly, it has to be done in real-time. The main focus of the text mining community has been improving the accuracy, and little attention has been paid to the development of real-time text mining techniques. Secondly, it must produce a high accuracy with partial content information. In network DLP scenarios, we don't have access to the entire document, but only to a fraction of the document. Therefore, the traditional text mining relying on the document content can not be applied.

- Usability study for DLP systems

The accuracy of DLP system is an important factor for the system to be adopted by the users. Another very important factor is the usability of a DLP system. If a DLP system is too intrusive to a user, the user may stop using the system. Understanding appropriate protection mechanisms for different scenarios, and users' reaction to the system can help us design a more usable DLP system.

- Development of advanced text mining systems for different DLP applications

There are many other DLP applications which require more customized text mining technologies for instance detection of documents with HIPAA-related information. This example task has been also regarded as a simple pattern-matching task, i.e., 'find medical condition or treatment names and personally identifiable information in a document'. However, to be more accurate, we need to determine if the medical condition mentioned in the document belongs to the person whose personal information is revealed in the document. This level of analysis requires more sophisticated relationship extraction.

References

[1] R. Mogull, "Dlp content discovery: Best practices for stored data discovery and protection," *http://www.emea.symantec.com/discover/downloads/DLP-Content-Discovery-Best-Practices.pdf*, 2008.

[2] Wikipedia, "Data loss prevention software," *http://en.wikipedia.org/wiki/Data_loss_prevention_software.*

[3] U. Congress, "Health insurance portability and accountability act (HIPAA)," 1996.

[4] P. Security Standards Council, "PCI DSS," *https://www.pcisecuritystandards.org/security_standards/pci_dss.sh* 2004.

[5] T. Joachims, "Text categorization with support vector machines: learning with many relevant features," in *Proceedings of ECML-98, 10th European Conference on Machine Learning*, pp. 137 – 142.

[6] D. D. Lewis, R. E. Schapire, J. P. Callan, and R. Papka, "Training algorithms for linear text classifiers," in *Proceedings of the 19nd annual international ACM SIGIR conference on Research and development in information retrieval*, 1996, pp. 298–306.

[7] F. Li and Y. Yang, "A loss function analysis for classification methods in text categorization," in *Proceedings of the Twentieth International Conference on Machine Learning*, 2003, pp. 472–479.

[8] A. Broder, "On the resemblance and containment of documents," in *Proceedings of the Compression and Complexity of Sequences 1997*, 1997.

[9] A. Chowdhury, O. Frieder, D. Grossman, and M. McCabe, “Collection statistics for fast duplicate document detection,” in *ACM Transactions on Information Systems*, vol. 20, 2002, pp. 171–191.

[10] G. Forman, K. Eshghi, and S. Chiocchetti, “Finding similar files in large document repositories,” in *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, 2005, pp. 394–400.

[11] A. Kolcz, A. Chowdhury, and J. Alspector, “Improved robustness of signature-based near-replica detection via lexicon randomization,” in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2004, pp. 605–610.

[12] P. Lakkaraju, S.Gauch, and M. Speretta, “Document similarity based on concept tree distance,” in *Proceedings of the nineteenth ACM conference on Hypertext and hypermedia*, 2008, pp. 127–132.

[13] G. Forman, K. Eshghi, and S. Chiocchetti, “Local text reuse detection,” in *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, 2008, pp. 571–578.

[14] V. Vapnik, *The Nature of Statistical Learning Theory*. New York: Springer-Verlag, 1995.

[15] C. Burges, “A tutorial on support vector machines for pattern recognition,” *Data Mining and Knowledge Discovery*, no. 2, pp. 121 – 167, 1998.