Document Security Identification Based on Multi-classifier

Kaiwen Gu(✉), Huakang Li, and Guozi Sun

School of Computer Science and Technology, School of Software,
Nanjing University of Posts and Telecommunications, Nanjing, China
mynamekevin@qq.com, {huakanglee,sun}@njupt.edu.cn

Abstract. Data leakage is a potentially important issue for businesses. Numerous corporate offer data loss prevention (DLP) solutions to monitor information flow, and detect such leakage. Adding a secret label to a document, DLP can use documents label to do securely control, effectively protecting data. With the increasing documents every day, manual labeling is time-consuming. To better solve the difficult task, recently researchers need to start use document security identification by machine learning quickly classify a large number of texts. The contribution of this paper is to explore dimensionality reduction by feature selection and combine two models to avoid the process of weighting different type of features. In contrast to training all features with one algorithm, our experimental results demonstrate that the combination of two models can improve the classification performance.

Keywords: Data leakage prevention · Document security identification · Feature selection · Machine learning · Model combination

1 Introduction

With the development of information technology, corporate security threats are becoming increasingly diverse. Data leakage can be divided into external leakage and internal leakage. In recent years, most significant data leakage incidents are caused by internal network security, such as legitimate users, have access to database information and spread to other companies. DLP [1, 3] is an important way to address detection and protection of data leakage. An effective way in DLP is to label the data files according to some sensitive detection, and then mapping document labels into visitors rank. However, an important issue is that if the manual set the document label error, will once again cause information leakage seriously. At the same time, as the number of enterprise documents continue to increase, and manual work needs to spend a lot of manpower and time, so this task is very important for information security and management efficiency of the company.

Security text classification is clearly a solution to the effectiveness of data leakage security method. Recently security classification for DLP purpose is supported in some techniques like fingerprinting of documents, keyword matching and regular expressions. Machine learning [5, 7] has become an important method for text classification. In this paper, we define security labels into three levels: top-secret - the highest level of risk,

J. Abawajy et al. (eds.), *International Conference on Applications and Techniques in Cyber Security and Intelligence*, Advances in Intelligent Systems and Computing 580,
DOI 10.1007/978-3-319-67071-3_18

confidential - medium level and internal - lowest level. Different operations in DLP for these three levels are shown in Fig. 1.

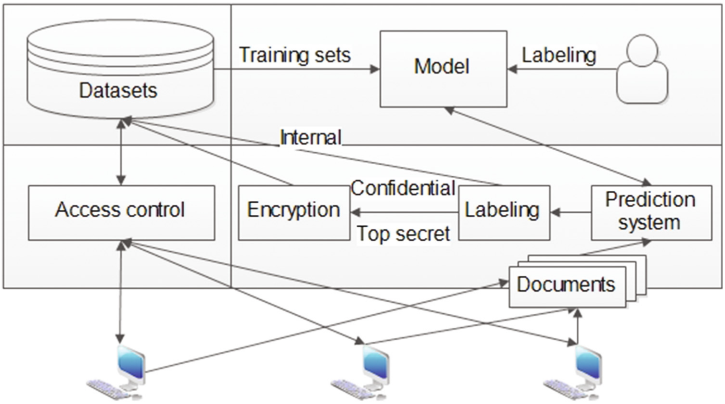

Fig. 1. After labeling un-label documents, internal document upload dataset directly, confidential and top-secret documents are encrypted before uploading.

Following previous works in [1, 3], our work is focused on how to automatically extract features from full-content. The space vector model (VSM) expresses the text information in the form of bag of words (BOW), but loss the semantic relation. Besides, features extracted from different part of text content is hard to weight. The contributions of this paper include: demonstrate our security text classification system is applicable to large documents in DLP to prevent data leakage; analysis the role of dimensionality reduction methods; extract feature from contents-based and security-based, training two type of features with different algorithm, combination results prove the state of the art method.

The remainder of this paper is structured as follows: Sect. 2 propose architecture and feature extraction method for the secret text; Sect. 3 follow the process of the classification of the class text to show experience results of our evaluations. Related work, conclusions and future work are discussed in Sect. 4.

2 System Architecture

The process framework contains three steps: the first step is text representation which imports the text into numerical features the algorithm can be identified, including preprocess, feature extraction, feature selection. Second step is training and evaluation. The results of the evaluation can be used to adjust the parameters and model. Third step is using the model to predict test set. The framework overview of our system process is shown in Fig. 2 left. More details of every part will be discussed in the rest sections. According to the importance of text information, we extract two types of textual features: security-based features (SBF) and content-based features (CBF). Because the different characteristics and importance of them, we use two algorithms to train, and finally do combination with two results as the final result (see the framework in Fig. 2 right).

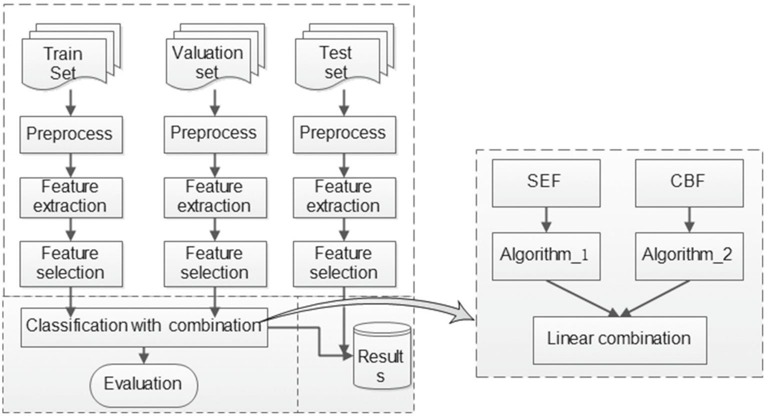

Fig. 2. The framework of security text classification.

3 Experience and Result

3.1 Dataset Preprocess

Our experience dataset is provided from Jiangsu Agile Technology Co., Ltd which leading data file system in encryption and control for large corporations in China. We choose three companies collections because they contain a mix of three security ranks. After removing empty, highly similar documents and documents with 30 words or less, we end up with 2270 documents in total, the dataset statistics show in Table 1.

Table 1. Documents dataset statistics

Datasets	Total	Top-secret	Confidential	Internal
Corporation1	965	283	308	374
Corporation2	738	193	242	303
Corporation3	567	173	125	269
Total	2270	649	675	946

Since the Chinese texts are different from other languages such as English, and the text contains many proper noun, preprocessing is an important step in classification. We use open source Jieba to cut words. All words cut by Jieba will be candidate features expect stop words which are insignificance. For domain associated terms(DATs), such as proper nouns, we define more than 2000 domain associated terms in our Chinese domain knowledge dictionary (CDKD) to achieve a more precise word division.

3.2 Feature Extraction

In contrast to the general method of feature extraction, the text representation model is divided into two parts. The first part of SBF including document title, the first paragraph, the end paragraph of the text, and the DATs. Title, the first and last paragraph are generally full text of the sentence, represents the higher level of secret characteristics than full text content. DATs are features we achieve from CDKD. They may have some association with sensitive information. The second part of features is content-based features. We first use common bag of words model, that is, a word as a feature, so that a text can be expressed with the Vector Space Model (VSM). But VSM has lost the context order, we also add the bigram and trigram feature to the second part.

To compare the method of feature extraction, we set several subtasks that training by the same algorithm Support Vector Machine (SVM). For each subtask (we remove one type of feature extraction), the system is automatically chosen the best performance from validation dataset. The average F1-measure in three training sets is shown in Table 2. When we remove unigram and bigram from CBF, the result has a certain degree of decline except trigram. If we remove features from title, DATs and first last paragraph respectively, the result also has some decline. So we believe these features have positive effect on the experience.

Table 2. Performance of experience by feature extraction

Features	Avg-F1	Descend rank	Features	Avg-F1	Descend rank
CBF	78.0%		SBF	74.3%	
CBF-unigram	73.2%	1	SBF-title	73.1%	2
CBF-bigram	76.5%	2	SBF-DATs	73.8%	3
Total-trigram	78.0%	3	SBF-para	72.8%	1

3.3 Security Feature Selection and Dimensionality Reduction

Due to the number of CBF cause excessive dimension disaster, makes the model computational complexity and not conducive to industrial, dimension reduction become an important step. In this paper, the feature selection method is based on unsupervised TF-IDF (term frequency - inverse document frequency) and label-based Chi-Square (χ^2).

Compared with the weight calculated by the word frequency, TF-IDF model can effectively exclude the interference of such high frequency words. χ^2 is used to determine whether there is a significant difference between the expected frequencies and the observed frequencies in one or more categories. The sum of quantity over all of the features is the test statistic.

For the above two methods, we designed four contrast methods, TF-IDF, χ^2, first TF-IDF then χ^2 and first χ^2 then TF-IDF in the four groups of experiments, and search for the optimum feature size by grid search method. We find select top 20–24% features by TF-IDF method or top 31–34% features by χ^2 method when the highest F1 value is 79%. When combine two methods can improve the accuracy of classification of text classification. First select top 81–84% by TF-IDF, and then select top 42–45% by χ^2 when F1 up to 85%, so we choose to first use TF-IDF and then use the χ^2.

3.4 Model Combination and Evaluation Result

Some machine learning algorithms have achieved great success in text categorization such as Naive Bayes, Support Vector Machine. For security text classification task, the feasibility of these classifiers is proved by [6] et al. Support Vector Machine (SVM) can efficiently perform a non-linear classification using what is called the kernel trick. Our experience demonstrate linear kernel has better performance than other kernels such as RBF kernel, polynomial kernel.

Our final system is merging two algorithm results. The Naive Bayes train first part features SBF and SVM train second part features CBF. The method is shown as follows. For the probability of three ranks of a text x, we have

$$\text{probValue}^{rank_i}(x) = \lambda P^{rank_i}_{Naive\ Bayes}\left(x_{SBF}\right) + (1-\lambda)P^{rank_i}_{SVM}\left(x_{CBF}\right) \quad (1)$$

where $P^{rank_i}_{Naive\ Bayes}$ and $P^{rank_i}_{SVM}\left(x_{CBF}\right)$ are the $rank_i$ probability value used Naive Bayes and SVM respectively in two type of features SBF and CBF.

We combine two models, Navies bytes training SBF and SVM training CBF with linear combination. The formula's parameter λ can be searched by cross validation. In Fig. 3b we could find the best performance when λ equal to 0.4. The Fig. 3a shows the performance of Naive Bayes training all features, SVM training all features, and linear combination with best λ. The combination model always performs better than Naive Bayes model and SVM model.

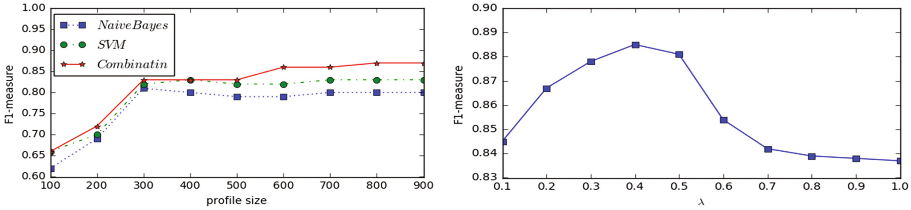

Fig. 3. Experience result: (a) Comparison of two models and combination model. (b) Search the best parameters.

4 Related Work and Conclusion

Some research has focused on the automatic security classification. In [3], their aim is to using methods from machine learning and information retrieval to detect misclassification. Paal E [2, 6] consider about dimension reduction and performance improvement, the accuracy drops to only around 74% with 18 words by lasso. This paper method is more practical with the combination of existing methods. For text presentation, Sultan [4] add common N-gram to category, the percentage of correct classification increased from 78.8 to 85% after modification. Khudran [1] through pruning procedure to improve performance of algorithms while reducing training set sizes, but not clear whether eliminate paragraphs would lead to better performance.

This paper explores the method of using the text categorization method to label the secret text on DLP. We propose a method of extracting two kinds of features, and make a combination with two model results. In the future, we intend to detect a class of documents that re-edit and use the same template, and do in-depth classifications of such documents. We also intend to use the word embedding to pre-train document features and then use convolution neural network training to compare.

References

1. Alzhrani, K., Rudd, E.M., Boult, T.E., et al.: Automated big text security classification (2016)
2. Engelstad, P.E., Hammer, H., Kongsgard, K.W., et al.: Automatic security classification with lasso. In: International Workshop on Information Security Applications, pp. 399–410. Springer International Publishing (2015)
3. Kongsgard, K.W., Nordbotten, N.A., Mancini, F., et al.: Data loss prevention based on text classification in controlled environments. In: Information Systems Security, pp. 131–150. Springer, Berlin (2016)
4. Alneyadi, S., Sithirasenan, E., Muthukkumarasamy, V.: Word N-gram based classification for data leakage prevention. In: 2013 12th IEEE International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom), pp. 578–585. IEEE (2013)
5. Hammer, H., Kongsgard, K.W., Bai, A., et al.: Automatic security classification by machine learning for cross-domain information exchange. In: Military Communications Conference, Milcom 2015, pp. 1590–1595. IEEE (2015)
6. Engelstad, P.E., Hammer, H., Yazidi, A., et al.: Advanced classification lists (dirty word lists) for automatic security classification. In: 2015 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC), pp. 44–53. IEEE (2015)
7. Sebastiani, F.: Machine learning in automated text categorization. ACM Comput. Surv. (CSUR) **34**(1), 1–47 (2002)