Analysis and Comparison of Fax Spam Detection Algorithms

Jaekwang Kim
School of Information and Communication Engineering,
Sungkyunkwan University, Suwon,
440-746, Republic of Korea
linux@skku.edu

Hyoungshick Kim
School of Information and Communication Engineering,
Sungkyunkwan University, Suwon,
440-746, Republic of Korea
hyoung@skku.edu

Jee-Hyong Lee
School of Information and Communication Engineering,
Sungkyunkwan University, Suwon,
440-746, Republic of Korea
john@skku.edu

ABSTRACT

Spam detection is one of the important problems in these days. Many spam detection methods were proposed, but fax spam detection is not popular. It not easy to directly use existing content-based spam detection methods for fax documents because the documents are processed as image rather than text. In this paper, we propose a fax spam detection framework which is based on keyword patterns by using an Optical Character Recognition (OCR) technique. To demonstrate how effective the proposed framework is, we analyze and compare three fax spam detection algorithms (rule based method, SVM based method, and naïve Bayesian based method) with 219 normal and 212 spam documents. Our recommendation is to use naïve Bayesian based method which is capable of achieving an accuracy of 92.49%.

CCS Concepts

• Security and privacy → Intrusion/anomaly detection and malware mitigation → Intrusion detection systems

Keywords

Fax Spam, Comparison, Detection, Rule Based Filtering, SVM, Naïve Bayesian.

1. INTRODUCTION

Fax is a data transfer technique which is used for a long time. Even it is not the newest but popular for transferring data. With its popularity, fax spam also exists frequently. Many spam detection methods were proposed; however, fax spam detection is not easy. Fax documents are processed as image. Therefore, we cannot use existing content-based spam detection methods with the keywords in a document.

Even though we recognize the words or characters in the fax documents, the number of words in a fax document might relatively be smaller than the number of words in e-mail documents, which makes existing spam detection algorithms harder. This paper proposes a fax spam detection framework that uses common keywords used in legitimate and spam documents being received via fax machines. With this framework, we analyze and compare three fax spam detection algorithms; rule based method, Support Vector Machine (SVM) based method, and naïve Bayesian based method. To demonstrate the effectiveness of our approach, we experimented with three well-known machine learning algorithms (rule based method, SVM based method, and naïve Bayesian based method). For experiments, we used a fax document dataset (legitimate: 219, spam: 212 documents) collected from a real fax machine. The experimental results showed that the proposed machine learning based detection methods could be used to detect fax spam documents. In particular, naïve Bayesian based method produced the best results, which is capable of achieving an *F*-measure of 87.74%.

The paper is organized as follows: in Section 2, we will briefly summarize the existing spam detection algorithms. Section 3 presents our proposed method. We present a performance evaluation of our work in Section 4. Finally, Section 5 concludes the paper.

2. RELTED WORK

There are lots of researches which are detecting spam in various domains; e-mail, Sort Message Service (SMS), and Social Network Services (SNS) as well fax. The most popular and powerful approach is contents (or keyword) based text mining. A fax spam detection method is also similar with other domains. There are similar and different features. We want to refer some related works in this section.

E-mail spam detection has widely studied in recent years [1][2]. But it cannot be applied to fax spam detection, directly. It is because that the most of these researches are based on contents analysis and text mining but fax are dealing with the documents as images.

Widely using of smartphones, SMS and SNS are abused for spammers [3][4]. SMS and SNS spam detection are more similar to fax spam detection. Because these also have a small number of words in the documents. But there are little missing characters in the documents.

There are not so many researches in fax spam detection. Qing et al. proposed a method for spam fax detection and classification based on clustering. The method took advantage of behavioral characteristics of spam fax [5]. Image-based methods and Text-based methods are also proposed but showed not so good in detection performance. Muhammad et al. had studied on the effectiveness of spam detection technologies with large volumes of data [6]. They compared some methods such as whitelisting, blacklisting, email signatures and different machine learning methods for detecting spams. And they concluded that there was no 100% secure systems around the world which could handle this

IMCOM '17, January 05-07, 2017, Beppu, Japan
© 2017 ACM. ISBN 978-1-4503-4888-1/17/01…$15.00
DOI: http://dx.doi.org/10.1145/3022227.3022284

problem. But they missed some brilliant techniques such as SVM and Bayesian based methods.

In this paper, we are willing to solve practical fax spam detection problem using representative machine learning techniques.

3. FAX SPAM DETECTION ALGORITHMS

3.1 Overview of Fax Spam Detection

A fax system consists of as follows:

- A transmitting device – it translates the image data into electrical signals according to a set pattern such as bitmap
- A synchronized receiving device - it retranslates the electrical signals to a duplicate of the original image data and prints them.

The fax spam detection should be conducted at the receiving device. To detect spam messages, we considered three contents based mining methods; a simple rule based spam filtering [7][8], SVM based spam filtering, and Naïve Bayesian based spam filtering methods.

3.2 Fax Spam Detection Framework

In this paper, we prosed three steps for fax spam detection framework. One is a preprocessing with OCR scan and morphological analysis. Another is a frequency analysis and feature selection. And the other is a contents based spam detecting with three spam detection algorithms. Figure 1 shows the overview of the fax spam detection.

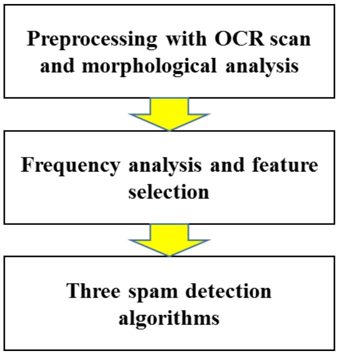
Preprocessing with OCR scan and morphological analysis
Frequency analysis and feature selection
Three spam detection algorithms

Figure 1. Overview of the proposed fax spam detection.

3.2.1 Preprocessing with OCR scan and morphological analysis

In order to extract words from scanned fax images, OCR scan is used. We use IRIS OCR for extracting words from documents which have English, number, and Korean. The raw data of scanned fax documents from OCR scanner contains unclear or un-recognized characters. The average recognition rate is 67.14%.

Because of low recognition rate of OCR scanning, we preprocess the raw data by eliminating words and selecting words with morpheme analysis. We used KLT: Korean Language Technology which is widespread [9].

3.2.2 Frequency analysis and feature selection

For the frequency analysis and the feature selection, we should analyze the frequency of words in fax documents. We collected 219 ham and 212 spam documents respectively. We select keyword features sorted by the keyword frequency. We found that the special words such as ‘loan’, ‘finance’, and ‘journey’ show up in spam documents. So we make controlled ham document. We add 229 legitimate documents with some keywords (e.g., ‘loan’ and ‘finance’) which frequently appear in spam documents.

For constructing SVM or Naïve Bayesian classifiers, we ought to select features of two classes: the fax spam class and the fax normal class. We arrange features in order of word frequency and use the frequency based feature selection which is widely used to text mining [10]. With the features, we build train and test sets.

3.2.3 Three spam detection algorithms

We use three spam detection algorithms; rule based method, SVM based method, and Naïve Bayesian based method.

Rule based method is a simple spam filtering method [7]. It firstly had a prohibited words list. The system extracts words from fax documents. If there is a prohibited word, the document is decided to a spam and blocked.

SVM is representative two-class classifier [11]. There are a lot of optimization algorithms for SVM classifier [12][13]. We here use the SVM light which is widely used in many real applications since it shows one of the best performances in the RBF optimization [12]. We do three fold cross validation and average the results.

Naïve Bayesian classifier is simple to implement and fast to make model. But it works well [14]. We train Naïve Bayesian model with 2/3 of whole data, and test the rest, 1/3. We randomly sample each train and test data for each trial. We try 50 times and average the results.

4. EXPERIMENTS

In this section, we compared the performance of three fax spam detection algorithms; Rule based, SVM based and naïve Bayesian based. We randomly selected 30 fax documents from normal and spam datasets, respectively. And we calculated the recognition rate of them manually. Table 2 shows the result. The average OCR recognition rate is about 67%. The text quality of fax documents for mining is not so high in comparing with e-mail and SMS documents.

In general, spam fax documents sequentially consist of company name, loan conditions, loan capacity, credit line, interest rate, and contact with boxes and tables. On the other hand, normal spam documents present affiliation sent the document, receiver of the document, title, contents with numbering and non-narrative expression and so on.

We measured four metrics; accuracy, precision, recall, and *F*-measure which are widely used in performance evaluation [15]. These are calculated by equation (1) ~ (4).

$$Acc. = \frac{TP+TN}{TP+FP+TN+FN} \qquad (1)$$

$$Pre. = \frac{TP}{TP+FP} \qquad (2)$$

$$Rec. = \frac{TP}{TP+TN} \qquad (3)$$

$$F-Measure \ = \frac{2\times Pre.\times Rec.}{(Pre.+Rec.)} \qquad (4)$$

Where *TP* = True positive, *TN* = True negative, *FP* = False positive, and *FN* = False negative.

We tested with two scenarios; one is to use the original fax documents (219 normal and 212 spam documents) and the other is to use the artificially created normal dataset by intentionally including 229 legitimate documents with some keywords which frequently appear in spam documents.

4.1 Performance Comparison of Fax Spam Detection Algorithms with the Original Documents

We firstly conduct the performance comparison of fax spam detection algorithms with the original data. The original data consists of 219 ham documents and 212 spam documents. Figure 2 shows that results of four measures (accuracy, precision, recall, and *F*-measure).

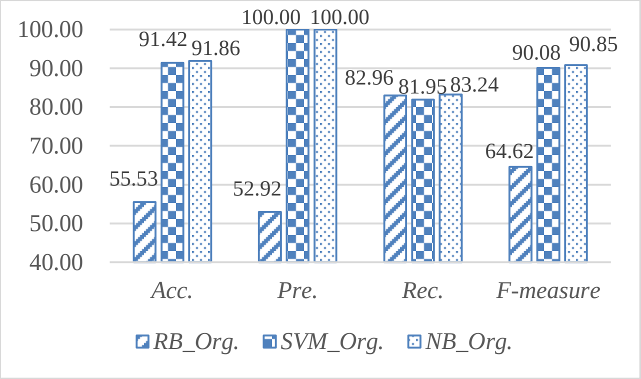

Figure 2. Performance comparison of three fax spam detection algorithms.

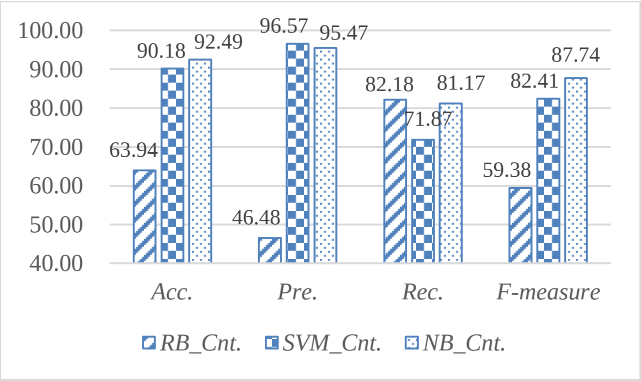

Figure 3. Performance comparison of three fax spam detection algorithms.

RB_Org. is a rule-based method with original data. *SVM_Org.* is a SVM-based method with original data. And *NB_Org.* is a Naïve Bayesian based method with original data. As the figure shows, *SVM_Org.* and *NB_Org.* are superior to *RB_Org.* in *Acc.*, *Pre.*, and *F*-measure. However, the performances are almost same in *Rec.*

4.2 Performance Comparison of Fax Spam Detection Algorithms with the Controlled Normal Documents

We secondly conduct the performance comparison of fax spam detection algorithms with the controlled normal documents data. The controlled normal documents are made by 219 ham documents plus 229 ham documents which are normal documents but contain many spam keywords such as ‘Loan’, ‘Foundation’, etc. Figure 3 shows the performance comparison of three fax spam detection algorithms with four measures (accuracy, precision, recall, and *F*-measure)

As Figure 3 shows, *RB_Cnt.* is a rule-based method with controlled normal document data. *SVM_Cnt.* is a SVM-based method with controlled normal document data. And *NB_Cnt.* is a Naïve Bayesian based method with controlled normal document data. Generally, the performance of all three algorithms dropped. And the tendency of *Acc.*, *Pre.*, and *F*-measure is similar to the original data. However, in the performance of *Rec. SVM_Cnt.* is worst. Overall, *NB_Cnt.* is superior to the others and *RB_Cnt.* is the worst.

Through these two experiments, we recommend the Naïve Bayesian based algorithm for detecting fax spam because of its outstanding performances with various conditions.

5. FIGURES/CAPTIONS

In this paper, we analyzed and compared three fax spam detection algorithms; Rule based, SVM based and naïve Bayesian based. We proposed the fax spam detection framework and evaluated three algorithms with two test scenarios. We also compared the performance of three algorithms with four metrics; accuracy, precision, recall and *F*-measure. The experimental results show that the Naïve Bayesian based algorithm is the best performance in all situations. Compared with the rule based filtering method, this algorithm shows an improvement more than 35% with same condition

6. ACKNOWLEDGMENTS

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (No. 2014R1A1A1003707). This research was supported by the MSIP (Ministry of Science, ICT and Future Planning), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2016-R0992-16-1006) supervised by the IITP (Institute for Information & communications Technology Promotion).

7. REFERENCES

[1] S.M. Lee, D.S. Kim, J.H. Kim, J.S. Park, “Spam Detection Using Feature Selection and Parameters Optimization,” *In Proceedings of the 2010 International Conference on Complex, Intelligent and Software Intensive Systems (CISIS)*, pp. 883-888, 2010.

[2] M.F. Saeedian, H. Beigy, “Spam Detection Using Dynamic Weighted Voting Based on Clustering,” *In Proceedings of the Second International Symposium on Intelligent Information Technology Application*, Vol. 2, pp. 122-126, 2008.

[3] J.W. Yoon, H. Kim, J.H. Huh, “Hybrid Spam Filtering for Mobile Communication,” *Computers & Security*, Vol. 29, Issue 4, pp. 446-459, 2010.

[4] S.J. Soman, “A Survey on Behaviors Exhibited by Spammers in Popular Social Media Networks,” *In Proceedings of the 2016 International Conference on Circuit, Power and Computing Technologies (ICCPCT)*, pp. 1-6, 2016.

[5] Q. Li, H. Yu, P. Li, “A Method for Spam Fax Detection and Classification Based on Clustering,” *In Proceedings of the 2010 2nd International Workshop on Intelligent Systems and Applications (ISA)*, pp. 1-4, 2010.

[6] M. Iqbal, M. M. Abid, M. Ahmad, and F. Fhurshid, “Study on the Effectiveness of Spam Detection Technologies,” International Journal of Information Technology and Computer Science, Vol. 1, pp. 11-21, 2016.

[7] T.S. Guzella, W.M. Caminhas, “A Review of Machine Learning Approaches to Spam Filtering,” *Expert Systems with Applications*, Vol. 36, Issue 7, pp. 10206–10222, 2009.

[8] E. Blanzieri, A. Bryl, “A Survey of Learning-based Techniques of Email Spam Filtering,” *Journal Artificial Intelligence Review*, Vol. 29 Issue 1, pp. 63-92 2008.

[9] http://nlp.kookmin.ac.kr/HAM/kor/

[10] T.M. Mahmoud, A.M. Mahfouz, “SMS Spam Filtering Technique Based on Artificial Immune System,” *International Journal of Computer Science Issues (IJCSI)*, Vol. 9, Issue 2, 2012.

[11] C. Cortes, V. Vapnik, “Support-vector Networks,” *Journal Machine Learning*, Vol. 20, Issue 3, pp. 273-297, 1995.

[12] T. Joachims, “Making Large-scale SVM Learning Practical,” *Advances in kernel methods*, pp. 169-184, 1999.

[13] S. Shalev-Shwartz, N. Srebro, “SVM Optimization: Inverse Dependence on Training Set Size,” *In Proceedings of the 25th International Conference on Machine Learning*, pp. 928-935, 2008.

[14] D.D. Lewis, "Naive (Bayes) at Forty: The Independence Assumption in Information Retrieval," *In Proceedings of the 10th European Conference on Machine Learning*, pp. 4-15, 1998.

[15] S. Ghosh, S. Mondal, B. Ghosh, “A Comparative Study of Breast Cancer Detection Based on SVM and MLP BPN Classifier,” *In Proceedings of the 2014 First International Conference on Automation, Control, Energy and Systems (ACES)*, pp. 1-4, 2014.