A Chinese Spam Filter Using Keyword and Text-in-Image Features*

Ying-Nong Chen[a], Cheng-Tzu Wang[b], Chih-Chung Lo[c], Chin-Chuan Han[d†], and Kuo-Chin Fan[a,c]

[a]Department of Computer Science and Information Engineering,
National Central University, Chungli, Taiwan
[b]Department of Computer Science,
National Taipei University of Education, Taipei, Taiwan
[c]Department of Informatics,
Fo Guang University, Ilan, Taiwan
[d]Department of Computer Science and Information Engineering,
National United University, Miaoli, Taiwan

Abstract

Recently, electronic mail(E-mail) is the most popular communication manner in our society. In such conventional environments, spam increasingly congested in Internet. In this paper, Chinese spam could be effectively detected using text and image features. Using text features, keywords and reference templates in Chinese mails are automatically selected using genetic algorithm(GA). In addition, spam containing a promotion image is also filtered out by detecting the text characters in images. Some experimental results are given to show the effectiveness of our proposed method.

Keywords: *spam, text/image features, genetic algorithm, evolution optimization, convolution neural network*

1 Introduction

Due to the wide use of Internet, E-mail is a conventional and effective communication among us in daily life. It is also an efficient promotion tool for companies. In such conventional environments, unwanted advertisement (i.e. spam) is congested in Internet. Lots of resources are wasted and users are disturbed during reading mails. Spam filter is an essential tool in filtering out those unwanted mails from mail servers.

Text classification is a popular research topic for information retrieval. Text keywords in documents are extracted and analyzed for classification. Many applications such as Web page classification, text-based searching agent, ..., etc, are widely developed using the classified results. Spam detection is an extension of text-based classification. Most of spam is filtered out using text features. However, a new type of spam is congested recently. An image mixed with text characters, like an advertisement DM, was attached in a mail. According to the reports of IronPort and CipherTrust Co. in 2006[1], the percentage of image-based spam was increased from 1% to 12%. This text-in-image spam will cheat the keyword-based spam-filter. In this paper, Chinese spam detection algorithms are proposed using the features of text keywords and image contents.

Soonthornphisaj *et al.*[2] proposed an anti-spam filtering system by using the centroid-based classification approach. They use the cosine function to measure the similarity of two mails. They also designed the Naive Bayes classifier and the k-nearest neighbor classifier to verify the testing samples are spam or not. The support vector machine-based(SVM-based) classifiers are designed to categorize the mails[3, 4]. All of them used the values of TF(term frequency) and IDF(inverse document frequency) to be the features in the classification process. In Chinese text classification, word segmentation is an essential step to extract the meaningful words. Lin[5] proposed an effective algorithm to extract the Chinese frequent strings without using a word dictionary. Chien[6] proposed the PAT-tree-based method to extract the keywords of documents for Chinese information retrieval. Chang[7] retrieved the information of web page by using the semi-structured patterns embedded in the pages. However, they are unsuitable and inefficient for verifying the mails in mail servers.

Instead of keyword-based approaches, Wu *et al.*[8] designed a behavior-based spam filter using enhanced

*The work was supported by National Science Council of Taiwan Under grant no. NSC 96-2629-E-431-001.

†To whom all correspondence should be addressed.

induction tree. The spamming behaviors include irrelevant subject to the content, the forged headers, the massive distribution, ..., and so on. Using those behavior-based features, the decision tree was generated by using an ID3 algorithm. Lai [9] made an empirical performance comparison for the spam categorization on three classification algorithms, a naive Bayes, a k-nearest neighbor, and a support vector machine. They summarized their conclusions in five points.

In this paper, Chinese spam detection is proposed using text classification and text-in-image detection techniques. The architecture of keyword-based filter is composed of two phases as shown in Fig. 1. In the training phase, E-mails are grabbed from mail servers and segmented by an MIME-based parser. The text contents of mails are collected and analyzed to generate the frequently-used Chinese words. Next, the keywords are found to characterize the documents. Each document is represented as a vector form of keyword frequency appearing in the documents. The vectors are analyzed to find the discriminant features for the classification of spam mails. These feature vectors with higher discriminant power are trained to obtain a classifier. In the testing phase, a new mail is parsed and segmented by the MIME-based parser and the word cutter, respectively. Chinese words are separated from the sentences in mails. Keywords are also extracted to obtain a feature vector. This feature vector is verified by the trained classifier. If it is a spam, an alarm is made by the detection system.

This paper is organized as follows. The preprocessing step is to parse the mail contents and to segment the Chinese sentences are described in section 2. In section 3, a GA approach is designed to find the discriminant features and the better referred templates. Moreover, an NN-based text-in-image detector is constructed in section 4 for detecting the new type spam. Some experimental results are given in section 5 to show the effectiveness of proposed method. Finally, conclusions are made in section 6.

2 Pre-processing: Parsing and Word Segmentation

Recently, mails were all transmitted in MIMIE-based format. Thus, the mails should be first parsed by an MIME-based parser to retrieve the title, sender's address, mail content, attached files, etc, from the original mails. Lots of spam filters built the black lists for filtering when a large number of mails are sent in a short time. However, the lists must be frequently updated. Mail contents analysis is another effective approach for spam filtering.

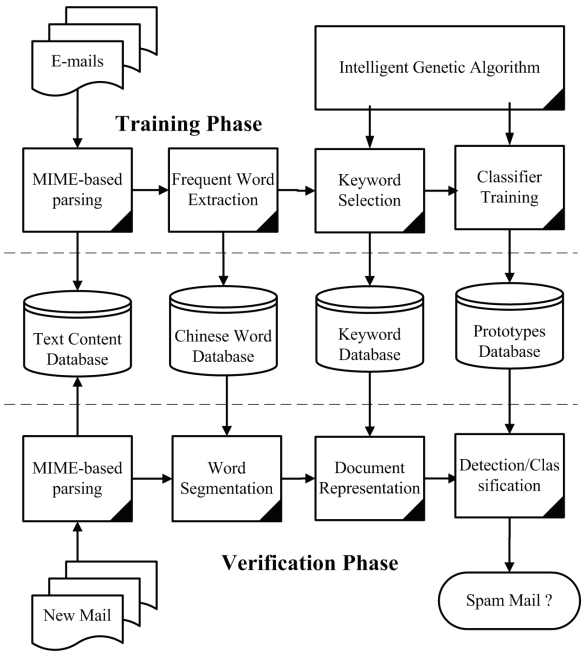


Figure 1: The architecture of text-based spam detection.

Word segmentation is the first step in analyzing Chinese mails. Some special symbols or space characters are frequently used to delimit the words in English sentences. However, there are no clear separators in Chinese sentences. It is a very time-consuming task to exactly cut the words for Chinese sentences. Long word priority rule, maximal matching rule[10, 11], and PAT tree-based rule[6] are the popular and efficient methods. In considering the efficiency of mail servers, the maximal matching approach was adopted by matching the pre-built word database. First, a Chinese frequent string(CFS) database was built from collected spam. In this study, the method of Lin and Yu's[5] was adopted to find the new words. The frequent words were extracted based on the statistical information among the words. In addition to the CFS database, two databases, an IIS[1] and a Tsai's[2], were included in our word database for word segmentation.

3 A Keyword-based Spam Filter

After the segmentation of Chinese sentences, lots of words are with less discriminant power for classification. Those lower discriminant words are of the

[1] A CFS database possessing 80,000 words was constructed by Institute of Information Science, Academia Sinica, ROC.

[2] A CFS database possesses 137,450 words, http://ftp.isu.edu.tw/pub/Windows/Chinese/phrase/wordlist.txt

frequently used words or the meaningless words in our life. They are widely spread in all class documents. Many researchers construct a *stop list* to exclude the frequent and meaningless words. However, the stop list was manually constructed and unsuitable for all applications. Therefore, high discriminant keywords were selected by a statistic-based criteria and a genetic algorithm(GA).

Keywords are selected to form feature vectors in the feature analysis module. Frequency computation is a statistical measurement to find out the importance of keywords in documents. TF and IDF are two most effective measurements in evaluating the importance. These two measurements are defined in the followings. Consider a set of documents $\mathcal{D} = \{d_1, d_2, \ldots\}$, and a set of keywords $\mathcal{W} = \{w_1, w_2, \ldots\}$, a document d_j in set $\mathcal{D}$ is represented in term of vector form $(d_j(w_1), d_j(w_2), \ldots)$. The TF is to compute the frequency of a specified keyword w_i appearing in document d_j as defined

$$TF_{d_j(w_i)} = \log\left(\frac{d_j(w_i)}{\sum_{w_k \in \mathcal{W}} d_j(w_k)}\right), \qquad (1)$$

where $d_j(w_i)$ denotes the number of keyword w_i appearing in document d_j. The IDF of keyword w_i is defined as

$$IDF_{w_i} = \log\left(\frac{|\mathcal{D}|}{|\mathcal{D}(w_i)|}\right), \qquad (2)$$

where, $\mathcal{D}(w_i)$ is the subset of set $\mathcal{D}$ containing the keyword w_i, and $|\cdot|$ represents the cardinality of set. The multiplication of values TF and IDF for keyword w_i is set to be an element of feature vectors.

Thousands of keywords are selected from the preceding section to represent the documents. The documents are represented as the feature vectors in terms of keyword numbers appearing in documents. Feature vectors with high dimensionality always complicate the design of classifiers. In addition to the keyword selection, the referred templates in the training samples are the key role in the matching process. In this study, both keywords and templates are selected by an evolution approach.

Different from the conventional IF-IDF weighting features, documents are represented by the TF features. In this section, nearest neighbor matching strategy is applied for determining the class of an input sample. In order to find the discriminant features and the better referred templates, a GA approach is designed. First, M features and N referred templates are encoded as a chromosome in term of binary vectors $[f_1, f_2, \ldots, f_M, R_1, R_2, \ldots, R_N]$. Two basic operations, crossover and mutation, are performed at each iteration. To increase the performance of multi-object evolution process, Chen *et al.* [12] designed an intelligent crossover operation. A multi-object GA based approach is proposed for a nearest neighbor classifier which maximizes classification accuracy and minimizes the template size and the feature number. More details could be referred in references[12]. Moreover, seven factors should be considered in a GA-based approach as summarized below.

1. **Initial population:** Initialize the population size P and randomize the values of the chromosomes to obtain the initial population.
2. **Fitness evolution:** The fitness function is designed as the mis-classified error.
3. **Elite set:** Store the better solutions and update the elite set S.
4. **Selection:** Select the chromosomes with the better fitness values from set S for the next generation.
5. **Crossover:** Generate the new chromosome by an intelligent crossover operation with a probability value ρ.
6. **Mutation:** Choose a single chromosome bit for the mutation with a probability η.
7. **Terminal condition:** The terminal conditions are designed as the accuracy rate is larger than a threshold or the generation iteration is larger than a number. The condition should be checked at iteration.

4 An NN-based Text-in-image Detector

In this section, a neural network-based detector is trained for detecting the text characters in images. This new-type spam could cheat and pass through the text-based spam filters. Different from the pictures we send to our friends, the advertisements mixed product pictures and promoted texts together. For promotion, artificial text characters in image-based spam are of higher contrast than the background texts in natural sense. Lots of POP characters were embedded into an image to highlight the functions of products. In this study, those text characters in images are identified by using an NN-based detector and a rule-based classifier. First, the NN-based detector was designed and trained to detect the patches of text characters. Next, the patches were clustered and classified by rules.

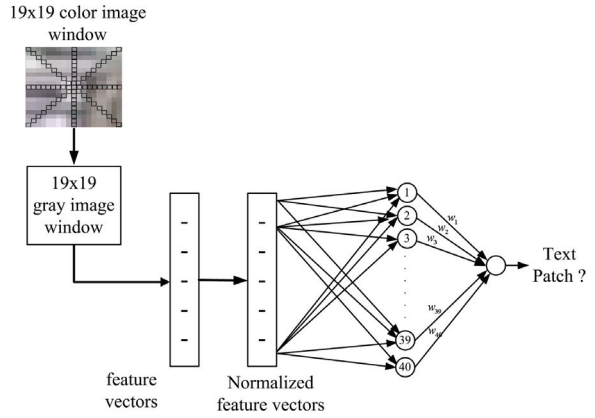


Figure 2: The architecture of NN-based text-in-image detector.

A window slid through an image from top to bottom and left to right. This window was verified by an NN-based classifier. Since Chinese characters are composed of lines, there exist clear edges and lines within a character. Unfortunately, the characters are too many and complex to design an effective detector that could detect the whole character within a window. Local and small patches with high contrast are detected by a trained NN. In order to efficiently detect the character patches, a feature vector with 73 gray values was extracted from a star pattern within a window of 19 by 19 pixels as shown in Fig. 2(a). This vector was inputted into the NN as shown in Fig. 2(b). The architecture of NN is a BPNN of 73 inputs, 40 hidden, and one output nodes. The weights were trained and a threshold was determined to determine the checked window was the text patch of a character or not. The text characters in a promoted spam image and the background characters in a conventional image are illustrated in Fig. 3(a) and Fig. 3(b), respectively. Lots of patches were detected and drawn in red rectangles in the spam image. On the other hand, only few patches were detected as shown in Fig. 3(b). In the training phase, 5000 text and 5000 non-text patches were collected for training the NN-based detector. In the verifying phase, the checked window was not slide through the whole image pixel by pixel. The main reason is that it is unnecessary to exactly identify the locations of text characters. After detecting the character patches, three classification steps are devised for identifying the text character.

1. Patch clustering: The detected patches were clustered and connected to obtain the larger regions.

(a)

(b)

Figure 3: The text-in-image detection results in a spam and a conventional picture.

2. Statistical data of a region: Made the statistical data for a clustered region. The mean and the standard derivation values were calculated.

3. Determination by threshold: The text regions contain the data with a high standard derivation. The threshold is determined from the training samples as follows. Calculate the standard derivation values of natural images and the text characters. The threshold is obtained from the average of two values.

5 Experimental Results

In this section, the experiments were conducted to show the validity of the proposed method. The system was developed and implemented on a personal computer with AMD Athlon 64 Dual Core CPU using the PHP language. The E-mails were parsed for testing. The samples were obtained from an internet web site CDSCE (CCERT data sets of Chinese Emails) for the performance evaluation. The ground truth of each sample was manually set by users. Three data sets were randomly selected for training the keyword-based filter. In each set, 100 spam and 100 non-spam samples were randomly chosen for training. Another 1000 samples, 500 spam and 500 non-spam, were collected for testing the trained filter.

Since the k-NN matching strategy is adopted in the keyword-based filter, is value k an essential role of classification? The first experiment was designed for evaluating value k. The parameters at the training phase were set as $P = 20$, $L_8(2^7)$, $|S| = 10$, $\rho = 0.6$, and $\eta = 0.05$. From the experimental results, the accuracy rates are achieved more than 94% after the second generation. It is unnecessary to choose a larger value k for more effectiveness. In the following experiments, value k is set to be 1 for efficiency. The comparisons of intelligent GA with the conventional GA are made in Fig. 4. Three data sets, I, II, and III, with 200 samples were trained and another 1000 samples were tested. The parameters for the intelligent GA are set to be $P = 20$, $L_8(2^7)$, $|S| = 10$, $\rho = 0.6$, and $\eta = 0.05$. The same scheme is performed in 30 times without loosing the generality. The performance of GA with IC outperform than the others.

In addition to the classification results, the simplification of keyword selection and template reduction is also achieved. Shown in Fig. 5, the curves of keyword selection are bounded in a range of [40%-65%] as drawn in a green line. The template numbers drawn in a red-dot line are reduced to a range of [35%-60%].

In the text-in-image filter, 100 images, 50 positive and 50 negative, were collected for testing the detection performance. Two sample images are shown in Fig. 3. The experiment was executed 30 times and the detection rates are in a range of 85.5% to 92.7%. In addition, the false accepted and false rejected rates are well reduced to the small values.

6 Conclusions

In this paper, a spam filtering system has been proposed based on text keywords and the text-in-image features. The designed scheme automatically finds the new keywords from Chinese mail corpus. Discriminant keywords and referred templates are selected by the evolutional approach. Finally, new mails were verified by the trained classifiers. Some experimental results are given to show the validity and effectiveness of the proposed system.

References

[1] "Spammers continue innovation: Ironport study shows image-based spam, hit & run, and increased volumes latest threat to your inbox," *http://www.ironport.com/company/ironport_pr_ 2006-06-28.html*.

[2] N. Soonthornphisaj, K. Chaikulseriwat, and T. O. Piyanan, "Anti-spam filtering: a centroid-based classification approach," in *Proc. IEEE Internal Conference on Signal Processing*, vol. 2, Auguest 2002.

[3] K. L. Li and H. K. Huang, "An architecture of active learning SVMs for spam," in *Proc. IEEE Internal Conference on Signal Processing*, vol. 2, Auguest 2002.

[4] H. Drucker, D. H. Wu, and V. N. Vapnik, "Support vector machines for spam categorization," *IEEE Transactions on Neural Networks*, vol. 10, pp. 1048–1054, 1999.

[5] Y. J. Lin and M. S. Yu, "Extracting chinese frequent strings without dictionary from a chinese corpus and its applications," *Journal of Information Science and Engineering*, vol. 17, pp. 513–523, 2001.

[6] L. F. Chien, "Pat-tree-based keyword extraction for chinese information retrieval," in *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrival*, pp. 50–58, 1997.

[7] K. Chang, K. W. Bowyer, S. Sarkar, and B. Victor, "Comparison and combination of ear and face images in appearance-based biometrics," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 9, pp. 1160–1165, 2003.

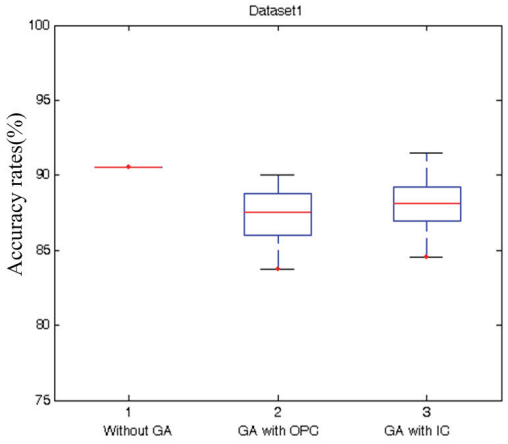
[8] Q. Wu, H. Zhang, W. Jia, X. He, J. Yang, and T. Hintz., "Car plate detection using cascaded tree-style learner based on hybrid object features," in *Proc. of 2006 IEEE International Conference on Video and Signal Based Surveillance*, pp. 15–15, 2006.

[9] C.-C. Lai, "An empirical study of three machine learning methods for spam filtering," *Knowledge-based Systems*, vol. 20, pp. 249–254, 2007.

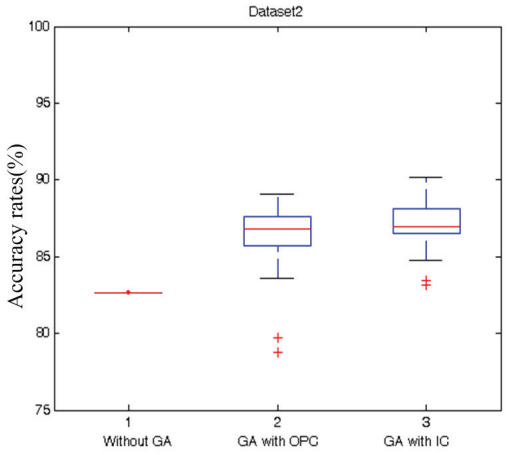
[10] K. J. Chen and S. H. Liu, "Word identification for mandarin chinese sentences," in *Proc. of 15th International Conference on Computational Linguistics*, 1992.

[11] K. J. Chen, "Lexical analysis for chinese-difficulties and possible solution," *Journal of Chinese Institute of Engineers*, vol. 22, pp. 561–571, 1999.

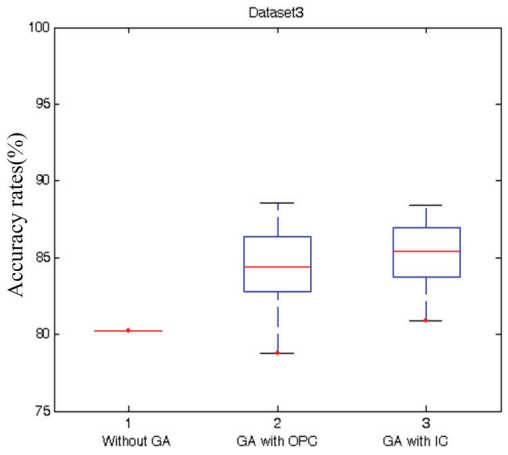
[12] J.-H. Chen, H.-M. Chen, and S.-Y. Ho, "Design of nearest neighbor classifier: Multi-objective approach," *International Journal of Approximate Reasoning*, vol. 40, pp. 3–22, 2005.



(a)

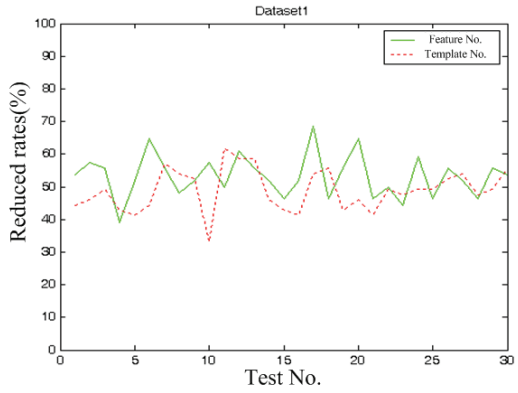


(b)

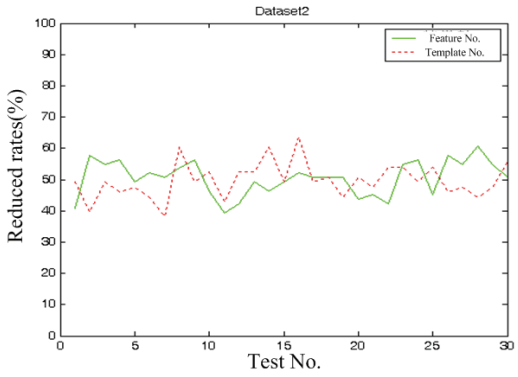


(c)

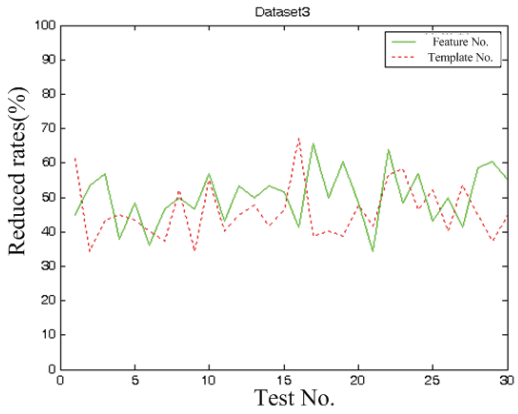
Figure 4: The comparisons for the non-GA, conventional GA and intelligent GA methods on three data sets.



(a)



(b)



(c)

Figure 5: The reduction of feature sizes and template numbers.