International Journal of Knowledge
and Language Processing
Volume 7, Number 1, **2016**

www.ijklp.org
KLP International ©2016 ISSN 2191-2734
pp.24–32

Spam Message Identification Based on Random Forest and Adaboost

Bi Yinlong, *Zan Hongying, Li Mengshuang, Li Runchuan

School of Information Engineering
Zhengzhou University
Zhengzhou 450001, China
zzubylong@gs.zzu.edu.cn, iehyzan@zzu.edu.cn

Selected Paper from Chinese Lexical Semantic Workshop

ABSTRACT. *Aiming at the problem of spam message identification, this paper extracts features from the content and structure of short messages, thus avoiding the high-dimensional and sparse feature vectors. Because of the unbalanced training data set, we adopt the method combined with Random Forest and Adaboost to reduce the impact of it. With the data set provided by "Spam Message Identification Based on Text Content" in 2015 China Good Idea Contest, better identification results have been achieved in our experiments.*

Keywords: Spam message identification, Random Forest, Adaboost

1. Introduction. With the wide use of mobile phones, short messages have become a very important part of people's daily communication. There are 361.22 billion messages sent during the first half of 2015. Although the short message brings convenience to users, lots of spam messages exert serious negative impact on people's life at the same time. According to "China mobile Internet security report in the first half of 2015" released by Baidu Mobile Guards, as of June 2015, the number of national spam messages had increased to 19.9 billion and each month there were seven spam messges received per capita. The spam message mainly includes ad-promotion, acting-invoice, fake certificates, erotic services, fraud, etc. In China, some areas have been seriously affected by the spam message. Thus, MIIT announced "The Regulations on the Communications Short Message Service". It definitely requires that the short message service providers shall not send commercial messages to their users without their consent or request. So it is of great significance to improve the identification of spam messages.

[*] Corresponding author: Zan Hongying, E-mail: iehyzan@zzu.edu.cn.

2. **Related Work.** The current techniques about identifying spam messages are mainly based on the following two methods: (1) Identify spam messages with black and white list; (2) Use classification algorithms based on text content to identify the spam message. In this paper, we mainly discuss the spam message identification based on text content.

In fact, the spam message identification based on content is making a binary classification for short messages and dividing them into spam messages and normal messages. Sohn et al. [1] used Maximum Entropy for the message classification on basis of the lexical and structural features. Karami et al. [2] further used the relationships between different structural features. Because of the unbalanced training data set, Akbari et al. [3] used GentleBoost to make a binary classification. Zhang et al. [4] identified spam messages by calculating the category membership degree. Referring to the application of Bayes classification algorithm in the junk mail filtering, Lihui et al. [5] used Minimum Risk Bayes Decision to identify spam messages. Jinzhan et al. [6] combined NaiveBayes and SVM to conduct feedback filtering on spam messages. Guanjing [7] considered the message length, punctuations and the inclusion of phone numbers in addition to the feature keyword of spam messages and achieved a good identification effect with the decision tree classification method. Manan [8] further used the feature keyword of normal messages on basis of literature[7] to reduce the false alarm rate of normal messages.

Considering the short length of messages, this paper extracts features from the content and structure of short messages with referring to the literature[7, 8]. In addition, we adopt the method combined with Random Forest and Adaboost to reduce the impact brought by the unbalanced training data set. The results of experiment show that the proposed method can identify spam messages effectively.

3. **Spam Message Identification Based on Random Forest and Adaboost.**

3.1. **Preprocessing.** There are many messages that are not standardized in their content. Especially, some spam messages often use some very non-standard editing formats to avoid to be filtered. The common non-standard editing formats include: (1) Inappropriate spaces and special symbols between the words, such as “@诚@信@办@证：电话：一五八.七八.一一一.四六七[@cheng@xin@ban@zheng: dian hua: one five eight.seven. one one one. four six seven]”;h(2) Traditional Chinese characters, such as “中國農業銀行xxxxxxxxxxxxxxxxxxxx 連力權(*Agricultural Bank of China: xxxxxxxxxxxxxxxxxxx Lian Liquan*)”; (3) Similar Chinese characters and homophones, such as “会员咔[hui yuan ka]” replacing “会员卡[hui yuan ka](credit card)”; (4) The mixture of half- width and full-width symbols.

Before segmenting the message text, this paper solves the above problem by removing the inappropriate space, converting traditional Chinese characters to simplified Chinese characters, correcting the similar Chinese characters and homophones and converting half-width symbols to full-width symbols. As to the special symbols, such as “【”, “≮” and “↓”, they are seen as stopwords in the traditional text classification, but they are important features of the spam message. In this paper, we extract these special symbols to help identify spam messages. Besides, there are more punctuations in spam messages than in normal messages, so it is important to consider the use of punctuations.

3.2. **Message Features.** Because of the short length of messages, if we directly use the bag-of-word model in the traditional text classification, the feature vector will be high-dimensional and sparse, which will affects the classification effect. Referring to the feature extraction method of literature[7,8], we extract features from the content and structure of short messages.

Message Length. One message is limited to 70 Chinese characters. In order to deliver more information to users, the sender of spam messages generally use up 70 Chinese characters. As to the messages which contain more than 70 Chinese characters, they are cut into two parts to be sent. In this paper, we take them as one complete message. The statistic result of the length of training samples is showed in Fig.1. 88.33% of normal messages contain less than 30 Chinese characters. However, 64.78% of spam messages contain between 61 and 90 Chinese characters. In addition, there are some spam messages containing between 91 and 150 Chinese characters. Thus the length can be used to identify the spam message.

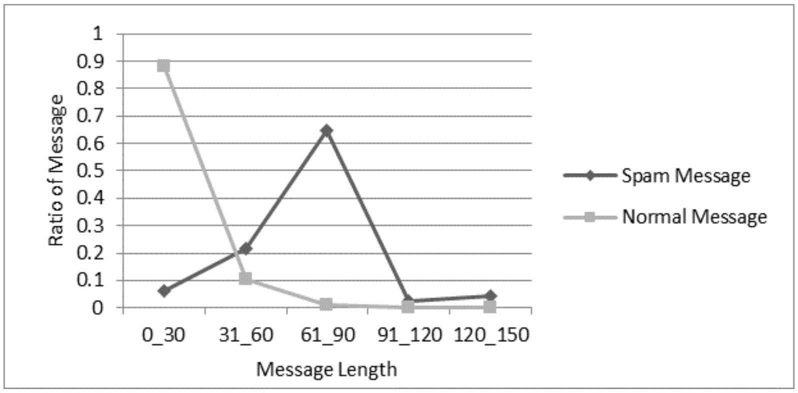

FIGURE 1. MESSAGE LENGTH

Phone Numbers, Credit Card Numbers, Commodity Prices and Dates. In order to make profits, there are lots of information about phone numbers, credit card numbers, commodity prices and dates in the spam message. Such as “温厚银祝您元宵节快乐！x月x日 火爆开盘。亲朋好友引荐热线xxxxxxxxxxx (*Wen Houyin wish you a happy Lantern Festival ! x month x day hot opening. Relatives and friends referral Hotline: xxxxxxxxxxx*)”, “农行卡：xxxxxxxxxxxx户名:徐阳杰(*The Agricultural Bank card number: xxxxxxxxxxxx and the account name : Xu Yangjie*)”. In this paper, we use the regular expression “[X|x] + .*[X|x]*” to describe such information uniformly. The distribution of “[X|x]+.*[X|x]*” in training samples is showed in Fig. 2. There are 92.86% of spam messages containing “[X|x]+.*[X|x]*”, but the percentage of normal messages that contain “[X|x]+.*[X|x]*” is 7.13%. So it is very important to consider whether one message contains such information.

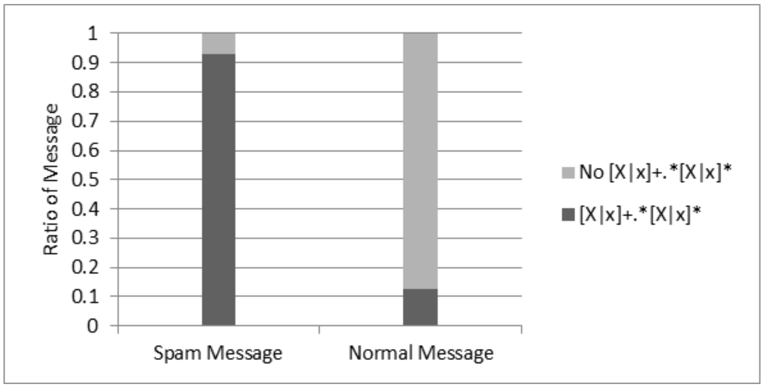

FIGURE 2. THE DISTRIBUTION OF "[X|X]+.*[X|X]*"

The Usage Rate of Special Symbols. Some spam messages use special symbols to escape the filteration mechanism or highlight the key content, such as "∢有⦠≠∢抵⦠∢增⦠≠∢扣⦠∢值⦠≠∢Ix%⦠[∢you⦠≠∢di⦠∢zeng⦠≠∢kou⦠∢zhi⦠≠∢Ix%⦠]", "代↓用开↓后 发↓付漂↓费[dai ↓ yong kai ↓ hou fa ↓ fu piao ↓ fei]", "【魅力上海】休闲商务会所带给您高端、私密的体验享受！(*Charm Shanghai leisure business club brings you a high-end, intimate experience*)". They contain special symbols "∢", "≠", "↓" and "【". However, these special symbols are rarely used in normal messages. As shown in Fig.3, 45.68% of spam messages contain special symbols, but only 6.85% of normal messages contain these symbols. In this paper, we consider the usage rate of special symbols when extracting features.

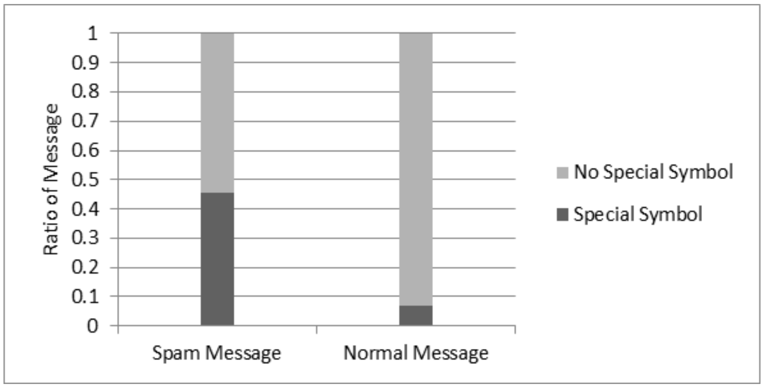

FIGURE 3. THE USAGE RATE OF SPECIAL SYMBOLS

Keywords. Spam messages generally contain some very obvious keywords. These keywords are very important for identifying spam messages. For example, one spam message about commercial advertisement: "本公司为了回馈新老顾客，全场打折，会员有礼品赠送(*In order to repay the new and old customers, the company will discounts all its goods and gives free gifts to its members*)", there are some keywords about commercial advertisement: "回馈(feedback)", "新老顾客(new and old customers)", "打折(discounts)" and "赠送(giving)". So it is very important to build the keyword set by selecting some representative words from messages. Finally, we select 1986 keywords from spam messages and 2873 keywords from normal messages.

Punctuations. As shown in Fig.1, the length of spam messages is generally greater than that of normal messages, which leads to the result that there are more punctuations in spam messages than in normal messages. In this paper, we consider the use of punctuations when extracting features.

Message Weight. Based on the above features, we further consider the weight of messages and calculate its value with the following equation (3) and (4) respectively.

According to the category label, we divide the training data into spam messages and normal messages and then build their corresponding bag-of-words model bag_1 and bag_0 respectively. For each word in bag_0 or bag_1, we put the word and its weight in the key-value pair list map_0 or map_1.

For one word t_i in bag_1, its weight can be calculated by equation (1).

$$\mathrm{weight}(t_i)=\frac{tf(t_i)}{\sum_{i=1}^{N} tf(t_i)} \tag{1}$$

Where, $tf(t_i)$ is the word frequency of t_i in spam messages, $\mathrm{weight}(t_i)$ is the weight of t_i and N is the size of bag_1.

For one word t_j in bag_0, its weight can be calculated by equation (2).

$$\mathrm{weight}'(t_j)=\frac{tf'(t_j)}{\sum_{j=1}^{N'} tf'(t_j)} \tag{2}$$

Where, $tf'(t_j)$ is the word frequency of t_j in normal messages, $\mathrm{weight}'(t_j)$ is the weight of t_j and N' is the size of bag_0.

For one message M that contains K words in data set, we use equation (3) and (4) to calculate its weight respectively.

$$\mathrm{weight\_1}(M)=\sum_{i=1}^{K} tf(t_i) \tag{3}$$

Where, $\mathrm{weight\_1}(M)$ is the weight of M in spam messages, $< t_i, tf(t_i) > \in \mathrm{map\_1}$.

$$\mathrm{weight\_0}(M)=\sum_{j=1}^{K} tf'(t_j) \tag{4}$$

Where, $\mathrm{weight\_0}(M)$ is the weight of M in normal messages,$< t_j, tf'(t_j) > \in \mathrm{map\_0}$.

3.3. **Algorithm Based on Random Forest and Adaboost.** Random Forest is an ensemble classification algorithm developed by L.Breiman[9] in 2001. The algorithm selcvects a certain number of training samples from the original sample set with the method of bootstrap resampling[10] at first, and then builds the decision tree. After several iterations, there are a group of decision trees to be built. When classifying the unknown samples, Random Forest combines the prediction of each decision tree to get the final category of them.

The building process of every decision tree in Random Forest is different from the general decision tree: (1) The training sample of each decision tree in Random Forest is randomly selected from the original sample set with bootstrap resampling, thus avoiding over-fitting; (2) The best split node of each decision tree in Random Forest is selected from the candidate feature subset according to GINI index[11]; (3) Each decision tree in Random Forest grows completely and doesn't need pruning.

Adaboost is an ensemble learning algorithm developed by Freund and Schapire in 1995[12]. It starts from a weak classification algorithm to get a series of weak classifiers through repeated iteration and then combines these weak classifiers to build a strong classifier. In each iteration, it changes the weight distribution of training samples constantly to make the misclassified sample be paid more attention in next iteration and puts weight on each base classifier according to its classification effect.

In this paper, we adopt the method combined with Random Forest and Adaboost and make Random Forest as the weak classifier of Adaboost[13]. The experimental results show that the method proposed by this paper achieves a better effect.

The algorithm is described as follows:

For training data set $D=\{(x_1,y_1),(x_2,y_2),\ldots,(x_i,y_i),\ldots,(x_N,y_N)\}$, $x_i \in R^n$ is one message of D, $y_i \in \{1,0\}$ is the category label of x_i, 1 represents the spam message and 0 represents the normal message.

Step 1: Initialize the weight distribution of training samples: $w_1 = (w_{11}, w_{12}, \ldots, w_{1i}, \ldots, w_{1N})$, where $w_{1i} = \frac{1}{N}$.

Step 2: Make K iterations with Random Forest, k=1,2,3,…,K:

1. Generate M training sample subsets randomly with bootstrap resampling: $D_1, D_2, \ldots, D_j, \ldots, D_M$.
2. For each training sample subset D_j, building its corresponding decision tree model T_j.

   When dividing the none-leaf node of T_j, the best split node is selected from the candidate feature subset according to GINI index. Because the training sample and feature are selected at random, every decision tree grows completely and doesn't need pruning.
3. Combine M decision tree models built in this iteration to build the prediction model M_k of Random Forest.
4. Put weight on M_k according to its classification effect on the current training set D and then update the weight distribution of training samples in D.

Step 3: Make a linear combination of the prediction model sequence $M_1, M_2, \ldots, M_k, \ldots, M_K$, which are generated in K iterations, to build the final prediction model $M = \alpha_1 M_1 + \alpha_2 M_2 + \ldots + \alpha_k M_k + \ldots + \alpha_K M_K$, where α_k is the weight of base classifer M_k.

4. Experimental Results and Analysis. In this paper, we use the public data set provided by 2015 Chinese good idea "spam message identification based on text content"[1]as the experimental data. The training data set includes 800000 messages with category label and the test data set includes 200000 messages without category label. In addition, there are 720000 normal messages and 80000 spam messages in training data set. We use *weka* as experimental platform and take Precision(P), Recall(R) and F value(F), which are provided by 2015 Chinese good idea, as evaluation indicators.

To validate the effectiveness of the method proposed by this paper on reducing the impact of unbalanced training data set. In this paper, we make a group of experiments on the balanced corpus at first. We select 40000 spam messages and 40000 normal messages from training data set and use them as training samples, then build training model with the method of 10-fold cross-validation. During the test, we select another 40000 spam messages and 40000 normal messages from the rest of training data set and use them as test samples. The experimental results are shown in Table 3.

TABLE 1. THE IDENTIFICATION RESULTS ON BALANCED CORPUS

Experiment Results ID	Classifier					
	NaiveBases	Logistic Regression	Multilayer Perceptron	LibSVM	Random Forest	**RandomForest +Adaboost**
Result_0	0.969	0.986	0.984	0.99	0.993	0.993

Table 1 shows that the method proposed by this paper has same recognition effect as Random Forest. They are both better than other classification methods and achieve the best recognition effect.

Then, we make two groups of experiments on the whole training data set that is an unbalanced corpus and then test on the whole test data set. The experimental results are shown in Table 2.

TABLE 2. THE IDENTIFICATION RESULTS ON UNBALANCED CORPUS

Experiment Results ID	Classifier					
	NaiveBases	Logistic Regression	Multilayer Perceptron	LibSVM	Random Forest	**RandomForest +Adaboost**
Result_1	0.915	0.968	0.973	0.976	0.983	0.986
Result_2	0.935	0.976	0.974	0.978	0.985	0.988

Result_1 shows that F value is 0.983 when we use Random Forest alone. Then we adopt the method combined with Random Forest and Adaboost, which uses the weight of samples in training process and puts weight on base classifiers according to their classification effect, to identify the spam message and F value increases to 0.986.

On basis of Result_1, Result_2 further considers the weight of messages when extracting features. The experimental results show that the comprehensive effect of all classification algorithms are improved. The method proposed by this paper achieves the best recognition effect and F value is 0.988.

[1] http://www.wid.org.cn/project/2015ccf/comp_detail.php?cid=227

5. Conclusions. Aiming at the identification of spam messages, in this paper, we extract features from the content and structure of short messages to avoid the sparse and high-dimensional feature vectors. Then we adopt the method combined with Random Forest and Adaboost to reduce the impact of unbalanced training data set. The experimental results show the method proposed by this paper is effective. Because short messages are short and with a limited amount of information, in our future work, we will further consider how to extract some more representative features to express the messages to improve the experimental result.

Acknowledgment. This paper is supported by the National Social Science Foundation Project (No. 14BYY096), the National key basic research and development program (No. 2014CB340504), Basic research project of Henan provincial science and Technology Department(No. 142300410231, 142300410308) and the Key scientific research projects of Henan Province Universities(No. 15A520098).

REFERENCES

[1] Sohn D N, Lee J T, Han K S, et al. Content-based mobile spam classification using stylistically motivated features.[J]. Pattern Recognition Letters 33(3), 364-369 (2012).

[2] Karami A, Zhou L. Improving Static SMS Spam Detection by Using New Content-based Features[J]. AIS Electronic Library (AISeL) - AMCIS 2014 Proceedings: Improving Static SMS Spam Detection by Using New Content-based Features (2014).

[3] Akbari F, Sajedi H. SMS spam detection using selected text features and Boosting Classifiers[C]// Information and Knowledge Technology (IKT), 2015 7th Conference on. IEEE (2015).

[4] Zhang Y J, Liu J L, Chang hui Y U. A spam short message classification method based on word contribution[J]. Journal of Shandong University (2012).

[5] Lihui, Zhangqi, Luhuchuan. Junk SMS Filtering Based on Context[J].Computer Engineering 12, 154-156 (2008). (in Chinese)

[6] Jinzhan, Fanjing, Chenfeng, Xucongfu. Spam Message self-adaptive filtering system based on Naive Bayes and support vector machine[J]. Journal of Computer Applications 3, 714-718 (2008). (in Chinese)

[7] Guanjing. Content-based junk short messages filtering in client side[D]. Beijing University of Posts and Telecommunications (2008). (in Chinese)

[8] Manan. Research on content based spam short messages identifying[D]. Beijing University of Posts and Telecommunications (2014). (in Chinese)

[9] Breiman L. Random Forests[J]. Machine Learning 45(1), 5—32 (2001).

[10] Efron B, Tibshirani R J. An introductin to the bootstrap[J]. Journal of Great Lakes Research 20(1), 1-6 (1993).

[11] Liukan, Yuanyunying, Liuping. A Weibo-users Indentification Model Based on Random Forest[J]. Journal of Peking University(Natural Science Edition) 2, 289-300 (2015). (in Chinese)

[12] Freund Y, Schapire R E. A desicion-theoretic generalization of on-line learning and an application to

boosting[M]// Computational Learning Theory. Springer Berlin Heidelberg, 23-37 (1995).

[13] Boinee P, Angelis A D, Foresti G L. Meta Random Forests[J]. International Journal of Computational Intelligence 2, (2006).