Journal of Computing and Electronic Information Management

ISSN: 2413-1660

A Multi-level Framework to Filtering Spam Messages based on Text Content

Jing Mi[a], Guisuo Guo[b]

Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, China

[a]mijing@bit.edu.cn, [b]guoguisuo@bit.edu.cn

Abstract: With the gradual popularity of Short Message Service (SMS) and the constantly updated forms of short message and text feature, filtering out spam message becomes more urgent. But for text content, traditional filter algorithms ignored the obvious text characteristics of spam message which influences the filter's performance. In this paper, we proposed a new framework for building classifiers that deal with filtering spam messages based on text. This framework extended features with Latent Dirichlet Allocation (LDA) topic model and specific properties and then splited spam messages in multi-level with decision tree and AdaBoost Classifiers. The datasets we used are real messages from public which can represent the varying proportion of spam and legal messages users received. We did a careful experimental procedure to evaluate the effect of this new spam filter in three aspects, 'spam' ,'legal' and 'weighted' respectively so as to analyze the result from different angles. Meanwhile we investigated the effect of training-corpus size, sub-classifiers number, feature set size on the filter's performance. The results proved that this filtering framework can effectively improve the accuracy of filtering spam messages based on text content.

Keywords: spam message filtering, feature extension, LDA, decision tree, AdaBoost, multi-level classification

1. Introduction

In our daily life, among short messages, nearly three quarters are spam ones. The survey of 12321 Acceptance Center for Bad Network and Spam Message shows that in 2014 in China, over 93% of mobile users used Short Message Service (SMS), while only 4.4% of them did not receive spam messages, which means that spam message has occupied a large part of SMS market. Due to its serious harm to society, it is imperative

to filter spam messages efficiently. There are some successful researches to deal with the problem presently. Cromack [1] performed experiments with top performing email spam filters, which showed that spam messages have something in common with junk email. Wu [2] proposed spam filters based on Cloud Computing. Nuruzzaman [3] proposed a way to train spam filters on the phone directly. Uysal [4] used CHI and IG to select features and classified spam messages with two classifiers. JinZ [5] designed an adaptive spam filter with Naïve Bayes and Support Vector Machine (SVM).

Although these methods have shown some success for general spam filtering problems, as for text content, they have not focused on the specific information removed in pre-processing stage which can be used in feature extension and classification. Thus we proposed a new method based on text that uses Latent Dirichlet Allocation (LDA) and specific properties which are quite different between spam and legal message text in general to extend features, then integrates decision tree and AdaBoost based on Naïve Bayes to classify messages in multi-level. To validate our work, we provided a detailed experiment to evaluate filter's efficiency. The results showed that the filter can work well in filtering spam messages. The main advantages of this framework are as follows:

- Enriching feature space effectively: After using LDA to predict training set, we can get more related features and make the data more topic-focused.
- Implementing easily: We chose decision tree and AdaBoost based on Naïve Bayes as classification methods. They are all simple algorithms that easier to implement than others.
- Filtering typical spam messages earlier: Using decision tree and specific properties of spam messages to filter out typical ones which may be misclassified later because these messages have less difference with legal messages after pre-processing.

In this paper, Section 2 describes the domain specific properties in detail; Section 3 presents the filter we designed; Section 4 describes feature selection and extension; Section 5 presents the multi-level classifiers; Section 6 discusses the experiments and results; Section 7 summarizes the conclusions and future work.

2. Domain Specific Properties of Spam Messages

Considering the specific problems of spam messages, we found that there are some particular features of spam message in its own text content. These features may provide powerful evidence to justify whether a message is spam or legal [6]. For example, a message which contains bankcard information is a spam message probably.

By comparing 342732 spam messages with 505735 legal messages, we got the comparison results listed in Table I, where the numerical results are average values. As shown in Table I, there are some special properties that clearly differ between spam and legal messages.

TABLE I. COMPARISON OF SPAM AND LEGAL MESSAGES

Item	Spam	Legal
The length of text	58	38
The number of nouns	9	4
The number of verbs	7	6
The number of pronouns	0	1
The number of stop words	7	8
Sensitive information	2	0
Special characters	3	0
Anti- interception	2	0

In total, we summarized five specific properties as follows:

a. Length: the average length of a short message.
b. Nouns: the number of nouns in a short message.
c. Sensitive information: the special information, such as QQ, URL, telephone number, bank card number in a short message.
d. Special characters: garbled characters, deformation characters or emotions, for example, " ┌┬┐┌┬┐ ", "℃○M".
e. Anti-interception behavior: the behavior of adding some special characters in the middle of sensitive term to avoid being filtered, for example, using wrongly written word to replace some sensitive word; insert the original complex form of a simplified Chinese character in a simplified Chinese word, etc.

3. Spam Filter Framework

In this section, we proposed a new framework that aims at building a multi-level spam message classifier. The framework is depicted in Figure 1 which consists of following main problems.

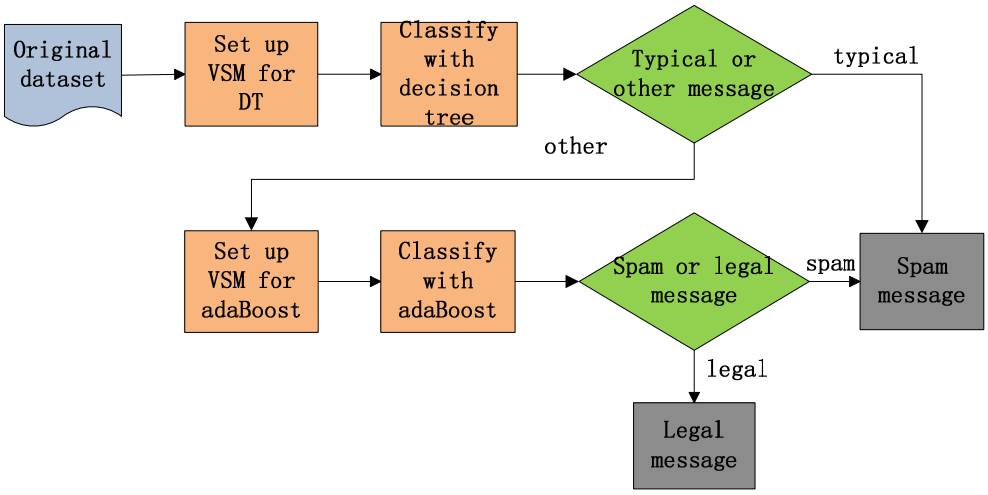

Figure 1. spam filter framework

1) Set up Vector Space Model (VSM) v1 for the original samples according to the pre-defined custom attributes.
2) Use decision tree classifier to divide the samples into typical messages and other messages.
3) Pre-process the original dataset.
4) Set up a new Vector Space Model v2 for "other messages".
5) Use AdaBoost Naïve Bayes classifier to classify "other messages", getting the final spam ones.

We must notice that the dataset v1 must be original without pre-processing. Its garbled characters and other special information should be retained. In step b, typical message is one kind of spam message which has enough special information. Before setting up v2, considering the sparseness of feature space caused by short message, extending features by LDA topic model and using special properties of spam messages are good choices.

4. Feature Selection and Extension

Owning to the fact that short texts have inherent defects such as length, weak signal and high ambiguity, the feature space based on key words must be sparse which is unfavorable to the later classification. So we use two extra tools: LDA topic model and custom attributes to get extended features.

LDA is a generative probabilistic model of a corpus first introduced by Blei [7]. The basic idea is that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words [8]. In this new framework, we train LDA topic model with extra related resources and then predict

original training set. In this way, we can find more synonyms which have similar meaning with the training set's words. Except for this, we also add domain specific properties to feature set so as to make use of inherent characteristics of spam messages. The main idea is shown in Figure 2.

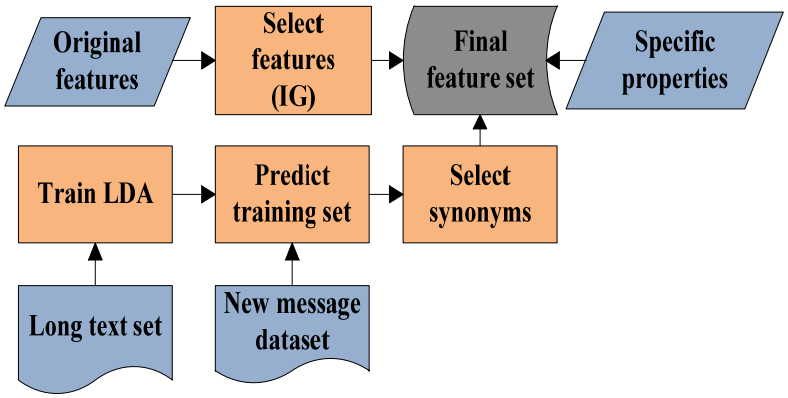

Figure 2. Feature selection and Feature extension

1) Select top n words according to Information Gain (IG) [9] as the first part (partA) of final set.
2) Train LDA topic model with additional long text set to get automatic summaries of topics in terms of a discrete probability distribution over words for each topic.
3) Predict each short message of original training set and get topic set according to each message's topic-probability.
4) Select top m words of each max-probability topic as the second part (partB) of final set.
5) Use specific properties as the third part (partC) of final set.
6) Calculate TF-IDF [10] as weight of each word in final set.
7) Normalize weight of each feature.

In step 2, the long text training set must be large and rich enough to cover a lot of words, concepts, and topics that relevant to the classification problem. Features of partA and partB use TF-IDF as their weight. For specific properties partC, calculation formulas are shown as follows:

$$\overline{score_{type}} = \frac{1}{n}\sum_{i=1}^{n} score(message) \tag{1}$$

$$w_{type} = \overline{w_{average}} + \overline{score_{type}} * count_{type}, \quad type \in \{special, purpose, length, nouns, sensitive\} \tag{2}$$

Equation (1) is used to calculate the score of custom attributes, and sub-score of each type is listed in Table II. Equation (2) is used to calculate the weight of custom attributes, $\overline{w_{average}}$ is the average weight of partA, $count_{type}$ is the number of special information belong to this type for a given short message.

TABLE II. THE SCORE OF SPECIFIC PROPERTIES

type		score
Nouns		0.1
Special characters		0.1
Anti- interception		1.0
Sensitive information	qq	0.5
	mobile	1.0
	bankcard	1.0
	phone	0.5
	email	0.5

5. Multi-Level Classification

Considering the multiple styles of spam message, there are a lot of extra information that can be used to distinguish them pref. We use two classifiers to process two sub-class samples. Decision-tree is proposed by Quinlan [11, 12]. It is commonly built by recursive partitioning. Naïve Bayes [13, 14] is based on Bayesian theorem and it currently appears to be particularly popular in text classification because of its simplicity and less training data required which are comparable to that of more elaborate learning algorithms. AdaBoost algorithm introduced by Freund [15] in 1995 and improved by Schapire [16] in 1999, is an iterative algorithm which combines many weak sub-classifiers into a strong classifier.

We add new label for original dataset as "typical" or "other". The main criterion is the score of special information a short message got. If the score is more than m points, the label is "typical", otherwise, the label is "other". We will discuss which m is proper in Section 6. The features decision tree used are five specific properties proposed in Section 2. Then we use AdaBoost to be the second classifier. The main steps of second classification are as follows:

1. Input: rest training samples with labels. $\{(x_1, y_1),....,(x_N, y_N)\}$, $y_i \in \{spam, legal\}$, sub-classifier
2. Naïve Bayes, the number of cycles T.

3. Initialize: the weights of training samples: $w_i^l = 1/N$, for all i=1,.....N.
4. Do for t = 1,..., T

1) Use Naïve Bayes to train a component classifier h_t , on the weighted training samples.

2) Calculate the training error of h_t : $\varepsilon_t = \sum_{i=1}^{n} w_i^t, y_i \neq h_t(x_i)$.

3) Set weight for the sub-classifier h_t : $\alpha_t = \frac{1}{2}\ln\frac{1-\varepsilon_t}{\varepsilon_t}$.

4) Update the weights of training samples: $w_i^{t+1} = \frac{w_i^t \exp\{-\alpha_i y_i h_t(x_i)\}}{C_t}, i = 1,2,...n$

5) Where C_t is a normalization constant, and $\sum_{i=1}^{n} w_i^{t+1} = 1$.

5. Output: $f(x) = \mathrm{sgn}(\sum_{t=1}^{T} \alpha_t h_t(x))$.

The main advantages of multi-level are listed below:

- Some messages contain representative properties such as anti-interception behavior, mobile number, etc. We can use decision tree to define criterion according to actual problems.
- As for spam filters, the price of misclassify a legal message should be far greater than the price of misclassifying a spam message. We can increase its weight of wrong classified samples for next training by AdaBoost.

6. Results and Discussion

6.1 Experimental Datasets

The description of datasets we used in this experiment is as follows:

- Training set for classifier: including 60,000 short messages, the proportion of legal messages and spam messages is 1:1.
- Testing set for classifier: including 10,000 short messages, the distribution is same with training set.
- Training set for LDA topic model: including 40 long texts.
- Here the 40 long texts are composed of 100,000 short messages without repeated samples with training set for classifier.

6.2 Experimental Tools

- SMS segmentation: ICTCLAS (Institute of Computing Technology, Chinese lexical analysis system)[17].

- Decision tree: C4.5.
- AdaBoost classifier: composed of 20 Naïve Bayes classifiers.

6.3 Experimental Results

The evaluation of experimental results are three indicators: precision (the accuracy of classification), recall (the recall rate of classification) and F-measure, here F-measure is measured by F1 = (2 * precision * recall) / (precision + recall).

We did the experiment based on different feature sets showed in Table III. By observing the results we can get the conclusion that adding additional features, especially these non-textual special properties give consistently superior results compared to just considering words in the messages. For the greater price of misclassifying a legal message, improving the precision of legal messages should be the most important target.

Table IV listed the results produced by using different classifiers. From this table we can find that decision tree classifier can filter a part of spam messages efficiently which contain evident special information. Spam messages filtered by decision tree are more likely to be misclassified as legal messages after removing this special information in pre-process stage. By using AdaBoost as major method to combine weak sub-classifiers, the performance of this filter improved effectively.

TABLE III. PRECISION, RECALL, F1 OF THE FILTERS BASED ON DIFFERENT FEATURE SETS
F1: KEY WORDS; F2: CUSTOM ATTRIBUTES; F3: KEY WORDS OF LDA TOPIC MODEL

Feature set	Spam (%)			Legal (%)			Weighted Avg(%)		
	P	R	F1	P	R	F1	P	R	F1
F1	85.6	85.1	85.4	78.9	79.6	79.3	82.9	82.8	82.8
F1+F2	91.6	90.8	91.2	87	88.2	87.6	89.7	89.7	89.7
F1+F3	85.7	85.5	85.6	79.4	79.6	79.5	83.1	83.1	83.1
F1+F2+F3	92	90.8	91.4	87.1	88.8	87.9	90	90	90

TABLE IV. PRECISION, RECALL, F1 OF THE FILTERS BASED ON DIFFERENT CLASSIFIERS
C1: NAIVE BAYES; C2: DECISION TREE; C3: ADABOOST

classifier	Spam (%)			Legal (%)			Weighted Avg(%)		
	P	R	F1	P	R	F1	P	R	F1
C1	84.6	67.2	74.9	63.8	82.5	72	76	73.5	73.7
C1+C2	85.7	85.5	85.6	79.4	79.6	79.5	83.1	83.1	83.1
C1+C3	87.3	92.5	91.2	87	88.2	87.6	89.7	89.7	89.7
C1+C2+C3	92	90.8	91.4	87.1	88.8	87.9	90	90	90

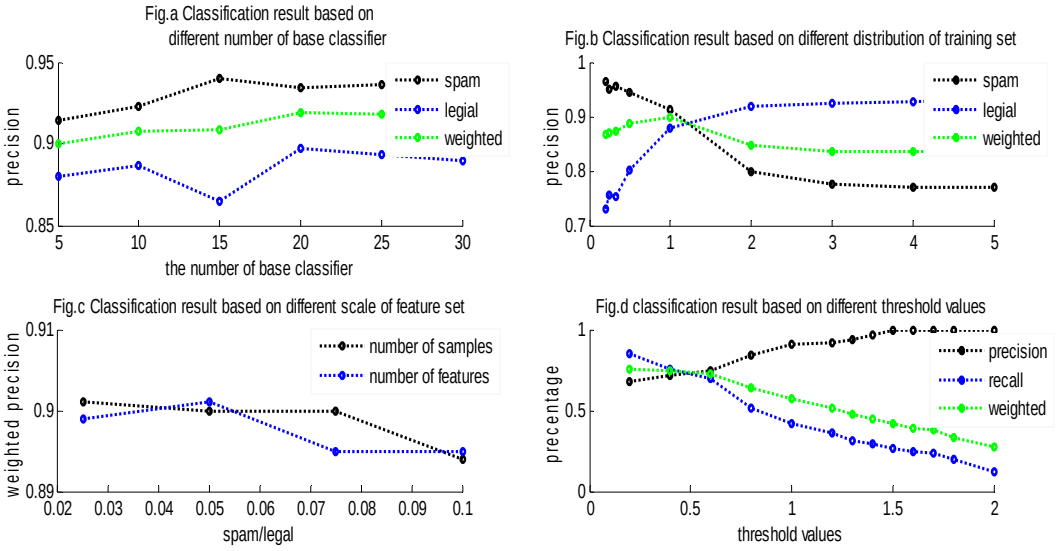

Figure 3. Classification result based on different conditions.

In Figure 3, Figure a gives the classification precision of broken lines based on different number of base classifier. Considering the computational complexity and the requirement for high accuracy, choosing 20 as the number of base classifier is proper for spam filters. The result of Figure b is produced by putting different proportions of spam and legal messages in training set. We can see that the more spam messages training set contains, the lower precision classification get, and the same appearance with legal messages. For higher precision, we should set up the training set by 1:1. In Figure c, the abscissa indicates the proportion of features of samples or original feature set. We can see from the figure that if choosing samples as reference, the most proper dimension of vector is 1/4 of samples; if choosing original feature set as reference, the dimension should be 1/2. Figure d showed the result of different threshold values as the scores of special information a message got. It is a criterion that labels a message "typical" or "other". The most important criteria of decision tree should be the precision. We can get that setting the threshold value to be 1.5 is proper.

7. Conclusion and Future Work

In this paper, for solving main problems of spam messages, we proposed a new spam filter framework based on feature extension and multi-level classification. Specific properties are the abstract features summarized after comparing a large number of spam and legal messages. By setting domain specific properties for decision tree, we can filter out those which have evident sign for spam message. LDA topic model can help to find more synonyms for training samples' key words to enrich the feature space. With the aid of AdaBoost classifier based on Naïve Bayes, we can pay more attention to the price of misclassifying legal messages by setting weight for samples and sub-classifiers. Experimental results show that this new filter has better effect on spam message filtering based on text. Besides, the filter framework is not limited to spam messages, which can also handle other similar short-text classification problems such as micro-blog, email, and twitter.

In future work, we seek to optimize the pre-process and classification algorithm proposed in this paper, for instance, how to process mispronounced characters which belongs to anti-filtering behavior, how to make use of special information more efficiently, how to recognize the users' anti-interception behavior more intelligently. We are also interested in clustering spam messages into different sub-classes and then design new specific classification method for particular types. In this way, we hope to train a better classification with higher accuracy.

References

[1] Cormack G V, Hidalgo J M G, Shnz E P. Feature engineering for mobile (SMS) spam filtering[C]ll 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, USA: Association for Computing Machinery, 2007: 87 1— 872

[2] Wu H L, Jiang Y H. SMS spare filtering based on "Cloud Security"[C]ll 2012 International Conference on Information Technology and Management Innovation, Ger— many: Trans Tech Publications, 2012: 263-266.

[3] Nuruzzaman M T, Lee C, Abdullah M F A, et a1. Simple SMS spam filtering on independent mobile phone [J]. Security and Communication Networks, 2012, 5(10): 1209— 1220.

[4] Uysal A K, Gunal S, Ergin S, et a1. A novel framework for SMS spam filtering[C]ll International Symposium on INnovations in Intelligent Systems and Applications, USA: IEEE Computer Society, 2012.

[5] Jin Z, Fan J, Cheng F, et a1. Spam message self-adaptive filtering system based on Naïve Bayes and support vector machine [J]. Computer Applications, 2008, 28 (3): 714—718.

[6] Sahami M, Dumais S, Heckerman D, et al. A Bayesian approach to filtering junk e-mail[C]//Learning for Text Categorization: Papers from the 1998 workshop. 1998, 62: 98-105.

[7] Blei D M, Ng A Y, Jordan M I.Latent dirichlet allocation [J].The Journal of Machine Learning Research, 2003, 3:993-1022.

[8] Blei D, Lafferty J. Correlated topic models [J]. Advances in neural information processing systems, 2006, 18: 147.

[9] Yang Y, Pedersen J O. A comparative study on feature selection in text categorization[C]//ICML. 1997, 97: 412-420.

[10] Salton G, Yang C S.On the specification of term values in automatic indexing [J].Journal of Documentation, 1973, 29(4): 351-372.

[11] J. R. Quinlan. Induction of decision trees. Machine Learning 1986, 1.

[12] J. R. Quinlan. C4. 5: Programs for machine learning. Morgan Kaufmann Publisher, San Mateo, CA, 1993.

[13] Scaling up the accuracy of Naïve-bayes classifiers: A decision tree hybrid - Kohavi - 1996

[14] Eyheramendy S, Lewis D D, Madigan D. On the Naïve bayes model for text categorization [J]. 2003.

[15] Y. Freund, R. E. Schapire. Experiments with a new boosting algorithm. In: Proc. the 13th Conf. Machine Learning. San Francisco: M. rgan Kaufmann, 1996. 148~156

[16] Schapire, R.E., Singer, Y., 1999. Improved boosting algorithms using confidence-rated predictions. Machine Learning 37 (3), 297–336.

[17] Zhang H P, Liu Q. ICTCLAS [J]. Institute of Computing Technology, Chinese Academy of Sciences: http://www. ict. ac. cn/freeware/003_ictclas. asp, 2002.