Unimodel-based Multi-source Portable Spam Filtering

Wuying Liu, Ting Wang
Department of Computer Science and Technology, School of Computer, National University of Defense Technology, Changsha, Hunan 410073, P. R. China
{wylui, tingwang}@nudt.edu.cn

Abstract

Various public information interactive processes, such as email, Instant Messaging (IM), Short Message Service (SMS), contain lots of advertising, obscene, illegal, and other spam information. Most of such spam information is text. From the computational linguistics perspective, textual information from different sources can be processed in a similar way. So the processing models or systems are expected to be portable on different information types. This paper introduces a unified spam filtering model for multi-source information, and proposes an approximate estimate method for the model portability. Based on the proposed model, a SVM has been used to classify the information. The experimental results show that the unified spam filtering model can be applied to multi-source information, and the SVM classification algorithm achieved encouraging performance.

1. Introduction

With the development of the computer network and mobile communication technology, the email, Instant Messaging (IM) and Short Message Service (SMS) have gradually become the important channels for information exchange. Among this information, there is always a large number of spam information, including the advertising information from information service providers, the illegal information such as obscenity, fraud, false, illegal sales. The email spam has already caused much attention, while until recently, spam IM and SMS gradually are becoming the highlighted problem as the increasing use of IM software (such as Tencent QQ, MSN, Yahoo Messenger, Google Talk, etc.) and mobile phones. The proliferation of spam information is not only a waste of network bandwidth, but also affecting people's normal information exchange. So currently the research of spam information filtering technology is hot in the field of information processing.

Most of the previous researches focus on filtering different sources information separately. From the computational linguistics perspective, textual information from different sources can be processed in a similar way. So the processing model is expected to be portable on different information types. After investigation to a variety of information sources, this paper proposes a unified filtering model. The portability of the model has been measured quantitatively. The experiments have confirmed the motivation of the proposed work.

The rest of this paper is organized as follows. Section 2 gives an overview of the related works. Section 3 introduces our unified spam filtering model for multi-source information. In section 4, the textual features of multi-source information have been developed for spam filtering. Section 5 describes the experiments applying Bayes and SVM classification algorithm and gives a discussion on the results. At last the conclusion and future work are given.

2. Related Works

People have started very early to consider the problem of spam information filtering, but what they focused on was to filter different sources information separately. For instance, the TREC\(^1\) from 2005 to 2007 carried out the spam email filtering task [1]. The AIRWeb\(^2\) from 2005 to 2007 carried out the spam web filtering challenge [2]. The spam track of TREC is almost text-based filtering, which uses a lot of experientialism methods of information retrieval and text categorization, while the spam web filtering challenge of AIRWeb is based on the links of web pages, which prefers to apply linking network analysis methods. In addition, the research of the content-based spam SMS filtering is also on the rise [3].

The spam information filtering for email, IM and SMS can be considered as a binary classification. Previous research investigates these tasks separately, such as developing different features for these different sources of information, but essentially, this information have common linguistic features. This paper argues that textual information from different sources can be processed in a similar way. So the processing models or systems are expected to be portable over different information types. This paper introduces a unified spam filtering model for multi-source information, and proposes an approximate estimate method for the model portability. Based on the proposed model, a SVM has been used to classify the information.

\(^1\) Text REtrieval Conference

\(^2\) International workshop on adversarial information retrieval on the web
3. Multi-source Portable Filtering

In order to implement multi-source portable filtering, first of all, we need to construct a multi-source portable filtering framework, in which the key point is to establish the language-feature-based unified model. The unified model can be refined in two aspects - text representation of information and filter profile representation. Within the framework, the portability of model can be measured.

3.1. Multi-source Portable Filtering Framework

The framework of spam information filtering basically follows the classic text filtering model [4]. The slightly difference from the classic model is that the filtering profile is relatively fixed, which is generally obtained from classified training data. In order to filter multi-source information, we designed a multi-source portable filtering framework as shown in Figure 1.

![Figure 1. Multi-source portable filtering framework](image)

In the multi-source analysis of information, text analyzer applies unified text representation model. While in the profile learning, unified profile representation model is applied. Finally the text and profile can be matched. It is just since the same language and similar application background of multi-source information that the unified text and profile can be represented from different sources of information.

3.2. Unified Text Representation Model

As email, IM and SMS applications are more similar, it is also quite similar to the model of that information. By extracting the text of multi-source information and analyzing their structured fields, we can map the fields to the unified representation model. This model, which is shown in the first column of Table 1, includes four fields such as Source, Object, Key Content and Body Content. For instance, the Subject field of email can be mapped to the Key Content field of the unified model. Since IM and SMS information excludes relative Key Content information, the GetKeyPhrase() function need to be increased, which can extract the key content of their message text and map it to the Key Content field of the unified model.

Table 1. Unified Text Representation Model

<table>
<thead>
<tr>
<th>Source/Object</th>
<th>Email</th>
<th>IM</th>
<th>SMS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source</td>
<td>From</td>
<td>From Messenger Id</td>
<td>Source Cellphone Number</td>
</tr>
<tr>
<td>Object</td>
<td>To</td>
<td>To Messenger Id(s)</td>
<td>Object Cellphone Number(s)</td>
</tr>
<tr>
<td>Key Content</td>
<td>Subject</td>
<td>GetKeyPhrase(Message)</td>
<td>GetKeyPhrase(SMS Message)</td>
</tr>
<tr>
<td>Body Content</td>
<td>Body</td>
<td>Message</td>
<td>SMS Message</td>
</tr>
</tbody>
</table>

The Source/Object information in the model is a very important role in behavior analysis and social network analysis. The content information of the model can be used for text-based classification and retrieval methods. Our previous research shows that applying the From/To information to construct From/To network can improve the accuracy of email spam filtering [5]. Because of privacy Source/Object information of IM and SMS corpus we collected, this paper only applies Body Content as a unified text model. In order to adapt to different filtering algorithms, we need further change unified text model to some algorithm-specific representation, such as word-bag for Bayes classification algorithm and features vector representation for SVM classification algorithm.

3.3. Unified Profile Representation Model

The unified profile model is associated with the filtering algorithm. For instance, the priori probability table of word is the unified profile model in Bayes classification algorithm and the support vector model is the model in SVM classification algorithm which can be learned form multi-source classified corpus. This paper used both Bayes and SVM classification method. For Bayes statistical method, we need to estimate the word frequency in the corpus. For SVM method, we choose three features: the number of spam words appeared in text, the number of ham words appeared and the length of text. Before SVM training, we must obtain the spam/ham words sets through classified corpus statistics. And then in SVM training and predicating we can count the number of spam/ham words appeared in text according to above spam/ham words sets.

3.4. Portability

Applying the unified model to filter multi-source information, the filtering profile and information source can come from different applications. For instance, we can use the filtering profile trained from email corpus to filter IM or SMS information. In the unified model, the ability of using a profile trained from corpus A to filter source B information can be defined as the profile’s portability. If the training and testing corpus are with the similar background of applications and use the same language, the portability of profile trained from training corpus can approximate concise to: (information scale of training set A/information scale of testing set B) and it is also simple to: (the length of training text A/the length of testing text B) in the practical application.
4. Multi-source Text Features

With the studies of email, IM and SMS, we think that these three kinds of information can be divided into two categories. Email is a delayed information transmission which is basically one batch processing of data sent. IM and SMS are real time transmission of information which uses interactive pieces to send data. The features, between delay and real time, between batch and interaction, are both reflected in three text features -- text scale, semantic integrity and repeatability of information.

4.1. Text Scale

Email is relatively complete data information. There are also a lot of convenient email tools to help people edit their emails. So generally email is longer than IM and SMS. IM and SMS information are sent by subparagraph data interactively, so the text of each message is a relatively small scale of data. Text data scale can be measured through the number of characters. Through characters statistics of three labeled corpus -- EMAIL, QQ and SM which are described in Section 5.1, we can analyze their scale features, which are shown in Figure 2.

![Figure 2. Text Scale Features of EMAIL, QQ, SM](image-url)

Figure 2 shows the data scale of email has the same distribution both in spam and ham corpus. But the text scale distribution of spam and ham is obvious distinctive both in QQ and SM corpus. This distribution is consistent with reality, because the length of email is not limited, may be longer or shorter. The QQ and SM are both interactive information, the length of most QQ and SM ham texts are shorter. But the length of QQ and SM spam text, in order to convey more complete information, must be longer relatively. From Figure 2 we can see that more than 90% of QQ ham information length is less than 30 characters, and most of ham SM message length is less than 50 characters. These two length heuristic rules can be used as an important feature for the filter.

4.2. Semantic Integrity

An important feature of the text is semantic information. The simplest semantic is word-level semantics. Generally we apply word-bag to express word-level content of text. In the word-bag-based Bayes model training, the probability of each word is needed to be statistics both in spam and ham corpus. Applying vector model of section 3.3 the two sets of spam words and ham words must be learned in advance. If the word-bag can cover the majority of semantic information, it is a good filter feature. Due to great difference between batch and interactive transmission methods, the body of email has relative semantic integrity, but each QQ or SM is a typical short text with little semantic information. Short text brings difficulties in the classification based on semantic. However, in section 4.1 we are glad to see that the short text is ham often.

4.3. Repeatability

For the repetition feature of spam and ham, we can improve filtering performance more by comparing the new text content with the stored classified one. The spam (or ham) repetition is defined as: the total number of spam (or ham) / the number of unrepeatable spam (or ham). The text repeatability of EMAIL, QQ and SM corpus is shown in Figure 3.

![Figure 3. Text Repeatability of EMAIL, QQ, SM](image-url)

Not all classified corpus can be stored because content comparison one by one is time-consuming. We get an observation that the same spam often appears in some period of time. So we can use a cache to store the last ones. For a higher content comparison speed we can compromises to compare numeric feature vectors instead of texts [6]. The improved performance of cache is dependent on the quantity of repeated information. If there are a lot of repeated ones, the cache improvement will be obvious; otherwise, there will be less improvement.

5. Experiment and Discussion
We have implemented a unimodel-based multi-source portable spam information filter. The experiment evaluated Bayes method and SVM method on three data sets.

5.1. Corpus and Evaluation

Firstly, we extract the body text of TREC06C email corpus to constitute our EMAIL corpus, collect real Tencent QQ chat text to build our QQ corpus, collect real domestic mobile phone text messages to build our SM corpus. All of these data have been labeled with golden judgment, spam or ham. The EMAIL golden judgment comes from TREC06C corpus, while the QQ and SM is labeled manually. The EMAIL contains total of 64,620 emails, the QQ includes total of 215,584 messages and the SM includes total of 10,193 mobile phone short messages. The text data scale of each corpus set and the partitions of spam (harmful)/ham (non-hazardous) are shown in Table 2.

<table>
<thead>
<tr>
<th>Table 2. Corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td>EMAIL</td>
</tr>
<tr>
<td>Total</td>
</tr>
<tr>
<td>Spam</td>
</tr>
<tr>
<td>Ham</td>
</tr>
</tbody>
</table>

Secondly, according to the 2:1 ratio of training and testing, the EMAIL, QQ and SM corpus were divided into EMAIL1/EMAIL2, QQ1/QQ2 and SM1/SM2 sets respectively. In addition, according to the length heuristic rules of QQ, we delete those messages whose length is less than 30 characters from QQ1/QQ2 sets to build the QQ3/QQ4 sets. Also we delete those short messages whose length is less than 50 characters from SM1/SM2 to form the SM3/SM4 sets. The spam/ham numbers of each set are shown in Table 3.

<table>
<thead>
<tr>
<th>Table 3. Corpus Partition</th>
</tr>
</thead>
<tbody>
<tr>
<td>EMAIL</td>
</tr>
<tr>
<td>Train Set</td>
</tr>
<tr>
<td>Total</td>
</tr>
<tr>
<td>Spam</td>
</tr>
<tr>
<td>Ham</td>
</tr>
</tbody>
</table>

Finally, we use four criteria -- the classification accuracy rate(P), the recall(R), the F1 value(F) and the average accuracy rate(AA), to evaluate the work. The first three of these criteria are often used in information retrieval. According to the characteristics of spam information filtering, this paper uses their classic criteria for spam and ham classification evaluation separately. Average accuracy rate is defined as: (number of correct classification of spam + number of correct classification of ham) / total classified number.

5.2. The Result and Discussion

The experiment includes three tests. The first one uses two methods -- the Bayes method and Bayes with length heuristic rule method. The Bayes method only uses word-bag feature of Body Content and can achieve the 96.32% average accuracy rate between EMAIL1/EMAIL2 training/testing sets. The P, R, F values of spam and ham are all high also. The previous 4.1 section shows the email corpus doesn’t have an obvious length heuristic rule. The results of the Bayes method in QQ1/QQ2 and SM/SM2 corpus sets are poor. Adding the length heuristic rule -- IF the length of message’s Body Content is less than experiential threshold (QQ 30, SM 50) THEN the message is ham ELSE it is classified by Bayes method, the average accuracy rate of QQ1/QQ2 and SM1/SM2 corpus sets separately jumped to 97.64% and 95.53% from the original 22.84% and 88.02%.

The test results show that, on the one hand, the Bayes method for short text classification is poor; on the other hand, the length of the text in the QQ and SM corpus is distinctive obviously and is a good feature for classification. This also accords with the actual situation, because the QQ and SM is the most immediate interactive text information, the length of each is often very short, some only a word. But the majority of spam information has to express the meaning of integrity in a long text. The specific results are shown in Table 4.

In order to test the portability of multi-source information, the second test uses three-dimensional vector to represent each message and uses libSVM-2.84 [7] to filter over multiple sources. The first dimension is the number of spam words the information contained, the second one is the number of ham words it contained, and the third one is the length of information. The specific results are shown in Table 5. In EMAIL1/EMAIL2, QQ3/QQ4 and SM3/SM4 corpus the third test only uses first two dimensions of the second test’s to represent each message. The specific results are shown in Table 6.

In Table 5, the AA of EMAIL1/EMAIL2 is 81.10%, while in Table 6, it is 90.00%. So the length of email is a bad feature which will reduce the accuracy of email filtering.

When train and test sets come from the same source, the filtering performance of SVM method is better than that of Bayes method both in QQ and SM corpus. Our text-based SVM method is more fit for short text and the Bayes method is more fit for long text.

In Table 5, the AA of EMAIL1/EMAIL2 is 81.10%, while in Table 6, it is 90.00%. So the length of email is a bad feature which will reduce the accuracy of email filtering.

In Table 5 and 6, the model from EMAIL is also good for filtering QQ and SM, but the model from QQ and SM is bad for EMAIL. In portable spam filtering across sources, if the portability of profile is higher, approximately the across filtering performance is better. For instance, the portability of EMAIL1/QQ2 is $80 \times \frac{2}{5} = 32$ and that of QQ1/EMAIL2 is $5 \times \frac{4}{80} = 0.125$. The AA of EMAIL1/QQ2 is 96.35%
and that of QQ1/EMAIL2 is 34.74%. So the unimodel-based multi-source portable spam information filtering is conditioned with a higher portability between profile and information source.

6. Conclusions

This paper studied the portable spam filtering method in multi-source information and proposed a unified model only based on word-bag feature. In open testing, this model can achieve a good filtering accuracy when the training and testing sets are from the same source. By adopting the unified text and profile representation model of multi-source information, the portable spam filtering can be done. Applying profile with high portability we can filter over different sources and the experiment result is encouraging.

Future research includes both in-depth analyses to the linguistics hypostasis of multi-source information and advanced portable filtering methods. We will add some important computable behavioral features, such as social network, to refine the unified portable model for better filtering effect.

References