Design and Evaluation of a Bayesian-filter-based Image Spam Filtering Method

Masahiro Uemura, Toshihiro Tabata
Graduate School of Natural Science and Technology, Okayama University
3-1-1 Tsushimanaka, Okayama-shi, Okayama, 700-8530, Japan
uemura@swlab.cs.okayama-u.ac.jp, tabata@cs.okayama-u.ac.jp

Abstract

In recent years, with the spread of the Internet, the number of spam e-mail has become one of the most serious problems. A recent report reveals that 91% of all e-mail exchanged in 2006 was spam. Using the Bayesian filter is a popular approach to distinguish between spam and legitimate e-mails. It applies the Bayes theory to identify spam. This filter proffers high filtering precision and is capable of detecting spam as per personal preferences. However, the number of image spam, which contains the spam message as an image, has been increasing rapidly. The Bayesian filter is not capable of distinguishing between image spam and legitimate e-mails since it learns from and examines only text data. Therefore, in this study, we propose an anti-image spam technique that uses image information such as file size. This technique can be easily implemented on the existing Bayesian filter. In addition, we report the results of the evaluations of this technique.

1. Introduction

Presently, the number of spam e-mails has increased and become a social problem. A recent finding has revealed that in the year 2006[1], 91% of all e-mails exchanged were spam. Due to the problems posed by increasing numbers of spam, after a certain time, it becomes necessary to distinguish between legitimate and spam e-mails; this task becomes a burden that needs to be imposed on the e-mail server. Furthermore, increases in the numbers of data flowing through a telecommunication line results in delays in the communication of an e-mail. In addition, recently, there has been an increase in the fraudulent e-mails, which are referred to as phishing e-mails. These problems have resulted in the deterioration of the reliability of e-mail. Therefore, technological measures to eliminate spam are needed to maintain the reliability of e-mail. Using a Bayesian filter is one such technical measure. The Bayesian-filter-based method can estimate whether a recently received e-mail is legitimate or spam based on the ones that we received in the past. This method has a high filtering precision, and, lately, its usage has increased.

However, there rapidly increase an e-mail which images contents of the spam to avoid text-based filtering such as Bayesian filter and SVMs. Such an e-mail is referred to as *image spam*. According to McAfee[2], by the end of 2006, image spam constituted 65% of all the spam e-mails. The Bayesian-filter-based method can memorize and evaluate only text data and not binary data such as an image. Thus, image spam can evade this filter better than text spam. Therefore, in this study, we focus on the information regarding an attached image, such as file size, as an anti-image-spam measure and suggest a method that uses both this information as well as a corpus of the existing Bayesian filter. The advantage offered by this method is that we can reduce the number of false negative (spam e-mail considered legitimate due to oversight) image spam. In addition, a proposed method does not necessarily alter the processing time of conventional filtering; it initially evaluates by employing the conventional method to examine an e-mail with an image attachment. Thus, it evaluates e-mails by considering image information only when necessary. The proposed method is responsible in determining whether an e-mail is legitimate by taking into consideration the evaluation result. Subsequently, the proposed method considers image information to evaluate an e-mail only when the evaluation results obtained by the conventional method suggest that the e-mail is spam. Thus, the proposed method can prevent the false positive (accidental classification of legitimate e-mail as spam) identification of legitimate e-mails with image attachments, thereby improving the filtering precision, because this method considers image information and only examines suspicious e-mail.

2. Bayesian Filter

First, the Bayesian filter computes the spam e-mail probability $p(w)$ of each token using the historic data of received spam and legitimate e-mails. Then, it calculates the spam e-mail probability $p(m)$ of the evaluation object using $p(w)$. If $p(m)$ exceeds a certain threshold, the e-mail is marked as spam. The methods presented by Graham[5]

or Robinson[6] methods have been applied well to compute $p(m)$. The proposed method uses the Robinson method. The Robinson method computes $f(m)$ as follows.

Firstly, it computes the $p(w)$ of each token.

$$p(w) = \frac{\frac{b}{n_{bad}}}{\frac{g}{n_{good}} + \frac{b}{n_{bad}}} \quad (1)$$

- g : frequency of token w in a legitimate e-mail
- b : frequency of token w in a spam e-mail
- n_{good} : total number of legitimate e-mails
- n_{bad} : total number of spam e-mails

$$f(w) = \frac{s \cdot x + n \cdot p(w)}{s + n} \quad (2)$$

Here, let x be the prediction probability, i.e., the probability of the first occurrence of a word in a spam e-mail, and s be the strength to provide the prediction. Moreover, let n be the number of occurrences of the word w. The values of x and s should be kept constant so as to optimize the performance of the filter. As a result of trial and error, it is concluded that the values of $x = 0.5$ and $s = 1$ are appropriate.

The Robinson method is superior to that presented by Graham with respect to its ability to handle a word w with fewer appearances. When w appears in a spam e-mail several times, the Graham method allocates 1 as the $p(w)$. However, assigning w the maximum $p(w)$ poses a problem due to the information of that degree. Therefore, when the number of appearances of w is less, the Robinson method can add on informationless to $f(w)$ by lowering specific gravity of $p(w)$. Subsequently, as the learning advances, n increases, and thus the value of $f(w)$ asymptotically approaches that of $p(w)$. In addition, $f(w) = x$ when $n = 0$. Furthermore, the probability that the e-mail being evaluated is a spam e-mail is provided by the subsequent inverse chi-square function. H indicates "Hamminess"; S, "Spamminess"; and I, the index unifying them.

$$H = C^{-1}(-2ln \prod_{w} (1 - f(w), 2n)) \quad (3)$$

$$S = C^{-1}(-2ln \prod_{w} (f(w), 2n)) \quad (4)$$

$$I = \frac{1 + H - S}{2} \quad (5)$$

In these computing methods, the spam e-mail probabilities for a token and e-mail are computed such that they assume a value between 0 and 1. When the value of this probability is almost 0, it indicates that the considered token exhibits characteristics of a legitimate e-mail and the corresponding e-mail is most likely to be legitimate. Conversely, when the value of the probability is almost 1, it indicates that the considered token exhibits characteristics of a spam e-mail and the corresponding e-mail is most likely to be spam. Bayesian filter can learn the word of the e-mail which received newly and update the data of appearance probability. Thus, the filtering precision improves with an increase in the number of words being memorized. For example, as the contents of e-mails sent by spammers vary, the standard of the spam e-mail intercepted by a filter also changes. Accordingly, a user with a filter can alter a filtering standard by adjusting the patterns of legitimate and spam e-mails. Thus, the Bayesian-filter-based method can detect several spam e-mails, which has led to its increased usage. However, spammers also manipulate the contents of spam e-mails to bypass the Bayesian filter[7]. In particular, in recent years, the number of image spam has been increasing rapidly, thereby posing a problem.

3. Image Spam

3.1. Present Conditions of Image Spam

According to an investigation conducted by McAfee[2], image spam first appeared in about 2005. At the start of 2006, image spam constituted 30% of total e-mail spam; however, this number rose to 40% in October and reached 65% by the year end.

With regard to the filtering of text-based e-mails, in particular, the Bayesian filter provides a high precision and is well known. However, it cannot extract a word, character string, or sentence embedded in an image, and neither can it cannot perform image measurements. Thus, the existing Bayesian filter can only be employed to examine the header and text portions of image spam.

Since the contents a spammer wishes to transmit are in the form of an image, there arise several cases that the text is included a few or not included. Hence, the information present in the e-mail header influences the filtering precision to a great extent. Since we can alter an e-mail header in a design of the SMTP, it is not a reliable filtering criterion. In addition, there exist several cases wherein the content being transmitted appears to be a legitimate e-mail because it contains words irrelevant to spam such as word salad when text is included. Thus, image spam poses a problem because it proffers several ways to evade filters, as compared to conventional text-based spam. Furthermore, as the size of a typical image spam is three to four times that of text-based spam, this could impose more burden on the e-mail server[2].

There exist several techniques for generating image spam. The most popular and widely used among these adds noise to the background of an image or a file name and changes the subject randomly. Further, by using animated GIFs and multi-layer image files, spammers can obscure an advertising message from filters to evade it.

Moreover, techniques capable of generating image spam

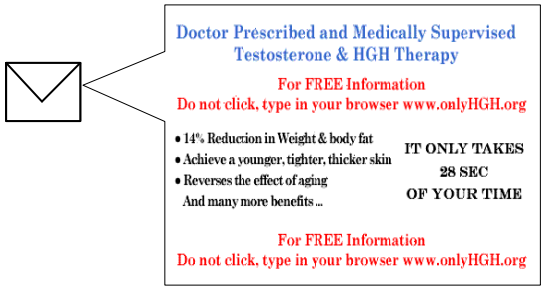

Figure 1. Example of an attached image contained in an image spam e-mail

Table 1. Details pertaining image formats

Image format	English Text	Japanese Text	Total
GIF	1,557	2	1,559
JPEG	814	15	829
PNG	41	0	41
Total	2,412	17	2,429

that is legible only to human beings have appeared[2]. These techniques have been developed to evade filters that employ OCR by adding processing that an image warps. Figure 1 shows an example of an attached image contained in an image spam e-mail. A spammer generates image content such as the one shown in Figure 1, and there exist several cases wherein certain text or a sentence is included with the spam image in order to masquerade it as a legitimate e-mail. There are the measures that used the difference of the image property as an anti-image spam measure[3][4]. As for these methods, there are problems at the false positive rate and the image processing speed.

3.2. Investigation of Attached Image Contained in Image Spam

We investigated 10,131 spam e-mails that one of the members of the laboratory received from May 2006 to February 2007. Table 1 lists our findings. Of these, 2,250 e-mails were image spam (22.2%), and they contained a total of 2,429 attached images. The image formats of the attached images were GIF, JPEG, and PNG, and GIF images constituted a total of 60% of all images. Based on language classification, English-language-based image spam constituted a total of 99% of the entire image spam number.

4. Proposed Method

The proposed method allows the existing Bayesian filter to learn image information and then evaluate the data based on the learning results. Thus, a high filtering precision is ensured.

4.1. Image Information to Incorporate in Corpus

In the case of the Bayesian filter, it is necessary to analyze the image information in order to incorporate it into a corpus to achieve high filtering precision. It is believed that legitimate e-mails that have image attachments tend to have fewer numbers as compared to legitimate text-only e-mails. In contrast to this, the image spam increases to avoid the text filtering. Therefore, it is very likely that an e-mail with an attached image is image spam. However, it is problematic to consider all e-mails with image attachments as image spam because images are also attached to legitimate e-mails. From this viewpoint, the proposed method proffers high filtering precision by letting the corpus learn the differences between the metadata of images attached to legitimate and spam e-mails, thereby empowering the corpus to differentiate between the two. In addition, we have realized that the occurrence of false positives can be reduced by allowing the proposed method to cooperate with text filtering.

As for images attached to image spam, there exist many cases wherein the image information contains several alphabets, which can be used as a text message. However, images typically depict information that cannot be represented using text, such as pictures, photographs, or illustrations. Therefore, it is believed that the metadata of images differ. As can be seen from Table 1, there exist several cases wherein the spam images being transmitted are in the GIF or JPEG format. These formats compress an image in order to lower its file size. Therefore, the compressibility of an image with few pigments and simple composition is high. In other words, it is believed that an image containing several alphabets has higher compressibility than a picture or an illustration. In such cases, we add the metadata to the corpus, such as file name, file size, area, compressibility, and state a characteristic that appears for each information entity.

4.2. Analysis of Image Data

The images considered for analysis in this study are GIFs (1,559), which account for most of the image spam, as can be seen from Table 1 and the GIF images (784), which we collected in Google[8] of the search engine by basing on the distribution of those areas. Figure 2 shows the findings pertaining file size; Figure 3, area; and Figure 4, compressibility. Further, we state a characteristic that appears for each information entity.

4.2.1. File Name

When a legitimate user attaches an image to an e-mail and transmits a message, it is assumed that the user does not transmit the same image to the same addressee several times. However, spammers transmit the same image several

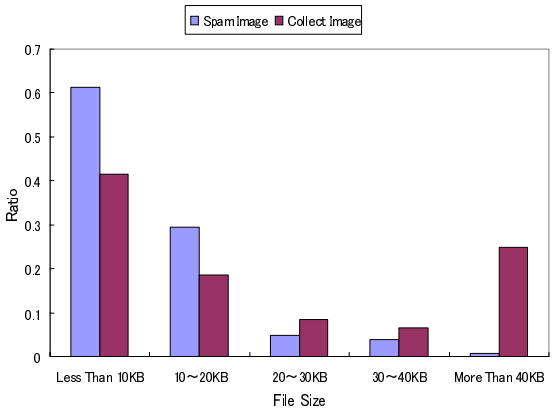

Figure 2. File sizes of spam and conventional images

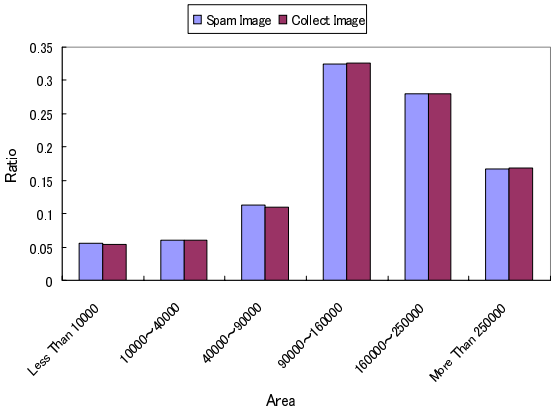

Figure 3. Areas of spam and conventional images

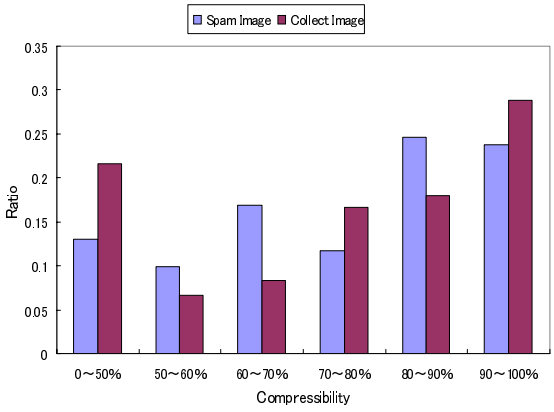

Figure 4. Compressibility of spam and conventional images

times because they send spam e-mails several times in large quantities. Therefore, it is assumed that high spam e-mail probability can be assigned to an image spam e-mail that has been sent with the same image attachment by allowing the Bayesian filter to learn the file name of the image. Actually, there exist some image spam that attached the image of the same file name.

4.2.2. File Size

Figure 2 reveals that an image attached to an image spam e-mail has a comparatively smaller file size. It should be noted that a spam image with a file size of less than 20 KB holds more than 90%, and there are many cases that the file size is smaller than a normal image.

4.2.3. Area of Image

We define the multiplication of the width and height of the image as an area. It is necessary for an image to be of a comprehensible size for it to be easily visible to human beings. The same can be assumed for both spam images and pictures or illustrations. In other words, it can be hypothesized that there the distribution of areas does not differ significantly. Thus, we can establish a distribution such as the one shown in Figure 3 to compare other information.

4.2.4. Compressibility

When converting the contents of a spam e-mail to an image, it is necessary for the image to have a comprehensible size. In addition, the file size of spam images tends to be small. Therefore, it is assumed that their compressibility might be higher than conventional images. Figure 4 reveals that in the case of images with compressibility of more than 50%, spam images are higher compressibility than collect images. This shows that spam images are more simple images than collect images.

4.3. Token of Image Data

Information such as file size, area, and compressibility assume numerical values. It is believed that a method capable of learning and evaluating single tokens in a certain mass range rather than yielding numerical values can determine the spam e-mail probability effectively. However, such a method cannot realize effective filtering because there are few cases that identical numerical values are obtained. For example, when the file size ranges between 10 and 20 KB, we incorporate it in the corpus as a token such as ”size10_20 KB”.

5. Implementation and Evaluation

5.1. Implementation Method

We implement the proposed method based on the bsfilter[9], which is the Bayesian filter being used currently.

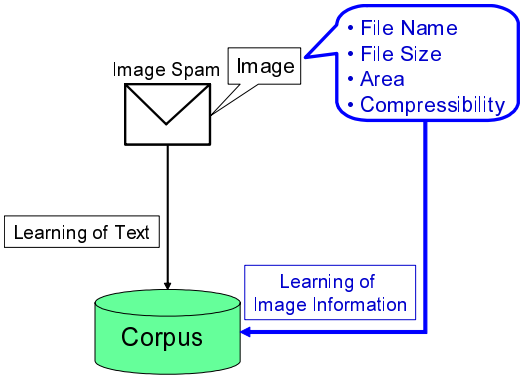

Figure 5. Learning flow

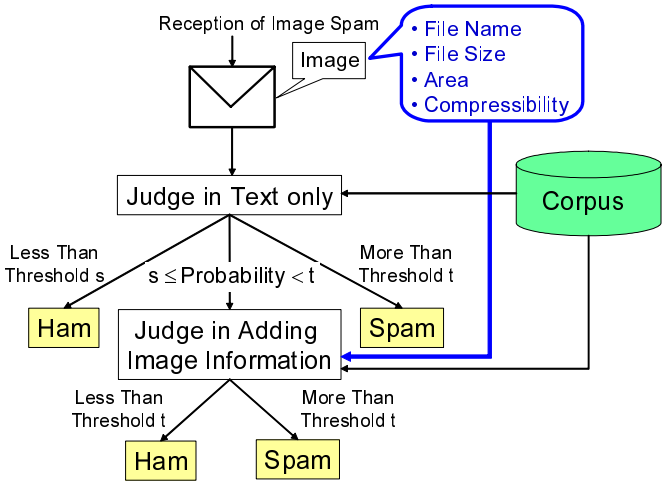

Figure 6. Judgment flow

Further, this method supports the following image formats: GIF, JPEG, and PNG.

5.1.1. Learning

Figure 5 shows the learning flow. The conventional method lets a corpus memorize only the header and text of an image spam e-mail. In addition to these, the proposed method lets a corpus memorize image information such as file name, etc.

5.1.2. Judgment

Based on past investigations, we can enumerate the following tendencies about e-mails with image attachments:

1. If a legitimate e-mail with an image attachment judges only text, the spam e-mail probability of it is low as it is almost same to a legitimate e-mail with no image attachments. This is because the texts contained in both these e-mails are similar.
2. The spam e-mail probability of text-only e-mails is around 0.5, even when the spam e-mail probability of image spam e-mails is low.

Based on these tendencies, below, we propose and implement a method to filter e-mails with image attachments.

Table 2. E-mails used for the experiment (English-language text)

	Learning	Judgment
Legitimate e-mail	300	300
Spam e-mail	200	200
Image spam	200	200
Total	700	700

1. We compute the spam e-mail probability of the text contained in an e-mail (by employing a procedure same to that of the conventional method).
2. When the spam e-mail probability is less than s, it is assumed to be a legitimate e-mail. When it is more than s, we calculate the spam e-mail probability by adding the image information (by employing the suggestion method).
3. When the spam e-mail probability computed by employing the proposed method is more than the threshold t, we judge it to be a spam e-mail. When it is under t, we judge it to be a legitimate e-mail.

Figure 6 shows the judgment flow. We set s to be 0.4, and the probability is calculated again when the spam e-mail probability of an e-mail with an image attachment is more than 0.4. In addition, when the filtering probability for texts is more than 0.9, we do not perform recalculation because it is assumed that the e-mail is almost spam, regardless of whether it has an image attachment. This experiment adds the number of image tokens that are being used for the evaluation to the ratio of the number of all tokens present in the text. For example, if the number of tokens present in the text is 100 and we assume that the number of image tokens employed in the evaluation is 10%; then, the number of tokens used for evaluating by the proposed method is 110 (file name: 1, others: 3).

5.2. Evaluation

The e-mails used in our experiment were English legitimate e-mails and English spam e-mails and English image spam ; these were received by one of the members of the laboratory. Here, spam e-mail is not image spam. Table 2 shows the total number of the e-mails that were used for learning and evaluation.

The evaluation results are shown in Table 3. The false negative rate of image spam when the threshold is assumed to be 0.9 is 13.5%, 9.5%, 3.5%, 0.5%, in that order, as shown in Table 3. A 10% addition of the number of image tokens does not result in a significant improvement as compared to the conventional method. On the other hand, a 50% addition results in considerable improvement. However, it is believed that the false positive rate increases because it is easily affected by the image information. Therefore, it is hypothesized that an addition of 30% addition is the most

Table 3. Results of evaluation after the addition of the number of image tokens to the ratio of the number of all tokens

Judgment probability	0.5 ∼ 0.6	0.6 ∼ 0.7	0.7 ∼ 0.8	0.8 ∼ 0.9	0.9 ∼ 1.0	1.0	False negative rate (if the threshold is 0.9)
Conventional method	17 (8.5%)	6 (3.0%)	4 (2.0%)	0 (0.0%)	25 (12.5%)	148 (74.0%)	27 (13.5%)
10% addition	10 (5.0%)	1 (0.5%)	4 (2.0%)	4 (2.0%)	33 (16.5%)	148 (74.0%)	19 (9.5%)
30% addition	4 (2.0%)	1 (0.5%)	0 (0.0%)	2 (1.0%)	43 (21.5%)	150 (75.0%)	7 (3.5%)
50% addition	0 (0.0%)	0 (0.0%)	0 (0.0%)	1 (0.5%)	35 (17.5%)	164 (82.0%)	1 (0.5%)

Table 4. Processing time per image spam e-mail(ms)

	Learning	Judgment
Conventional method	62	120
proposed method	94	270

appropriate. It can be observed that the false positive rate of the proposed method is almost same to that of the conventional method as it is employed for evaluating only text initially. In other words, the proposed method has a higher precision than the conventional method because the false negative rate of the former is lower than that of the latter, while their false positive rates are almost same. Table 3 reveals that the false negative rate of the proposed method is comparatively lower when the threshold value is assumed to be 0.8. Therefore, it is hypothesized that the threshold value of 0.8 is the most appropriate value and 30% of image information should be added to the number of all tokens. In this case, the false negative rate is 2.5%.

We used a machine equipped with Pentium III (1.26 GHz) and measured the approximate processing times. These results are shown in Table 4. The proposed method is slow during the learning and evaluation of one image spam e-mail 32 ms and 150 ms more than the conventional method, respectively. It can be assumed that this delay is within tolerable levels, assuming that this lag is necessary for the movement and deletion of e-mails based on false negatives.

6. Conclusion

In this paper, we describe the design and evaluation of a Bayesian-filter-based anti-image spam measure. The Bayesian filter cannot detect image spam as it capable of learning only text. Therefore, we focus on the image information of an image spam e-mail and let the Bayesian filter learn image information and suggest measures by using it to detect spam. Further, we evaluate the detection capabilities of the proposed method with regard to image spam containing GIF images and show that the proposed method can realize a false negative rate lower than that of the conventional method. In addition, it is said that the image spam filtering precision of the proposed method is better than that of the conventional method because the proposed method determines if image information needs to be added based on its judgment results obtained for only text.

For future studies, we need to investigate measures for detecting image spam for mobile phones and develop a more effective corpus embodiment method for image information and the decision method of the threshold.

Acknowledgement

This research has been partly supported by the grant of the foundation for C&C Promotion.

References

[1] postini : Email Monitoring + Email Filtering Blog. http://www.dicontas.co.uk/blog/quick-facts/email-spam-traffic-rockets/65/

[2] McAfee : McAfee Avert Labs Blog. http://www.avertlabs.com/research/blog/?p=170

[3] Wu, C. T., Cheng, K. T., Zhu, Q., Wu, Y. L., "Using Visual Features For Anti-Spam Filtering, " 2005 IEEE International Conference on Image Processing (ICIP 2005), pp. 509–512, 2005.

[4] Byun, B., Lee, C. H., Webb, S., Pu, C., "A Discriminative Classifier Learning Approach to Image Modeling and Spam Image Identification, " Fourth Conference on Email and Anti-Spam(CEAS 2007), 2007. http://www.ceas.cc/2007/papers/paper-66.pdf

[5] Graham, P.: A Plan for Spam. http://paulgraham.com/spam.html

[6] Robinson, G.: Spam Detection. http://radio.weblogs.com/0101454/stories/2002/09/16/spamDetection.html

[7] Toshihiro Tabata, "SPAM mail filtering : commentary of Bayesian filter, " The journal of Information Science and Technology Association, Vol.56, No.10, pp.464-468, 2006.

[8] Google : http://www.google.co.jp/

[9] bsfilter : http://bsfilter.org/