International Journal of Information and Computation Technology.
ISSN 0974-2239 Volume 3, Number 8 (2013), pp. 801-810
© International Research Publications House
http://www. irphouse.com /ijict.htm

Filtering Image Spam: A Survey

Shina and Preeti Anand

Computer Science and Engineering, Chitkara University, Punjab, India.

Abstract

Electronic mail being one of the most prevalent methods for exchange of digital messages has been facing the biggest threat that is spam. Since the text spam can be easily detected so a new variant of spamming came into being. The new variant of spam is the image spam that is the trending method of spamming. This paper investigates the several classifiers used in the image spam classification like the Decision Tree and the Support Vector Machine (SVM) and a combination of the both.

Keywords: Image Spam, Decision Tree, SVM.

1. Introduction

Image spam is an alias for the email spam as the job of spam emails is done by it efficiently because the main aim of a spam email is to convey the message to the recipient. But the text spam filters have emerged to be very successful in classifying the illegitimate emails directly into the spam folder of the mailbox. Hence the spammers have begun using the image spam methodology, where the text that is to be conveyed to the recipient is entrenched within the Image. Thus as soon as the recipient opens the mail the text in the image is readable. Several techniques are been developed in order to combat these superfluous mails. The main reason for the eradication of such mails is the sheer wastage of time that is spent by the recipient to chunk out them also the wastage of resources like bandwidth that is lost in sending and receiving such mails.

According to [1] spam accounts for 14.5 billion messages globally in a single day. Hence, out of every 100 mails 45 are spam. In fact some researches state that spam emails constitute even a larger portion of mails i.e. around 73 mails out 100 is indeed spam. The United States stands first as the initiator of spam, while Korea bags the second position. It had reached a climax where more than 50 mails out of 100 where spam in 2006 to 2007[2]; and the amount of image spam was about 15% ~ 22% on April, 2009[3]. Therefore, an efficient image-spam filtering system is the need of the hour.

Spam e-mails consists of several kinds of advertisements like forged lottery winning announcements, obscene content, fake lucky draws, online bogus educational institutes' and health products and false financial schemes. So to overcome such troubles a high-quality solution was required.

One of the techniques proposed [4] is extracting Mail Header from the email and depending upon the threshold values deciding it to be a spam or not. Another methodology [5] involves using the visual features of the image and then comparing it with the benchmark and then judging a mail as spam. Whereas some decide [6] on the basis of the File Properties of the image like contrast, color etc. The Optical Character Reader (OCR) technique is also a useful scheme to mine out the text from the image and then applying the routine text spam filters to classify the spam mails. But spammers retaliated by applying some modified methods, like "**C**ompletely **A**utomated **P**ublic **T**uring test to tell **C**omputers and **H**umans **A**part" (CAPTCHAs), to prevent OCR detection.

Other solutions include actions from the internet service providers by blocking spam emails by the use of blacklists IPs that contain an inventory of IP addresses that have history of spamming the users.

The image-spam filtering problem is a data mining problem that includes the classification of the illegitimate mails. The well-known classifier algorithms include Naïve Bayesian Classifier [7], Support Vector Machine (SVM) [8], Decision Tree [9], as well as Neural Network [10], etc.

The remaining of the paper is organized as follows. In Section II, we discuss the actual definition of image spam. Section III shows the datasets that have been used to fetch experimental results. Section IV discusses the several proposed systems. The conclusion is given in Section V.

2. Image Spam

The mail that is sent to a large cluster of individuals so as to forcefully impose the message on to them who would not have otherwise subscribed for it is categorized as the SPAM EMAIL. The spam email that contains an image in order to convey the message is hence known as the Image Spam Email. Not every mail that contains an image is a spam email. It depends on the type of content that has been embedded into it that decides whether it is a spam email or not. Receiving spam is a frequent complaint of all Internet users. In fact, it is a inconvenient problem as the spammers are finding ways to invade users' personal data like passwords of email accounts, net banking that has led to the need of efficient anti-spam techniques. There are broadly two categories of spam emails.

1. The emails that are spam but contain images that are not spam
2. The emails that are spam and the spam message is contained in the image itself.

Classifying the email as spam is a step-wise job wherein the first step is to demarcate an email as spam. If an email has been marked as spam then it has to be checked whether it contains an image. Then finally by judging the contents of that image it can be classified as image spam email.

3. Dataset

There are several corpuses of images available as datasets that can be used to perform experiments and fetch the results and analyze them. These include images of formats JPEG, PNG, GIF and Animated GIF. Now we shall elaborate on the datasets that have been used by the various authors.

In Image Spam Classification based on low-level Image Features, Wang et al. [11] used Spam Archive as an input corpus which consists of 9280 images, along with it Personal Spam corpus consists of 3202 images, Personal Ham corpus has 1786 images and the Personal Find Ham comprises of 1371 images.

Yan Gao et al.[12] prepared a sample spam set on their own from image spam mails received in a period of 6 months which comprises of 928 spam images from the real emails. Also they have collected 810 images from popular image website Flickr.com along with it they have used 20 scanned documents.

In Using Visual Features For Anti-Spam Filtering, Ching-Tung Wu et al. used the SpamArchive as used by [5] corpus which contained 122,877 emails out of which 46,395 emails have images. Another dataset used is the Ling-Spam which is a popular anti-spam filtering corpus.

Spam Assassin is another popular corpus that is commonly used as a dataset for spam filtering. This has been used by Gargiulo et al. [13].

Table 1: List of data sets used.

Author Name	Data set	No. of spam images
Wang	SpamArchive	9280
	Personal Spam	3202
	Personal Ham	1786
	PersonalFind Ham	1371
Gao	Spam Images from real emails	928
	Image collected from popular website	810
Tung Wu	SpamArchive	46,395
Liu	Text Retrieval Conference (TREC) dataset	6728
	Sansone dataset	20263
	Dredze dataset	3297

Liu et al. have used a combination of several datasets. The first one being TREC dataset (Text Retrieval Conference) that has 6728 image emails. Sansone dataset, which comprises of 20263 spam emails. Dredze dataset is also used where in there are 2006 legitimate images and 3297 spam images. The last corpus being, Image Spam Hunter dataset which has been used by Yan Gao et al. also.

All these dataset have been used in the discussed papers in order to give optimum results regarding filtering of image spam using their devised techniques.

4. Discussion

4.1 Decision Tree Classifier

In this paper [12] the main thought used is there is no requirement for extracting the text from the image. Since the image spam has been artificially generated, it differs from the natural or real images in terms of the visual features. Thus a probabilistic boosting tree has been used to classify the spam images from the legitimate images by using their color and gradient orientation histograms.

The two histograms used are color histogram and the gradient orientation histogram. The histograms of natural and real images are continuous and those of artificially generated images have isolated peaks. The distributions of gradient orientation for natural images are smoother and noisy as compared to the spam images.

In the color histogram, 2D color histogram in a color space. Since the main concern is to shape of the histogram so the bins are sorted in decreasing order and only top D bins are extracted from it.

In the gradient histogram each pixel is calculated using the Sobel's operator by using a threshold value which is equal to 50 in order to quantize the gradient histogram from 0-360 degrees. The accuracy of the method used is 89.44% which is a considerable amount of precision taking into consideration that only visual features have been taken into account.

Table 2: Based on Decision Tree classifier.

	Image Spam Hunter	**Visual and textual features**
Features used	1. Color Histogram 2. Gradient Orientation Histogram	**1. Visual** Contrast, energy, entropy, correlation, homogeneity and perimetric complexity **2.Textual** Text_length, word_number, ambiguity, correctness, special _length and special_distance.

Algorithm Used	Decision Tree	Probabilistic Boosting Tree
Accuracy in detecting Spam	89.44 %	τ=0 94.8% τ =1 99.8%

In this paper [13] the proposed model consists of two stages. The first stage is extracting the global properties from the suspected image by applying the image processing operations. The second stage of the proposed model is a collection of two sub-processes where first the text from the image is extracted by the help of an Optical Character Reader (OCR), the second sub-process involves the scrutiny that if the text in the image was purposely concealed. The OCR is used to get a fingerprint of the distorted image and then a set of features is obtained from the image. These features are then subjected to the Decision Tree classifier in order to distinguish the illegitimate emails from the legitimate ones.

Visual features: The six visual features extracted from the co-occurrence matrices [14] are: contrast, energy, entropy, correlation, homogeneity and perimetric complexity

Textual features: When the text has been extracted from the suspected image with the help of the OCR the text is generally in the disguised form, which is not understandable by the machines but can be read with the human eye. This is because some special characters have been used in the text. In the proposed model a set of special characters has been created that comprises of {!,“.#, $, %, &, ‘,(,),*,+,,,-,…,/,@,^}. The following textual features have been used:Text_length, word_number, ambiguity, correctness, special _length and special_distance.

The classifier used in the experiment is the decision tree C4.5 (J48) that available in the open source data mining tool Weka[15]. The values of the visual and the textual features are input to the classifier along with a threshold valueτ. By varying the value of τ suitable results are obtained and the suspected mail is classified as legitimate or illegitimate. Using the τ value 0 the precision is 94.8 and when value is 1the precision rate increases to 99.8%.

Thus, the features used by the above mentioned algorithms along with their efficiency has been depicted in Table I.

4.2 SVM Classifier

In [11], the basic identify image spam idea is divided into two steps as:

Step 1: The first step is to get extremely low computational cost file properties and color features and texture properties. These are also called file, color, texture (FCT) . Raw image features used in this paper are Image Size, Width, Height, Bit Depth, and Image File Type. Based on these raw features, paper generate a 10 dimensional feature vector (f_1-f_{10}) as Width, Height, Aspect ratio, File size, Image area, Compression Binary: JPEG image, Binary GIF image, Binary: PNG image, Bit Depth. The image file type features are binary features that are set to 1 if the file is of the specified type and to 0 otherwise.

Color is a general feature used for describing image in the manner of simplicity and intuition for basic attribute description of image. To characterize the color features of image spam, the paper selected 5 parameters for representing color features as Number of Colors, Variance, the Most Appear Number of Different Colors in the image, Primary Color of the image, and Color Saturation. Texture is a reaction to an image in a region of the spatial distribution of pixel gray-level properties, the inherent properties of the structure of this space neighborhood pixels can be directly related to portray. The simplest way to describe the demographic characteristics of the texture is the moment of gray histogram. In the histogram of the n-order moments, variance is a measure of gray scale contrast, an expression of the curve relative to the mean of the distribution.

Step 2: In this step a Support Vector Machine (SVM) classifier is run with Radial Basis Function (RBF) kernel as the kernel function to classify image spam. A classification task usually involves with training and testing data which consist of some data instances. Each instance in the training set contains one "target value" (class labels) and several "attributes" (features). The goal of SVM is to produce a model which predicts target value of data instances in the testing set which are given only the attributes.

In [5] a set of features from the images contained in the email were extracted. The set of features was then used for classification, with one-class Support Vector Machines (SVM) being used as the base classifier. Three sets of features as extracted as follows: Embedded-text features, Banner and graphic features, Image location features.

Table 3: Based on SVM Classifier.

	Visual feature anti spam	**Image spam low level**
Features used	1. Embedded-text features 2. Banner and graphic features 3. Image location features	1. Image Features 2. Color Features 3. Textual Features
SVM package	LIBSVM	LIBSVM with the RBF kernel
Accuracy in detecting Spam	81.40%	More than 95%

Spam emails are embedding text messages in images to get around text-based anti-spam filters. To detect such devious techniques, [16] developed a text-in-image detector which is capable of detecting the text region(s) in an image. Text-in-image detector is used to scan through each image in the email and derive the following embedded-text features: (1) the total number of text regions detected in all images in the email, (2) the percentage of images with detected embedded-text regions, and (3) the pixel count ratio of the detected text regions to that of the overall image area.

Many of the images in spam emails are banners and computer generated graphics which are part of advertisements. Developed a banner detector and a graphics detector, banner images are usually very narrow in width or height. Also, banner images usually have a large aspect ratio vertically or horizontally. Graphic images, however, usually contain homogeneous background and very little texture. Using these detectors, we can extract the following banner and graphic features: (1) the ratio of the number of banner images to the total number of images, and (2) the ratio of the number of graphic images to the total.

Spammers usually put their images behind web servers and create references in the emails to save server and network resources. This is in contrast to personal emails, where images are usually attached to the emails.

In previous approaches, the anti-spam filtering problem has typically been treated as a two-class or multiple-class classification problem. One difficulty with the two-class and multiple-class classification is the need for multiple sets of training samples. [17] Proposed one-class SVM as the base classifier. SVM classifier maps the data from the input space to a higher dimensional space, called the feature space, and constructs a hyper plane in the feature space which separates the data with a maximal margin. In [5], the support vectors construct a probability-dense region which encompasses the training data in the input space.

Based on the SVM classifier the features extracted by the above mentioned papers along with their experimental results have been depicted in Table III.

4.3 Combination of Decision Tree and SVM Classifier

In this paper [13] the basic process that has been applied is the feature extraction. The features extracted are height, width, image type and the file size. From these four features a matrix of nine features is created. The matrix consists of the following features: Image height, Image width, Aspect Ratio, Binary GIF image, JPEG Image, PNG Image, file size, image area and compression. If the image type is PNG then the value of the attribute is set to 1. Similar is the case for the JPEG and PNG files. Then in order to extract the information from these features the Signal to noise ratio is calculated.

For classifying the images as spam or ham there is a combination of two classifiers that is used C4.5 algorithm for creating a decision tree and the SVM algorithm for creating a support vector machine along with a RBF kernel. The C4.5 decision is available in the open source tool Weka [15] and the SVM is available through the LIBSVM [18]. The outcome of the experiment performed is having an efficiency of 60% with a low value of false-positive (classifying spam as legitimate email).

Based on our study we have deduced that the use of visual and textual features in the better choice when using the Decision Tree classifier. Whereas when using the SVM classifier the low level features of the image give a better result. But the use of both Decision Tree and the SVM classifier has not been able to deliver as the algorithms have delivered when working independently.

5. Conclusion

In this paper, we surveyed the several techniques used by the various authors for filtration of the most disturbing and the time consuming problem related to the email. The image spam is the latest drift of spamming since most of the text spam filters are quite precise and proficient in complying with their jobs. Moreover the concept of Blacklist IPs has also considerably condensed the count of Spam that is visible in the email. So the spammers have tried to overcome these constraints and devised the new tactics for spamming i.e. image spam. But the research to eliminate the very arrival of spam to the mail accounts is in its full swing. As discussed in Section IV certain papers have used the visual properties of the images in order to label out the spam out from the legitimate mails. Also some use the low-level properties while the others employ the high-level features to extract the spam. Certain papers discuss the classification based on the header and the file properties of the image concerned. Whilst some researchers have used the combination of the two methodologies mentioned above. Different algorithms have been used in the different papers like C4.5, Naïve Bayesian, SVM classifiers in order to classify the mails. We hope that our study will assist to augment the understanding of this topic, also be informative for future researches in the same direction.

References

[1] http://www.spamlaws.com/spam-stats.html

[2] M86 Security Whitepaper, http://www.m86security.com/newsimages/trace/RiseandFallofImageSpam_March08.pdf

[3] IBM X Force Report, http://blogs.iss.net/archive/image-spamrebirth.html

[4] Tzong-Jye Liu ; Wen-Liang Tsao ; Chia-Lin Lee, A High Performance Image-Spam Filtering System, Ninth International Symposium on Distributed Computing and Applications to Business Engineering and Science (DCABES), 2010, pp: 445 – 449 doi :10.1109/DCABES.2010.97

[5] Ching-Tung Wu ; Kwang-Ting Cheng ; Qiang Zhu ; Yi-Leh Wu" Using visual features for anti-spam filtering", IEEE International Conference on Image Processing, 2005. Volume: 3, pp: III - 509-12, doi: 10.1109/ICIP.2005.1530440

[6] Krasser, Sven ; Yuchun Tang ; Gould, J. ; Alperovitch, D. ;Judge, P." Identifying Image Spam based on Header and File Properties using C4.5 Decision Trees and Support Vector Machine Learning" Information Assurance and Security Workshop, 2007,pp: 255 - 261, doi: 10.1109/IAW.2007.381941

[7] M. Uemura, T. Tabata, "Design and Evaluation of a Bayesian-filter based Image Spam Filtering Method," International Conference on Information Security and Assurance, pp. 46-51, May 2008.

[8] H. Drucker, D. Wu, V. N. Vapnik, “Support Vector Machines for Spam Categorization,” IEEE Transactions on Neural networks, vol. 10, no. 5, September 1999.

[9] J. R. Quinlan, “Introduction of Decision Tree,” Machine Learning,vol. 1, pp. 81-106, 1986.

[10] C. Wu, “Behavior-based Spam Detection Using a Hybrid Method of Rule Based Techniques and Neural Networks,” Journal of Expert Systems with Applications, vol. 36, pp. 4321-4330, April 2009.

[11] Chao Wang ; Fengli Zhang ; Fagen Li ; Qiao Liu, “Image spam classification based on low-level image features”, International Conference on Communications, Circuits and Systems, 2010 pp: 290 - 293

[12] Yan Gao ; Ming Yang ; Xiaonan Zhao ; Pardo, B. ; Ying Wu ;Pappas, T.N. ; Choudhary, A., “ Image spam hunter”, IEEE International Conference on Acoustics, Speech and Signal Processing, 2008, pp: 1765 - 1768, doi: 10.1109/ICASSP.2008.4517972

[13] Gargiulo, F. ; Sansone, C. , Combining visual and textual features for filtering spam emails, 19th International Conference on Pattern Recognition, ICPR 2008. , pp: 1 – 4, doi: 10.1109/ICPR.2008.4761828

[14] J. Shi and J. Malik. Normalized cuts and image segmentation. In IEEE Transactions on Pattern Analysisand Machine Intelligence, volume 22, pages 888{ 905, 2000.

[15] http: //www. cs.waikato.ac.nz/ml/weka/.

[16] C.-T. Wu, et al. A Novel Embedded-Text-in-Image Detector and Its Applications. UCSB Technical Report, January 2005

[17] B. Scolkopf, J. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson. “Estimating the support of a high-dimensional distribution” .Technical Report 99-87, Microsoft Research, 1999

[18] http://www.csie.ntu.edu.tw/~cjlin/libsvm