International Journal of Research, Science, Technology & Management
ISSN Online: 2455-2240, Volume 3 Issue I, April 2016

PARTIAL IMAGE SPAM E-MAIL DETECTION USING OCR

Akanksha Mishra, Mayank Rajpoot

Abstract— **Even though there a number of inbuilt E-mail filtering software associated with every e-mail service providers like Gmail, Yahoo mail they identify only text spam and image spam. As a result they could not detect the mix image spam. So this paper concentrates on identifying and avoiding partial image spam whenever it is received to our e-mail inbox by an unknown source across the Internet.**

Keywords—**E-mail service provider, Partial image spam.**

1. INTRODUCTION

The image spam is a type of junk e-mail that replaces literal with images as significance of simpleton spam filters. The image spam usually embeds into an e-mail. Users will receive this spam e-mail by double clicking at the e-mail or whenever the e-mail is open.

When an image spam is distributed over a network, it is a larger drain in the network resource than the literal spam because an image file is larger than the text file and it requires a higher bandwidth. As a consequent, it causes a grater degradation of transfer rates. The image spam has the different formats, such as .gif, .bmp, .jpg, .png, .gif etc, most images are in the format of .gif and .jpg. In fact, the process of classifying the image spam groups because an image contains many properties, for example brightness, contrast, and radian. Presently, there are tremendous methods in the image spam detection process. Unfortunately, these methods can detect only detect only images of texts, or humans or bodies appearances.

Therefore, this paper proposes a new method that enhances the image spam detection. So, the image spam, not only the images of text or human pictures, but also other image spam such as images of advertisements, can be detected.

II. PROBLEM DEFINITION

The primary focus of this paper is on identifying and avoiding an image spam e-mail across the web. Security has always been an important aspect of quality of the service provided. Our aim is to develop a novel framework using OCR which helps to ensure that we won't have spam in our e-mail inbox. By identifying the spam it relives the user from worrying about the spam e-mail. There are a number of providers who provide the e-mail filtering software to whom the clients will use like Google, Yahoo, and Rediff etc. When these software provides provides inbuilt filtering software , there is a major threat of mix image spam for a number of reasons like e-mail message from unknown source , link leads to an unknown web pages. The message that are sent by any unknown source which must be prevented. Before going to the rest of this paper, all necessary definitions must be stated.

1) Image spam e-mail

An image is classified as an image spam if and only if the following conditions are true more than one.

1. The image is delivered by an unknown source.
2. The image is attached with the main body of e-mail.
3. The image is a hyperlink to an unknown web Page.

2) *Format of an Image Spam*

The format of an image spam is the same as standard format of image over the Internet like Subject, Message Body, and Receiver Address.

3) Characteristics of an image spam e-mail

An image spam can be classified in two categories: pure image, or mixed image. The pure image spam is the spam contains only image(s); the mix image spam consists of images and a text message attached to an e-mail.

Everyday we are encountering new Internet spam which is much obfuscated and it becomes a serious problem across the network. So an interesting research problem is to find a novel technique to decrease spam. Various solutions have been proposed to detect different types of spam; however, a spam has developed itself and moved from a text spam to be an image spam. This image spam causes numerous problems because there are varieties of images to be protected.

In the year 2007, Zhe et al. [14] Proposed a detection technique for an image spam using near-duplicate

detection technique. This technique produces a non-spam image repository. Then, when users receive an image, it will be compared with images in the image database for the spam filtering process. The received image will be eliminated when the feature vector of it is different from the feature of the images in the image repository. Similar to this research, Battista et al. [2] proposed a method to classify an image spam by comparing the received image with the original image in the database. The received image will be terminated when there is a difference between these two images. These two techniques have limitation in detecting image spam because some spam consists of both texts and images. Thus, the proposed techniques cannot be applied. Therefore, [12] proposed a method to detect the image spam that consists of both texts and images. This technique uses one-class Support Vector Machines (SVM) to classify a spam from the received e-mail. Moreover, the researchers separated three sets of features to detect the image spam; these features are Embedded-text features, Banner and graphic features and image location features. The filtering process [1] tries to recognize an image spam using the OCR tool. This technique is similar to [12] because it detects spam images , except that these spam flow via e-mail and the detection uses a non-uniform background, pixels of different colors for each character, and distortion of text lines or single characters.

Francesco et al. [4] proposed two different image processing techniques which are used to detect the image spam that composed of both text and image. The components-based method on [5] SIFT method to detect the image spam. This method detects image spam that was modified from the converted of content text to an image and the embedded spam message through an e-mail. Some image spam was identified using a boosting tree which is learning based prototype system announced by [13]. The detection system is called as the Image Spam Hunter. The other method that applied a tree structure to classify the spam is proposed by Sven et al. [11] uses a decision tree and a support vector machine as the classifiers. The advantage is that it can detect a large amount of image spam. A matching mechanism proposed by [7] uses using user specified image content focuses on text message hidden in spam images using SIFT algorithm that led many researches to use these properties as attributes of an object to detect as the image spam. A new method of Jordan [6] will converts JPEG images to ASCII using JP2A. Peizhou et al. [10] identifies the image spam by using properties of an image as attributes. He applied file properties and a

Histograms algorithm for image spam detection. This method is called as the FH algorithm. This algorithm is the first part of 2-step image spam classification while the second part is the comparison of the histogram, both gray and color histograms, models are used for image testing. Although the image spam is rapidly grow over the Internet, another unwanted image massage called ham also causes problem to the Internet users. Therefore, [3] proposed an image spam filtering mechanism named content obscuring techniques. This technique is based on the use of image classifiers. The method aims to distinguish between ham and spam images through the low-level characteristics of image texts. Moreover, three main kinds of image texts are determined. First, the presence of small fragments around characters; second, the presence of large fragments around characters, and the last is large background shapes overlapping with characters. A method proposed by [9] identifies the image spam by creating corpus, which is used as the metadata that consist of information such as file name, file size, compressibility and area of image. Although the detection mechanism had been growed rapidly, the detection time plays a major problem. To overcome this [8] proposed an approach that automatically classify an image spam e-mail. As a result we identify either it is a spam or not. A corpus of images was created and used in the classification process. So the advantage of this method is that the image spam can be identified across multiple dataset and classification models. Consequently, the detection process is reduced but the Accuracy to predict an image spam is high. From the researches mentioned above, all the detection mechanisms require 100% accuracy of image mapping between the incoming message and the standard image in the database or corpus. This constrain limits the efficiency of detecting spam because the spam can modify itself from the original pattern to varieties patterns. Thus, the method proposed in this paper will unlock this limitation as described in the following section.

IV. APPROACH

This paper proposes a method to detect a spam from the body of an e-mail, called Partial Image Spam Inspector (PIMSI) using OCR. Thus, the certified e-mail can be distinguished from the spam e-mail. Whenever an e-mail arrives at the e-mail server, it will be sent to the convert classification module to separate the content according to its characteristics which can be either text or images. The result from this classification module will be ended up at the evaluation module where the e-mail will be determined as a certified e-mail or an image spam e-mail by determines its type using two databases, called as corpuses; these two corpuses are a spam keyword and a spam image. Figure 1 shows the system architecture of the PIMSI

International Journal of Research, Science, Technology & Management
ISSN Online: 2455-2240, Volume 3 Issue I, April 2016

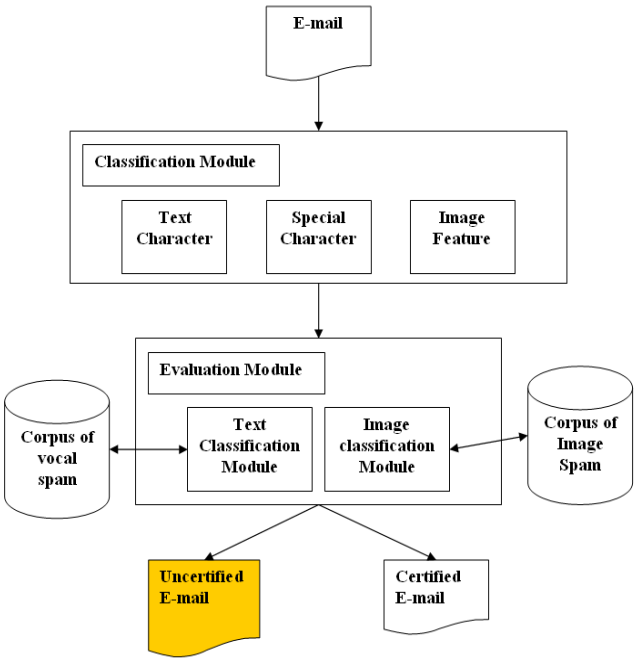

Figure 1. System Architecture of the Partial Image Spam Inspector (PIMSI)

A. OCR

Optical character recognition, usually abbreviated to OCR, is the mechanical or electronic translation of scanned images of handwritten, typewritten or printed text into machine-encoded text. It is widely used to convert books and documents into electronic files, to computerize a record-keeping system in an office, or to publish the text on a website. OCR makes it possible to edit the text, search for a word or phrase, store it more compactly, display or print a copy free of scanning artifacts, and apply techniques such as machine translation, text-to-speech and text mining to it. . Figure 2 illustrates the example of an image spam e-mail.

Figure 2. Sample of an image spam e-mail

OCR is a field of research in pattern recognition, artificial intelligence and computer vision. OCR software application is capable of reading black and white pixels on any image and can distinguish the accurate alpha character or numeric number. This way, the latest OCR technology is quite useful when we need coding our legal papers. When people scan any document it's stored in form of an image, which can't be edited. Whilst OCR has made it possible to scan any printed document and relocate it into word-processing software, such as MS word where we can easily edit it as per our need. OCR is an advanced technology that allows people to alter different sorts of documents including scanned paper documents, PDF files or pictures captured by a digital camera into editable and searchable data as well. Using OCR the system recognizes the static shape of the character. It is necessary to understand that OCR technology is a basic technology also used in advanced scanning applications. Figure 3 shows how when Figure 2 is split into text and image using OCR.

Figure 3. Shows the separate text and image when OCR applied

B. E-mail process

This module includes sending and receiving a mail system. Mail composing page have e-mail address of recipient, subject and the content. All the mail received in the corresponding mail inbox

C. Classify Module

In this module, whenever an e-mail arrives at the e-mail server, the OCR tool converts or extracts the received e-mail according to the content based on its characteristics. Since there are 2 types of outcomes from the previous module, texts and images, then details of each type must be separately defined. When considering the text message obtained from the original image, there are 2 different types of characters: special symbols and keywords. Examples of the special symbols appear as a part of the text message are such as !, %, $, #, and *;

Examples for keywords are such as prices, win, trust, and join now.

D. Corpus

According to a sample of an e-mail in Figure 2, the spam Database, called as corpus, must be divided into 2 types: a corpus of vocal spam and a corpus of object image spam.

1) Vocal spam corpus (keywords)

In this corpus, all keywords of the advertized spam are recorded. A popular resource to collect these keywords is the Internet, such as trash box of web mails, the spam box of web mails, shopping web site, including the Internet site of spam list. Figure 4 shows sample of Vocal spam corpus.

Figure 4. Sample of vocal spam corpus (keywords)

2) Object image spam corpus

C	D	E
call anywhere	dating	earn a college degree
came up a winner	day-trading	earn a degree
career opportunity	debt free	earn big
career singles	degree program	earn extra money
carisoprodol	depression	earning potential
casino	discreet meeting	easy money
casinos	discreet meetings	eliminate your debt
chatroom	discreet ordering	escorts
cialis	doctor approved	
click here	doctor prescribed	

This repository stores only images that are counted as spam. Thus, this corpus consists of small images, and its properties such as RGB colors, contrast, radian, brightness, etc. Figure 5 show some sample image spam in image spam corpus.

Figure 5. Sample of an image spam e-mail

E. Evaluation Module

The evaluation module is the last module in the detection architecture where all texts and images obtained from the e-mail will be identified as a certified message or an image spam message. In doing so, the corpuses defined previously will be used, all converted texts and images will compare with data stored in these corpus.

1) Vocal spam corpus (keywords)

Since there are large amount of spam in the vocal corpus, the searching speed is a significant issue that must be concerned; otherwise, the mailing system will have a huge problem in managing the arrival queue. One searching mechanism that is very efficient when searches in a large amount of a data volume, named the Van Emde Boas tree mechanism. By using Van Emde Boas tree mechanism the search mechanisms of the vocal spam corpus that the best performance of this standard mechanism is $O(\log(\log(n)))$.

2) Object image spam corpus

Content-based image retrieval (CBIR) is the technique of

Computer graphic or computer vision to the image retrieval problem that is the problem of searching for images in to corpus, similar means that the search will be analyzed the true contents of the image. The comparison method starts from retrieving images from the corpus with the CBIR, where all images are searched and analyzed the content of images by comparing all attributes with the image spam attributes in the corpus for example contract, brightness, speckle, radian and etc. if there is three-fourth similarity ratio in between the converted images and the image spam in the corpus, then the received image is identified as the image spam, otherwise, it is green message. There is no need to have 100% mapping when comparing the received image with the corpus image.

V. DISCUSSION

Today spam is unavoidable on the Internet and spams are evolving from text to image over the years each having different attributes. Even though many researchers have given different approaches to detect the uncertified e-mail spam, the problem is the time taken to detect the image spam and also requires 100% accuracy while mapping the spam image with our corpus. In our proposed technique we do not need require 100% accuracy while mapping the received image with the corpus image and the time to detect the receive spam with our data in the corpus is less . The only overhead is constantly searching for image spam signatures over the Internet and storing it in our corpus.

VI. CONCLUSION

Spam is crucial problem across the Internet because it is evolved from text to image. Some of the E-mail spam filtering software could not identify the partial image spam. So this paper proposed a new technique to indentify and avoid the received image spam across the web and this paper does not require 100% accuracy while mapping with the received image spam with the corpus. As a result the uncertified image spam is sent to the spam folder. Finally it relieves the user from worrying about the received image spam.

VII. REFERENCE

[1] B.Battista, F.Giorgio, P.Ignazio, and R.Fabio, “Image Spam Filtering Using Visual Information”, in Proc.14th International Conference on Image Analysis and Processing (ICIAP 2007), Department of Electrical and Electronic Engineering, University. Of Cagliari, Italy, 2007.

[2] B.Battista, F.Giorgio, P.Ignazio, and R.Fabio,“Image Spam Filtering by Content Obscuring Detection ” in Proc. 4th International Conference on E-mail and Anti-Spam(CEAS 2007), California, USA, April 2008.

[3] B.Battista, F.Giorgio, P.Ignazio, and R.Fabio, “Improving Image Spam Filtering Using Image Text Features”, in Proc.5th International Conference on Email and Anti-Spam (CEAS), California, USA. , 2008.

[4] G.Francesco, P.Antonio, PI.Antonio, and S.Carlo, “Using heterogeneous features for anti-spam filters”, in Proc. 19th International Conference on Database and Expert Systems Application, pp.670-674, September 2008.

[5]H.Hailing, G.Weiqiang, and Z.Yu, “A Novel Method for Image Spam Filtering”, in Proc. 9th International Conference for Young Computer Scientists, Zhang Jia Jie, Hunan China, pp.826-830, November 2008.

[6] N.Jordan, M.Daniel, C.D.Nunes, and A.John, “Image Spam-ASCII to the Rescue!” , in Proc. 3rd International Conference on Malicious and Unwanted Software (MALWARE), pp.65-68, 2008

[7] C.Junwei, Z.Lichun, and L.Yueu, “Application of Scale Invariant Feature Transform to Image Spam Filter”, in Proc. 2nd International Conference on Future Generation Communication and Networking Symposia, IEEE Computer Society, vol. 3, pp.55-58, 2008

[8] D.Mark, G.Reuven, and B.E.Ari, “Learning Fast Classifiers for Image Spam”, in Proc.4th International Conference on Email and Anti- Spam(CEAS 2007), Microsoft Research Silicon Valley, Mountain View, California, USA, 2007.

[9] U.Masahiro, and T.Toshihiro, “Design and Evaluation of a Bayesian-filterbased Image Spam Filtering Method”, in Proc. International Conference on Information Security and Assurance (Isa 2008), pp. 46 51, Bussan, Korea, 2008.

[10] H.Peizhou, W.Xiangming, Z.Wei, and L.Xinqi, “Filtering Image Spam Using File Properties and Color Histogram”, in Proc. International Conference on Multimedia and Information Technology, pp.276-279, 2008.

[11] K.Sven, T.Yuchung, G.Jeremy, A.Dmitri, and J.Paul, “ Identifying Image Spam based on Header and File Properties using C4.5 Decision Trees and Support Vector Machine Learning”, in Proc. IEEE Workshop on Information Assurance United States Military Academy, West Point, NY, June 2007.

[12] W.C. Wu, C.T. Kwang, Z.Qiang, and W.L.Yi, “Using Visual Features For Anti-Spam Filtering”, in Proc. IEEE International Conference on Image Processing (ICIP 2005), Genoa, Italy, Vol. 3, pp. III-509-12, September 11-14, 2005.

[13] G.Yan, Y.Ming, Z.Xiaonan, P.Bryan, W.Ying , N.Thrasyvoulos, Pappas, and C.Alok, “Image Spam Hunter”, in Proc. Acoustics, Speech and Signal(ICASSP2008), pp. 1765-1768, March 31 2008-April 4 2008

[14] W. Zhe , J.William, L.Qin, C.Moses, and L.Kai, “Filtering Image Spam with Near-Duplicate Detection”, in Proc. 4th International Conference on E-mail Anti-Spam(CEAS 2007), California, USA, April 2008.