Enhanced ReP-ETD Anti-Spamming Technique

Himanshu Bagwaiya
School Of Information Technology, U.I.T. R.G.P.V, Bhopal, M.P.

Varsha Sharma, PhD
School Of Information Technology, U.I.T. R.G.P.V, Bhopal, M.P.

Sanjeev Sharma, PhD
School Of Information Technology, U.I.T. R.G.P.V, Bhopal, M.P.

ABSTRACT
The Internet is a widely used paradigm where sharing of multimedia content is a major task. Spam image (an image that contains obscure or irrelevant content) is often discovered in web data available in servers and worldwide search engines. Techniques for spam filtering and finding or detecting obscure content in multimedia data (such as .JPEG, .png format of images) are available in the literature. This paper reviews different existing techniques to deal with obscure images and presents an enhanced ReP-ETD (Repetitive Pre-processing technique for Embedded Text Detection) technique in order to detect the obscured content in image data. The technique proposed in this paper first pre-process the multimedia image data using a Linux image script and further on OCR (Optical Character Reader) is used for the spamming image detection and depth analysis. The main contribution of this paper is to discover and perform spam word extraction from the embedded obscured image.

Keywords
K OCR, obscure images, CAPTCH

1. INTRODUCTION
The first spam email is believed to be sent over on May 1978 by a DEC marketing representative named Gary Thuerk to a great number of ARPANET users along the west coast of the US inviting them to a product introduction. Spam is the employment of electronic messaging systems (including to the highest degree broadcast media, digital delivery systems) to transmit unsolicited bulk messages indiscriminately [1]. This article considers the e-mail spam, also recognized as junk e-mail or unsolicited bulk e-mail (UBE) which is a subset of spam that involves nearly identical messages posted to numerous receivers by e-mail [2]. Day by day the amount of incoming spam increase and, spammer attacks are becoming targeted and thus more of a threat.

Current spam filters are built to work on and to resist the various spammer related data, such as a mail attachment, coping body and other signature treated as spam filtering activity. In the past ten years, the machine learning research community has been concerned with the spam filtering task & in particular with e-mail spam detection. Text categorization techniques with the potentially higher generalization capability have been evolved and are now employed in many spam filters. Using Image and OCR image detection tools to pull out content which is text embedded into images, & processing it using text categorization, was exhaustively investigated by the authors in [3]. It was found that this approach can be effective for clean images. A much more uncomplicated approach is based on a keyword search on text extracted using OCR mechanism used for the spammer image detection using text detection technique from the image. The spam intruder started to convert the image into spam image text to defeat OCR tools. For this purpose, it is worth noting that spammers could work to their advantage techniques used to create CAPTCH, which was inserted precisely to defend against robot spamming. Although we think that content obscuring techniques are not likely to be applied in every sort of spam (for example, phishing e-mails should look as if they come from reputable senders, and therefore, should be as "clean" as possible), so that spam trickling modules based on OCR techniques can even be useful, we also argue that different techniques should be invented for the shells in which noise introduced by spammers likely to make standard OCR tools ineffective.

For effective spamming, spammers are employing new and innovative techniques like email appending, image spam, blank image and backscatter spam. Email appending is done if the mailing person or company is having the huge database of customers such as email, contact information & other important information relating to their personals, they can pay to have their database which matched towards the externally available dataset containing electronic mail addresses. The company can commit the marketing mail to the masses who are generally not subscribed to the mails from the composition and may become a victim of receiving spam mails [5].

Image-based spam [6] [4] is an obfuscating method in which the text of the message is stored as a GIF or JPEG image & exposed in the email. This prevents text-based spam filters from detecting and blocking spam messages. Often emails contain the meaningless word, without a subject line or body part or some un-necessary links to navigate towards a different URL which may get the victim of spam directly.

A blank email with a link may get or capture the unnecessary, but important info such as contact email and other information from your mail account.

Blank spam may also come about when a spammer forgets or otherwise fails to add the payload when he or she puts up the spam run. Often blank spam headers appear truncated, suggesting that computer glitches may have taken to this problem with poorly written spam software to malfunctioning relay servers, or any problems that may truncate header lines from the message body. Some spam may come out to be blank when in fact it is not. An example of this is the VBS. Davinia. B an email worm [7] which propagates through messages that hold no open communication channel and appears blank, when in function, it uses HTML code to download other files, load other files.

2. SPAM AND TYPE CAST
In general, spam refers to the role of electronic messaging to send unasked messages to a big group of addresses arbitrarily through e-mail spam is the most widely recognized spam type, it also appears in many other electronic media such as chats, internet telephony, social nets, the web spamming, etc. The cost of transmission of these messages is borne by the users who get it and the ISP who cannot aid the spam traffic and are pressured to increase bandwidth to accommodate the traffic.

The spammers only need to manage the mailing lists that they target. Some common instances of spam are:

- Advertisements that are in the form of pop-ups selling products or making free downloads when we click any link on a World Wide Web page.
- Unsolicited emails with incompatible content, offers, political aspects
- Redundant calls on IMs like Skype offering mortgages, loans with low-interest rates
- Links on social nets that adopt you to free downloads, easy income, pornography
- Unsolicited text messages offering loans, low priced products etc.

2.1 Typecast of spam

- E-mail Spam
- Instant Messaging and texting
- Search Engine spam

2.1.1 E-mail Spam:

Also known as unsolicited bulk email, it is the most common form of spam we see. Mostly the motive behind these messages is to advertise and sell products, steal information or phishing, express political views, pornography & malware injection. The initiative e-mail spam is said to be sent in 1978 by DEC by sending an invite to all ARPANET addresses inviting them to the reception of their new DEC-20 machine.

2.1.2 Instant Messaging and texting:

Spam is also there in instant messengers like Skype, yahoo messengers generally come in the form of friend requests from strange people. It is less sparse than the e-mail spam. Text spam is promotional offers, advertisements.

2.1.3 Search Engine spam (spamdexing):

It refers to spamming web pages to falsely increase their page ranking results in the browser.

2.2 Image spam

Image spam is a variant of email spam where the spammers actually embed the spam message in an image instead of directly putting it as mail content to circumvent junk email filters. Spam filters look for certain keywords like Viagra, cash, money which are ordinarily linked to spam emails. Even so, when the message is inside an image the spam filters cannot effectively filter these messages. There are many techniques which spammers use to obfuscate spam filters like-

- Adding random words before HTML
- Use white text on a black background
- Using characters like M*oney
- Adding bogus HTML tags with a heap of text
- Adding spaces in words like "l o w I n t e r e s t R a t e"

As stronger filters were developed to cut through these messages, spammers came up with more novel techniques like image spam, using pdf documents to send spam etc. With the use of Optical Character Reader (OCR) filters it was potential to take out the contents of the images and then suss out if the image had spam content.

3. RELATED WORK

Among the developed techniques to block spam, filtering is a significant and popular one. An exercise applied for mail filters admit organizing email & removal of unsuitable content and computer viruses. A less common purpose is to inspect outgoing email at some companies to assure that employees pursue with appropriate laws. Users might also use a mail filter to prioritize messages, and to separate them into folders based on subject matter or other standards. Mail filters can be set up by the user, either as separate programs or as part of their email program (electronic mail client). In email programs, the user may apply for spam filter technique manually. The author in research Bob West [6], have delivered the performance of five commonly used machine learning methods in spam filtering. Most electronic mail programs now also suffer an automatic spam filtering function.

Chao Wang, Fengli Zhang [13] have used low-level features like elevation, aspect ratio, image width and file size which are all extracted from image header and employed with another set of feature and used SVM classifiers to classify images. Yan Gao, Ming Yang [14] used a semi-supervised approach of regularized discriminant EM algorithm, to detect image spam emails.

Fumera, Giorgio, Ignazio Pillai [6] [15] have proposed an approach to antispam filtering which exploits the text information embedded into images sent as attachments. Their approach is based on the application of state-of-the-art text categorization techniques to the analytic thinking of text extracted by OCR tools from images attached to emails. The strength of the suggested approach was experimentally assessed on two huge corpora of spam e-mails.

While going through email communication, many users have also experienced email spam. Congfu Xu,Kevin Chiew Chen and Juxin Liu [16] have proposed a hybrid image spam filtering framework to identify spam images based on both image features and extracted text.

Preventing text recognition using OCR tools and imposing various challenges in filtering different types of spam. Basheer Al-Duwairi et al. [17] have proposed an image spam filtering technique by using low-level image feature for image characterization. Using different classifiers like C4.5 Decision tree, Multilayer perception, Naïve Bavs Random forest and SVM.

Detecting image spam turns out to be an interesting example of the problem of substance-based filtering of multimedia data in adversarial environments, which is gaining increasing relevance in various applications and media. Battista Biggio, Giorgio Fumera [4] gave a comprehensive study and classification of computer vision and pattern recognition techniques proposed so far against image spam and constructed an experimental analysis and comparison of some of them on real, publicly available data sets.

Image spam poses a large threat to email communications due to high volumes, bigger bandwidth requirements, and higher processing requirements for filtering. Krasser, S., Yuchun Tang, Gould J [8] represented a feature extraction and classification framework that operated on features that can be extracted from image files in a very quick way. The characteristics considered were thoroughly analysed involving their data gain, with classification performance results for C4.5 decision tree and support vector machine classifiers. Lastly, they compared the performance that can be

accomplished using these fast features to a more composite image classifier operating on morphological features extracted from fully decoded images. The proposed classifier is able to find a large measure of malicious images while being computationally inexpensive.

Aradhaye et al. detected text embedded in digital photos. The text was analysed on extraction and features like colour saturation, colour heterogeneity was computed. SVM classifier was used to classify images. An accuracy of 85% [9] was achieved.

Mark Dredze, Reuven Gevaryahu and Ari Elias Bachrach [10] introduced Just in Time (JIT) feature extraction, which builds features at classification time as needed by the classifier. They demonstrated JIT extraction using a JIT decision tree that further increase system speed. Image spam classification provides accuracy of 99% and a method to learn fast classifiers.

Enhanced ReP-ETD Anti-image spamming is a pre-processing over obscure image data and image data set, by which improved result is obtained by text detection from the anti-image spamming tool. Rep-ETD proposed by Asha. S. Manek [11] is the pre-processing technique involving several steps. During its operation, all the corrupt and invalid images are to be removed from the information set, by which OCR tool will pick out the picture. After conversion of all images to grayscale to reduce noise overhead, all detected text is to be inserted into text transformation to lowercase so that textual sensitivity will be solved and then all types of filtering are done later on this. Now all collected words are to be set in the 'bag of keywords' so that successfully distinguishes between ham and spam. Choice of words is done by Bayes Theorem for calculating the probability of a natural event. Repeating all this process resulted a number of times of the keyword show the type of image spam/ham.

4. PROPOSED METHODOLOGY

The primary step in the Enhanced ReP-ETD includes a focused pre-processing stage over obscured image data, followed by classification and performance analysis.

Step 1: Removal of corrupted images from the data set that is taken for analysis and resize it into a processable format. To ensure that we do not use duplicate images in our training set we use a script used in Artificial neural networks as a tool for identifying spam [12] to compute the checksum of each single picture. The checksum is a mere way to check data integrity during transmission of data. MD5 (Message-Digest algorithm 5) is a useful cryptographic hash function with a 128-bit hash value.

Step 2: Personal data set of obscure image content is to be pre-processed with the assist of a Linux script of text cleaner.

Step 3: Removal of images which is not known as a valid image format of our OCR tool.

Step 4: Convert the entire obscured selected image into the desired format for OCR.

Step 5: Conversion of all images into grayscale to reduce the noise overhead.

Step 6: The OCR tool used in our work is Tesseract OCR API with default settings to retrieve the images in Different image formats for further processing in classification. Applying this technique, we ensured manually that there is no loss of data by our OCR technique.

Step 7: The crucial stage is to take the 'bag of keywords' which successfully distinguishes between spam and ham.

Step 8: Selection of keywords is done by using Bayes theorem for computing the probability of occurrence of "to be called spam words" from the extracted text embedded in the pictures.

Step 9: The occurrence of one of the keywords from the bag of keywords and the number of times it takes place is seen into consideration to determine the nature of input image (as spam /ham).

5. FLOW DIAGRAM

The process flow chart of proposed work is described as follows in figure 1:

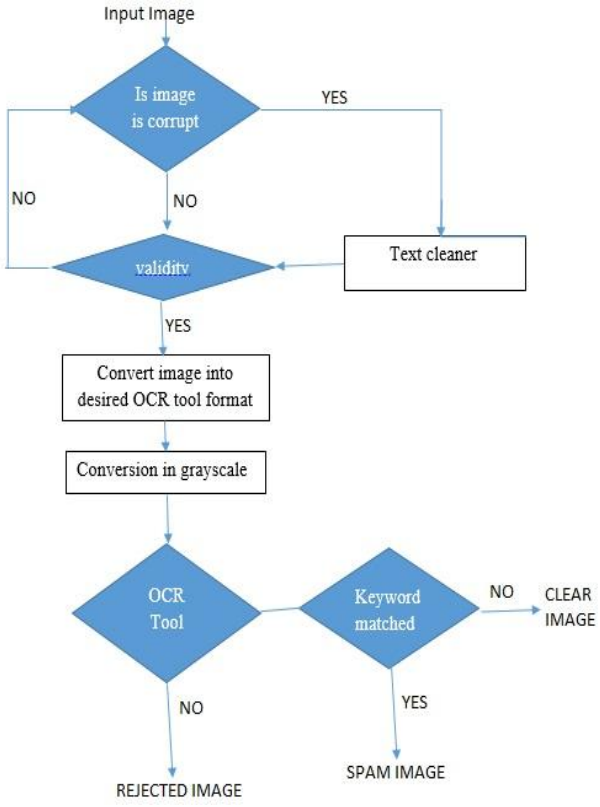

Fig 1 Flow chart of Enhanced ReP-ETD Anti-Image spamming technique

6. EXPERIMENTS AND RESULTS

Data Source: In order to perform analysis data source is to be extracted from Dredze, ISH and the personal data set as shown in table 6.1

Table 6.1 Statistics of the images collected to form the corpus

Dataset Source: Image dataset Identification	Ham Images	Spam Images
Dredze	1000	3171
ISH	784	896
Personal Dataset(Obscured Dataset)	163	191

Table 6.2 Comparison of existing algorithm and proposed algorithm

Algo Name	Accuracy	Detection Rate	Precision	Recall
Existing algo	70.972645	83.50354	72.307694	83.50253
Proposed algo	73.95219	68.57823	76.75841	58.57923

In order to understand the patterns and trends in image spam above table show various pre-processing steps done over normal spam image and obscured images, results of precision, accuracy and recall are calculated in table 6.2.

7. CONCLUSION

In this work an approach for anti-spam filtering, which exploits the text knowledge embedded into images sent as electronic mail attachments, is proposed. This approach extracts embedded text from attached images. The effectiveness of this approach has been evaluated more efficiently by evaluating some special kind of obscured images. For future study, analysis of obscure images from other data sets can be done.

8. REFERENCES

[1] Wikipedia,"Spam"http://en.wikipedia.org/wiki/Spam_(electronics)

[2] Wikipedia,"Emailspam"http://en.wikipedia.org/wiki/Emailspam

[3] Hope, P., Bowling, J. R., and Liszka, K. J., "Artificial Neural Networks as a Tool for Identifying Image Spam", The 2009 International Conference on Security and Management (SAM'09), pp. 447-451, July 2009

[4] Battista Biggio, Giorgio Fumera, Ignazio Pillai, Fabio Roli, "A survey and experimental evaluation of image spam filtering techniques, Pattern Recognition Letters". Volume 32, Issue 10, pp 1436-1446, ISSN 0167-8655, 15 July 2011

[5] Bob West, "Getting it Wrong: Corporate America Spams the Afterlife", Clueless Mailers. (January 19, 2008) Retrieved 2010-0923

[6] Fumera, Giorgio, Ignazio Pillai, and Fabio Roli. "Spam filtering based on the analysis of text information embedded into images", The Journal of Machine Learning Research vol.7, pp. 2699-2720, 12/2006.

[7] Symantec.com, "symantec.com", Retrieved 2012-12-10.

[8] Krasser, S.; 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. IAW '07. IEEE SMC, vol., no., pp.255-261, 20-22 June 2007.

[9] Aradhye, Hrishikesh B., Gregory K. Myers and James A. Herson "Image analysis for efficient categorization of image-based spam email." In Document analysis and Recognition, 2005. Proceeding. Eighth International conference on, IEEE pp. 914-918, August 2005

[10] Mark Dredze, Reuven Gevarvahu and Ari Elias-Bachrach, "Lerning fast classifers for Image Spam", CEAS 2007

[11] Asha S Manek, Shamini D.k, Bhat and Shenoye" ReP-ETD: A Repetitive Preprocessing Technique for Embedded Text Detection from Images in Spam Emails"978-1-4799-2572-8/14/$31.00c@2014 IEEE

[12] Hope, P., Bowling, J. R., and Liszka, K. J., Artificial Neural Networks as a Tool for Identifying Image Spam, The 2009 International Conference on Security and Management (SAM'09), July 2009, pp. 447-451.

[13] Chao Wang, Fengli Zhang, Fagen Li, Qiao Liu, "Image spam classification based on low-level image features", Communications, Circuits and Systems (ICCCAS), 2010 International Conference on, pp.290-293, 28-30 July 2010.

[14] Yan Gao, Ming Yang, and Alok Choudhary, "Semi Supervised Image Spam Hunter: A Regularized Discriminant EM Approach", In Advanced Data Mining and Applications, pp. 152-164. Springer Berlin Heidelberg, 2009.

[15] Giorgio Fumera, Ignazio Pillai and Fabio Roli, "Spam Filtering Based On The Analysis Of Text Information Embedded Into Images", Journal of Machine Learning Research 7 (2006) 26992720, Submitted 3/06; Revised 9/06; Published 12/06

[16] Congfu Xu, Kevin Chiew, Yafang Chen and Juxin Liu, "Fusion of Text and Image Features: A New Approach to Image Spam Filtering", Y.Wang and T. Li (Eds.): Practical Applications of Intelligent Systems, AISC 124, pp. 129–140. springerlink.com Springer-Verlag Berlin Heidelberg 2011.

[17] Basheer Al-Duwairi, Ismail Khater and Omar Al-Jarrah, "Detecting Image Spam Using Image Texture Features", International Journal for Information Security Research (IJISR), Volume2, Issues3/4,pp344353,September/December 2012.