Extraction of Advertising Words from Text and Image based Spam mails for Classification

Rahul Bansod
MTech 2nd year
BDCOE, Sewagram

R. S. Mangrulkar
Associate Professor
BDCOE, Sewagram

V. G. Bhujade
Assistant Profe
BDCOE, Sewagram

ABSTRACT
Email is a very popular way of communicating with others over the internet. As number of internet users are growing rapidly, many people are finding, email communication an inexpensive way to send their data and for the communication. Almost every website ask for email id so as to complete their registration and making users more and more prone to get affected by the spam mails. These uninvited bulk emails occupies consumes large amount of network bandwidth and it also requires server storage space. Recently, Image spam is kind of spam invented by the spammers where advertising details are specified in the image or picture files. In the proposed system the technique of extracting promotional (Advertising) words from text and image based spam is discussed.

Keywords
Spam, Image Spam

1. INTRODUCTION
Internet has increasingly become a favorite support of malicious programs. The e-mail communication is very attractive for direct marketers due to the increasing number of internet users and its low cost. In the past few years the volume of unsolicited commercial mails has grown enormously. These unsolicited bulk emails are termed as Spam. Anti-spam filters are posing a great challenge in detecting these spam emails where advertisement text embedded in images [1]. Text-based spam filtering techniques are failed to detect the spammer's new approach. Spammer's advertisements have become a part of an embedded image file attachment rather than the body of the e-mails. E-mail management has become a vital and growing problem for individuals and organizations as it is prone to misuse. The invasion of image spam into email has created problem for spam classifiers. Statistics says , up to 25% of spam being sent today are image spams and this number is expected to increase soon. Therefore, it is desirable to develop a systems to detect and filter image-based spam.

An optical character recognition (OCR) system [4] that extracts and recognizes embedded text, followed by a text classifier , is one of the possible way to detect image-based spam.

(a) Image containing only text (b) Image with photographic elements

Fig 1. Image spam

(a) Natural scene photo (b) Greetings e-card

Fig 2. Normal image

To overcome the issue of email spam , spam filters[3] are introduced. One of the methods of filtering spam is using neural networks to intelligently classify an email as a spam or a ham. In order to use a Neural Network it has to be trained first to get the optimal weights.

2. RELATED WORK
Today's internet world is facing many challenges for image spam classification [5]. Many algorithms have been proposed for classifying spam and legitimate emails[6]. The fidelity of the spam classifiers introduced by service providers are handled by the spammers through various randomization techniques. However many spam classification approaches have been put forth by the researchers to assist the developers of anti-spam detectors. Liu Ming et al. proposed a scheme [7] for filtering the image spam based on the method of spam behavior recognition filtering. A model is developed based on Bayes technique which identifies spam according to the behavior of mail sent. This approach adopts filtering of spam by stages, by utilizing the least risk of Bayes technique in order to recognize the image spam as early as possible. Ngo Phuong et al. [8] proposed a fast method to detect spam images. The proposed method uses an edge-based feature vector, which can be computed efficiently, to represent major shape properties of the image instead of extracting embedded text from an image. This method uses the edge-based feature to compute a vector of similarity measures from an image to a small set of gold standards. These similarity vectors are then served as input for SVMs training and classification. This method is fast because it does not use computationally expensive image processing and text recognition steps. Ms. Soranamageswari et al. [5] have concentrated on the measures working on statistical image feature histogram and mean value of a block of image. Based on color histogram and mean value of a block of an image a classifier is trained, trying to classify spam images from legitimate ones with minimal effort.

3. THE HARMS CAUSED BY SPAM
With the rapid development of Email services, spam messages are increasing rapidly, and the contents of spam messages are related to all aspects of life. The numbers are very large . The problems of spam messages has been seriously disrupting people's normal work and life.. Spam messages brought a very bad influence on social harmony.

1) Biggest nuisance on the net which affect the social stability

Spam messages are causing serious problems which flood our email boxes in quick time. Companies are wasting much time on detecting and removing spams.

2) To interfere with normal communication

Spam messages can be send in mass, so the transmission time will take up network bandwidth, causing congestion, affecting the performance of the network and people's normal communication.

Internet usage is expected to continue growing and e-mail has become a powerful tool intended for idea and information exchange. Negligible time delay during transmission, security of the data being transferred, low costs are few of the multifarious advantages that e-mail enjoys over other physical methods.

4. PROPOSED SYSTEM

An ongoing challenge is to develop a classifier that can distinguish spam and ham.

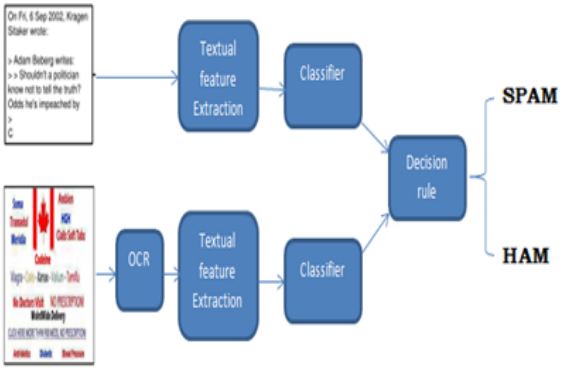

Fig 3. Proposed Approach for Spam Classification

4.1 Black Listing and White listing

All those web pages and domain that are widely known for sending spam mails and are not trusted, go onto the black list. If a domain that matches from this list, the mail is predicted spam without any further processing.

4.2 Extracting words from Image

Image spam is a kind of email spam where the message text of the spam is presented as a picture in an image file. Users have an option of attaching image to their mails. The image is passed through the OCR tool for extracting words from image. Prime accuracy is achieved for a strong resolution image and more common fonts like Times New Roman.

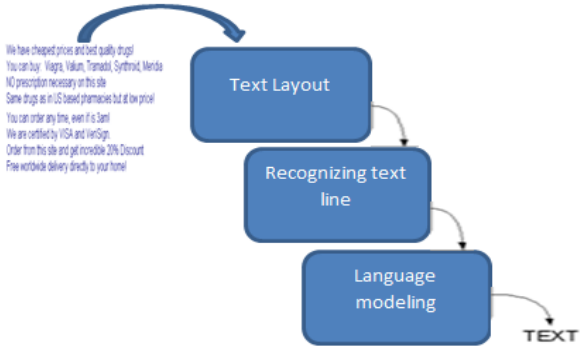

Fig 4. Architecture of OCR

The OCR is a feed-forward system with three major components: layout checking, text line recognition, and statistical language modeling;

- Layout analysis is responsible for identifying text lines, text columns, text blocks and reading direction.
- Text line recognition is accountable for recognizing the text contained within each vertical or right-to-left line.
- Statistical language modeling assimilates recognition guesses with former knowledge about language, vocabulary and grammar of the document.

4.3 Pre-processing of Dataset

Both datasets require the preprocessing before giving it to the system as the mails available in the datasets are not in proper format for training the system for classification. Two sets for each dataset is created for training and testing.

Messages are usually sent in the form of plain text or HTML marked up text. First strip the message of all HTML tags if the message is not plain text. Mail headers are also separated by creating separate features for each of the mail headers. While tokenizing messages, the words are converted to a single case. Punctuation is removed and all other tokens are broken down into constituent words.

```
Message-ID: <413-220028422154219900@freesource>
X-Priority: 1
To: "1" <thecashsystem@firemail.de>
From: "TheCashSystem" <Thecashsystem@firemail.de>
Subject: RE: Your Bank Account Information
Date: Thu, 22 Aug 2002 10:42:19 -0500
MIME-Version: 1.0
Content-type: text/plain; charset=US-ASCII
X-MIME-Autoconverted: from quoted-printable to 8bit by webnote.net i
Content-Transfer-Encoding: 8bit

------------------------------------------------------------------
A POWERHOUSE GIFTING PROGRAM You Don't Want To Miss!

  GET IN WITH THE FOUNDERS!
The MAJOR PLAYERS are on This ONE
For ONCE be where the PlayerS are
This is YOUR Private Invitation

EXPERTS ARE CALLING THIS THE FASTEST WAY
TO HUGE CASH FLOW EVER CONCEIVED
Leverage $1,000 into $50,000 Over and Over Again

THE QUESTION HERE IS:
YOU EITHER WANT TO BE WEALTHY
OR YOU DON'T!!!
WHICH ONE ARE YOU?
I am tossing you a financial lifeline and for your sake I
Hope you GRAB onto it and hold on tight For the Ride of youR life!
```

Fig 5. Spam email from Dataset

4.4 Preprocessing of Data

Stemming is the process of moving any word to its root value. Some steps of this process are:

- Remove the plurals and -ed or –ing suffixes
- Deal with suffixes , -full, -ness etc.
- Take off -ant, -ence etc.

Many words in natural language occur with high frequency but have low information content, such as ”a,” ”an,” ”the,” and most prepositions and conjunctions can be removed with the assumption that no serious loss of information will occur. These so called stop words are specified in a list and can be removed from the token stream. A database of all the words that occur in each mail with the frequency of the word stored in each column is maintained. For each preprocessed mail, TF (Term Frequency) is calculated.

	id	word	
	1	save, free, family, email...	64 b...
	2	ilug, adclick, ws, http, ...	55 b...
	3	www, adclick, o, ws, s, c...	36 b...
	4	advertisers, don, receive...	229 b...
	5	com, www, freeyankee, cgi...	60 b...
	6	don, received, free, time...	156 b...
	7	don, received, free, time...	156 b...
	8	home, com, link, www, loo...	44 b...
	9	com, save, jm, netnoteinc...	130 b...

Fig 6. Database of extracted words

After the database with the stemmed words, with each mail name in one column and the frequency of occurrence of words will move on to the next phase for classification.

4.5 Classification

The words with the highest frequency in a document will be supplied to the neural network for the classification of spam or ham mail. The overall weight of the document will be considered by the neural network for the decision making.

5. DATASET

Reliably classified and widely available email data sets are very difficult to obtain. Most developers use their own private data and do not release it for public use. For text based spam detection, SpamAsssassin [9] is the open source dataset which contains 500 mails of various category. For image based spam detection, SpamArchieve[10] is the open source dataset which contains more than 700 images of various category and format.

6. CONCLUSION

This paper discourses a side effects of the spam mails and proposes a solution to detect this problem. The mails could be text based or image based. In this technique a neural network will be train to be able to distinguish between spams and legitimate emails. OCR engine extracts text from image. After the text extraction from image the classification process is almost same as the text mail. This approach can efficiently perform classification of spam mails.

7. REFERENCES

[1] Harisinghney A. ; Dixit A. ; Gupta S. ; Arora A. “Text and Image based spam email classification using KNN, naïve Bayes and Reverse DBSCAN algorithm” Optimization, Reliability and Information Technology(ICROIT) , 2014 International Conference on DOI:10.1109/ICROIT. 2014. 6798302, page(s):153-155, 2014

[2] J. D. Brutlag and C. Meek, “Challenges of the email domain for text classification,” in ICML, 2000, pp. 103–110.

[3] N. Nhung and T. Phuong. "An Efficient Method for Filtering Image-Based Spam E-mail". Proc. IEEE International Conference on Research, Innovation and Vision for the Future (RIVF07), IEEE Press, Mar. 2007 , pp. 96-102.

[4] R. Smith, "An Overview of the Tesseract OCR Engine," in Proc. International Conference on Document Analysis and Recognition, 2007

[5] M. Soranamageswari and Dr. C. Meena “Statistical Feature Extraction for Classification of Image Spam Using Artificial Neural Networks” Second International Conference on Machine Learning and Computing, 2010

[6] Ms. D. Katrina Renuka and Dr.T.Hamsapriya “ Spam Classification based on Supervised Learning using Machine Learning Techniques” Process Automation, Control and Computing (PACC), 2011 International Conference on DOI: 10.1109/PACC.2011.5979035, 2011, Page(s): 1 – 7

[7] Liu, G., & Yang, F. “The application of data mining in the classification of spam messages” In Computer Science and Information Processing (CSIP), 2012 International Reverse Conference on (pp. 1315-1317), IEEE.

[8] N. Nhung and T. Phuong. "An Efficient Method for Filtering Image-Based Spam E-mail". Proc. IEEE International Conference on Research, Innovation and Vision for the Future 10.1I09 /RIVF.2007.369I41.

[9] http://spamassassin.apache.org/publiccorpus

[10] www.cs.jhu.edu/~mdredze/datasets/image_spam/spam_archive.tar.gz.