ANALYSIS OF AN IMAGE SPAM IN EMAIL BASED ON CONTENT ANALYSIS

Meghali Das[1] and Vijay Prasad[2]

[1] Dept. of Computer Science & Engineering and IT, Don Bosco College of Engineering and Technology, Guwahati, India
[2] Dept. of Computer Science & Engineering and IT, Don Bosco College of Engineering and Technology, Guwahati, India

ABSTRACT

Researchers initially have addressed the problem of spam detection as a text classification or categorization problem. However, as spammers' continue to develop new techniques and the type of email content becomes more disparate, text-based anti-spam approaches alone are not sufficiently enough in preventing spam. In an attempt to defeat the anti-spam development technologies, spammers have recently adopted the image spam trick to make the scrutiny of emails' body text inefficient. The main idea behind this project is to design a spam detection system. The system will be enabled to analyze the content of emails, in particular the artificially generated image sent as attachment in an email. The system will analyze the image content and classify the embedded image as spam or legitimate hence classify the email accordingly.

KEYWORDS

Spam Filtering, Image Spam, Content Based Filtering

1. INTRODUCTION

As the scope and use of Internet grows the type of information has been more multimedia enriched to attract larger number of users. Electronic mail is currently the most efficacious and sought-after mode of communication. However, like any other dynamic medium, it is prone to misusage. Such an instance of misuse is the blind posting of unwanted email messages, also known as spam, to large and random recipients. Spam messages are sent using bulk-mailers and address lists gathered from web pages and newsgroup annals. Radicati in the year 2009 estimated that 247 billion email messages were sent per day predicted to double by 2013 [Radicati 2009] [1, 2].

Spam, or unsolicited bulk mail, though sent out in various shape and form, nevertheless, may possess a number of similar characteristics in terms of structure, content, and distribution approaches. The generation and distribution is justified from a spammer's prospect as the effort and cost involved in sending a large number of emails is minimal and the probable return considering the large number of email users is huge. According to a survey the overall cost incurred in preventing spam in 2009 was estimated to be 130 billion U.S. dollars [1].

Like any other electronic means of communication the genre of email content has been upgraded from text-based to visual-based or combinations of the two. Consequently, this has greatly reduced the effectiveness of existing text-based anti-spam filters. As a matter of fact, legitimate message-senders have sought to enrich the messages by adding multimedia content like images. The spammers at the same time have started using images to hide the fraudulent messages and

10.5121/ijnlc.2014.3313

combat the text based anti-spam filters at the same time by using HTML formatting [3]. Understanding the increasing trend of multimedia enriched email messages, it is the need of time to use such information to defeat the spammers' intention.

Image based spam is a quantum leap from the spammers' perspective. It involves in embedding the spam message into images which are sent as email attachments. The aim is to surpass the spam filters whose analysis is based on only the emails' textual content. Such spam emails may get misclassified by the filtering program, but the hidden message becomes visible to the recipients when opened by them. Usually, spam images are generated by employing discrete changes to a template image making signature-based detection techniques ineffective, and are obscured to prevent optical character recognition (OCR) tools from reading the embedded text [4].

This work proposes a framework for a spam detection system. The proposed detection system attempts to extract embedded text together with the visual feature like color, texture, shape and hence used to calculate a similarity measure with a query image. The extracted features are used to train a classifier which would work online in labelling an incoming message as spam or legitimate. The paper is structured as follows. The basics of spam message and spam filtering are mentioned in Section 2. Section 3 emphasizes the prevalent spam filtering techniques. A comparison of existing image spam detection methods is focused in Section 4. Finally, in Section 5 a spam detection system is proposed focusing on the attached spam images.

2. Spam Message and Spam Filtering

The Spam Track at the Text Retrieval Conference (TREC) defines email spam as [5]:

"Unsolicited, unwanted email that was sent indiscriminately, directly or indirectly, by a sender having no current relationship with the recipient."

Email spam, also known as unsolicited bulk Email (UBE), junk mail, or unsolicited commercial email (UCE), can thus be defined as the system of sending innumerable undesired email messages, featuring commercial content to an indiscriminate set of recipients.

In general, the following features qualify an email to be classified as spam [6]:

- Unsolicited: The receiver is not interested in receiving the information.
- Unknown sender: The receiver does not know and has no link with the sender.
- Massive: The email has been sent to a large amount of addresses.

Spam, has recently posed to be a serious problem for email users. Several anti-spam filtering solutions have been proposed till date. In general, these approaches treat the email spam filtering problem as a text classification or categorization problem, employing various machine learning techniques to solve the problem. Image spam is the most recent trap developed by spammers. It is a simple and effective way of cheating spam filters since they can detect only text. An image spam email can be defined as a HTML formatted document, which usually constitute non-suspicious text and an embedded image sent as an attachment. The message is conveyed by the embedded image and most email clients show the message with full totality.

2.1. Spam filter

Spam filtering is designed to distinguish between legitimate and spam. Spam filters are a specialized technical fix against spam which helps end-users to keep their mailboxes clean. Spam filters can be operated on Internet Service Providers (ISPs), email servers, or users' email

clients[7]. They consist of several modules which analyze different characteristics of input emails, like address of the sender and the recipient, textual content, header format and mail attachments. The output of each module is combined to label an email as spam or legitimate.
A spam filter is defined to be an automated technique to identify spam. The decision of an ideal spam filter is based on features like the content of the message, characteristics of the sender and the receiver, knowledge as to whether the receiver or others consider the particular genre of messages as spam, or the sender as a spammer, etc. But perfect training is non-existent and it is therefore necessary to constrain the filter to use well-defined information sources such as the content of the message itself[5]

In general, a spam filter can be defined as an application which implements a function [3]:

$$f(m,\theta) = \begin{cases} c_{spam}, \textit{ if the decision is "spam"} \\ c_{leg}, \textit{ otherwise} \end{cases}$$

where m denotes the message to be classified, a vector of parameters is represented by θ, and c_{spam} and c_{leg} signify the labels to be assigned to the messages.

Usually, spam filters are designed on the basis of machine learning classification techniques. In such a technique the vector of parameters θ is the result of training the classifier on a pre-collected dataset [4]:

$$\theta = \Theta(M),$$
$$M = \{(m_1, y_1), ...(m_n, y_n)\},\ y_i \in \{c_{spam}, c_{leg}\},$$

where m_1, $m_{2,}...m_n$ represent the set of collected messages, y_1, $y_2...y_n$ signify the corresponding labels already assigned, and the training function is denoted by Θ.

2.2. Structure of a spam filter

Incoming messages are handled by the filter one at a time and classified as legitimate or spam. Legitimate messages are destined to the recipient's inbox which is read frequently. Spam is quarantined which is infrequently searched for any misclassified legitimate messages. If any misclassification is found—either spam in the inbox or legitimate message in the quarantine — the errors may be reported to the filter to improve its performance. The filter while classifying a message, make use of the content of the message, its built-in knowledge and algorithms, and its memory of previous messages, feedback from the user, and external resources such as blacklists or reports from other users, spam filters, or mail servers[5]. The filter may either work on the user's system or on a server serving the same purpose for multiple users at a single time.

3. SPAM FILTERING TECHNIQUES

Content-based filtering is one of the most popular technical methods used to combat spam. Internet users opt for spam filters that classify messages on the basis of analysis of the contents of the messages. The positive outcome of content-based filters has forced spammers to derive increasingly complex attacks which can surpass these filters and reach the users mailbox. These filters may involve hand-made rules, also known as heuristic filters, or trained using Machine Learning algorithms.

Learning-based filters possess the learning capability from spam and legitimate example messages which allows these filters to customize the spam detection. Learning-based filters also have the potential to learn and enhance the self- performance at real-time, as they can adapt themselves to the wide genre of spam and legitimate email a user receives. The spammers adopting complex and costly spamming techniques illustrate the success of learning-based filters establishing such filters as the current state of the art of email filtering.

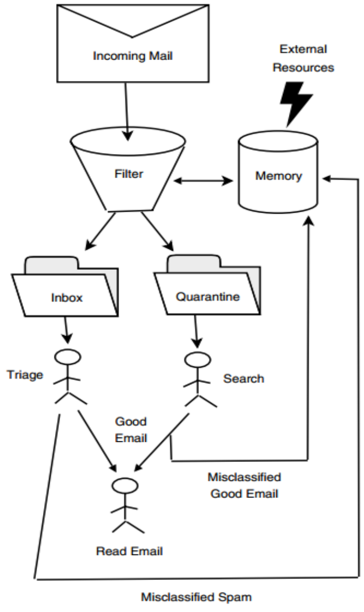

Figure 1. A typical spam filter [5]

3.1. Learning Algorithms

The most important part in a spam classification system is the learning algorithm. A large genre of learning algorithms have been used in spam classification, like the probabilistic Naïve Bayes, rule learners like Ripper, Instance Based k-Nearest Neighbors (kNN), Decision Trees like C4.5, linear Support Vector Machines (SVM) classifiers committees like stacking and Boosting, and Cost Sensitive learning.[6]

3.1.1. Probabilistic Approaches

Probabilistic filters are the frequently used filters for spam classification because of its simplicity and the accuracy achieved. These classifiers are based on the Bayes' Theorem which computes the probability for a document *d* to belong to a category c_k as [6]:

$$P(c_k|d)=\frac{P(c_k)\cdot P(d|c_k)}{P(d)}$$

When used for determining probabilities for spam classification the denominator can be ignored since there are only 2 categories (spam and legitimate), and one document can be classified only into one of them, so the denominator remains the same for every message k.

3.1.2. Decision Trees

The main disadvantage of the probabilistic approach is that the results are confusing. In the domain of Machine Learning, there exist some learning algorithms, which achieve interpretable results. One such algorithm is the Decision Tree family of learners.

A Decision Tree is defined as a finite tree structure where branches represent the tests, and the leaves denote the categories. The classification is done by expanding from root to leaf in the tree, and only selecting conditions in branches that are evaluated as true. Evaluations are repeated until a leaf is reached, assigning the document to the category that denotes the leaf reached.

There are many algorithms used for computing the learning tree. The most popular ones are ID3, C4.5 and C5.

3.1.3. Rule Based Learners

Conditional rules are the basis of the concept description languages most popular in the Machine Learning domain.

Rule based learning algorithms involve conditional rules consisting of a logic condition as the premise and the consequent as class name. The premise is usually a Boolean expression assigning weights for words that appear in the document representation. For binary weights, conditions stand rather simplified for binary classification where rules contain premises, if certain combination of terms appears or not in the document. Ripper is the most popular and effective rule learner applied to spam filtering.

3.1.4 Support Vector Machines

Support Vector Machines (SVM) is very popular in spam classification considering the accuracy with these algorithms. SVMs are defined as an algebraic formula generating maximum margin hyper-plane to separate training instances, with polynomial kernels.

Training instances need not be linearly separable. The main goal is to build a hyper-plane for separation. The basic form of hyper-plane can be generated as a linear function of the attributes. SVMs are fast to learn and highly effective in learning-based spam filters.

3.1.5 K-Nearest Neighbors

Another alternative approach in learning algorithms involves storing training instances after being pre-processed and represented. Consequently, when a new instance needs to be classified, it is matched to the stored documents and assigned the more appropriate category based on its similarity to those stored in every class.

The most popular one in this category is k-Nearest Neighbors (kNN). The value of k designates the number of neighbours used for classification. A significant step of this method is the choice of similarity function between messages. The method frequently used to compute the similarity measure between messages is the “cosine distance”, where cosine is defined as the angle between the vectors representing the compared messages. This distance function normalizes the length of the messages, and hence considered effective.

3.1.6 Classifier Combinations

The Classifier Combinations approach is based on the concept of implementing various methods to the same data and the output merged to obtain single and best results. Bagging, boosting and stacking are major techniques of this category. The theory of bagging is based on the idea of combining the different predictions of multiple models, or the same model for various genre of learning data.

Another advanced Machine Learning procedure which provides the concept of weighted prediction, also called voting, is the Boosting technique. The concept of boosting (applied to spam detection by Carreras et al. [2001])[6] generates multiple classifiers (for prediction or classification), extracting weights and uniting the predictions from these classifiers into a single predicted classification.

The idea of stacking (Stacked Generalization)[8] is another approach of combining the predicted output of several models. It is especially used for situations where the type of classifiers included is very distinct. It is evident that integration of the predictions from disparate methods often would yields higher accuracy compared to the results obtained from a single method[6]. The predictions from different classifiers are input into a meta-learner, which combine the predictions to generate a final best predicted classification.

3.1.7 Cost-Sensitive Learning

In the domain of spam filtering, the cost incurred of a false positive (a legitimate message classified as spam) would be higher than a false negative (a spam message classified as legitimate). This can be illustrated as the risk of missing significant and genuine messages because messages considered spam be removed or, preferably saved in a quarantine that can be later searched.

Thresholding is another method for making algorithms cost-sensitive. A prediction classifier is generated using a set of pre-classified instances (the training set). This classifier is then used to calculate a threshold that optimizes the cost on a different set of pre-classified instances (the validation set). When a new instance needs to be classified, the threshold decides the category of the instance as positive (spam) or negative (legitimate). The cost is computed using a cost matrix that assign cost 0 to the hits, a positive cost for misclassification. The false positives are costlier than the false negatives. This ensures the classifier optimizing the cost.

3.2. Summary

The learning ability of a filter is of great advantage, because if every user receives different email, every filter is different (in terms of stored data and model learned), and it becomes very difficult for spammers to design counter-attacks capable of avoiding the filters of all users simultaneously. As spammers benefit depends on the number of spam messages read, they are forced to prepare very sophisticated techniques to breakthrough different filters equipped with different learned models.

The methods of spamming and spam filtering are improving mutually. Spammers attempt to decrease filtering effectiveness by deriving new spamming techniques. The reactivity of spammers asks for countermeasures from filter developers, which in the field of spam filtering may be termed as opposing reactivity. The advent of image spam can be considered as a part of the reactivity, and thus the image-based spam filtering as such can be considered an opposition to reactivity.

4. IMAGE SPAM DETECTION TECHNIQUES AND A COMPARISON

Wu et al. (2005) [9] is the first one who proposed an image classification technique [7]. The proposed technique computed the chosen features on all the images attached to an email. Features related to the presence of text are:

- the number of detected text regions,
- the fraction of images with detected text regions, and

- the relative area occupied by text (usually denoted as text area).

The ratio of the number of banner and of graphic images to the total number of attached images were also used as features based on the assumption that many spam images are banners and computer-generated graphics (which are part of advertisements). Banners detection was done considering the aspect ratio, height, and width. For graphics detection it was assumed that computer-generated graphics contain homogeneous background and less texture. A one-class classifier (a SVM) was proposed in this work, since a representative set of legitimate emails was unavailable.

The main goal in Dredze et al. (2007) [10] was to build a fast classifier [7]. In order to achieve this, the image processing operations which involved more time and effort was avoided. The proposed technique exploits a large set of features which are considered easy to compute: image metadata, and information like image height, width, aspect ratio, format extension (e.g., gif, jpg), and file size [7]. Visual features like average red, green and blue values, features based on edge detection were also considered. Three classifiers namely, Maximum entropy, Naïve Bayes classifiers and decision trees were used for classification. The advantage of decision tree technique was it required accessing only a subset of features for every testing image. The processing time at classification stage was reported to vary from 2.5 to 4.4 ms, based on the classifier being used.

Mehta et al. (2008)[3] argued that spam images are artificially generated, and contain clearer and sharper objects than ham images; thus, their color distribution should be less smooth.[9] Moreover, the researchers propounded that the low-level features helped the recipients to achieve the highest discerning capability. It was also stated that these attributes make it difficult for a spammer to randomize the images making them passable through the filter and get them misclassified into the inbox. The features collected consider the color, shape and texture of an image. A two class SVM classifier with the RBF (non-linear) kernel was used [7].

Wang et al. (2007) [11] in their work used numerous set of features, extracted by existing image spam filters (in this case they are used only as feature extractors, not as classifiers). [7] The similarity measure is calculated individually for each set of features. The distance measure is then compared to a threshold. The threshold is set different for each feature space. Based on the threshold value it is decided if the image is spam or legitimate. The labels from different feature spaces are then combined using logical operators (OR, or AND), or by voting [9]. The features considered for experimental evaluation were extracted from color histograms, Haar wavelet transform, and edge orientation histograms. The processing time for an image was reported to be in milliseconds.

4.1. Comparison

Wu et al. (2005)[9] proposed a one class SVM classifier. Several one-class classifiers were trained using different percentage of outliers for each classifier. Due to privacy issues, a set of legitimate emails was difficult to collect. Moreover, lower values for false-positive are considered more significant than high detection rates. Therefore, the one-class SVM classifier with 20% outlier was chosen as the base classifier [9]. A comparison of a text-based and the proposed visual-based anti-spam filter revealed that a text-based Bayesian filter, trained using Ling-Spam dataset, identified only 47.73% emails as spam. The proposed visual-based anti-spam filter however, detected 81.40% of the spam emails, an increment of 36.87% detection rate compared to the text-based Bayesian filter [9].

Dredze et al (2007)[10] performed evaluation on three different sets of data. Results were measured using both accuracy and the spam F1 score. It was stated that Maximum Entropy achieved an accuracy exceeding 89% on all datasets. For PHam/PSpam, performance was found to improve to 98% in accuracy. Naive Bayes performed with an accuracy of above 76%, worse than Maximum Entropy while the decision tree showed an accuracy of 85% for PHam/SpamArc. Another set of experiments was performed without deleting the duplicate images. This increased the number of images in the Spam Archive dataset by a high value. Accuracy increased to 98% on PHam/PSpam. Interestingly, accuracy on PHam/SpamArc increased to 97% for the Maximum Entropy technique, reaching a level comparable to personal spam.

Mehta et al (2008)[3] observed that visual features can be used to identify spam images; the method achieved a prediction accuracy of over 95% in all categories. Comparing the results achieved by [Dredze et al. (2007), Fumera et al (2006)][10,12] the system was reported as recording an improvement above 6%. Experiments performed on the Personal spam dataset, comparable results were achieved as [Dredze et al. (2007)]. Evaluating the impact of resolution it was found that at 400 × 400, the system achieves an accuracy of more than 99.6%. Higher accuracies were deemed impossible to implement considering human classification fallacy. [3]
The results achieved by the Wang et al. (2007) [11] show that all three filters individually achieved high accuracy in targeted category (categories that the filter is designed to handle). The color histogram filter achieved 100% detection rates for 76% categories, and more than 96.7% for the remaining 4categories. The false positive rates of all categories (except shift) were 0.0006% [11]. The wavelet filter achieved 100% detection rate for all targeted categories and the false positive rate were below 0.0009%. The orientation histogram filter also achieved 100% detection rate for 4 categories, while the false positive rate was below 0.0007%.[11] To understand the multiple spam filter aggregation three aggregation methods have been used: AND, OR and VOTE. Out of the three methods it has been reported the VOTE method as more accurate compromising the false positive for detection rate keeping the false positive rate consistent below 0.0002% for all targeted categories with good detection rates. The researchers further emphasized that multiple filters can work better than an individual filter thus justifying the design goal of making the system commercial.

Table 1. Comparison of results achieved by previous researchers

Papers	**Feature set**	**Data set**	**Classifier**	**Accuracy**
Wu et al. (2005)	Text Area and Low Level Features	Spam Archive and Ling Spam personal emails	One-class SVM	84.6
Dredze et al. (2007)	Image Metadata and Low Level Features	Personal Ham and Spam collection, Spam Archive	Maximum Entropy, Decision Tree, Naïve Bayes	93 - 96
Wang et al. (2007)	Low level features and similarity of images	Personally collected dataset	Nearest neighbour detection using Manhattan distance	88-96

Mehta et al. (2008)	Low level features and similarity of images	Spam Archive, Princeton, dataset by Dredze et al, personal ham collection	SVM and Nearest neighbour detection using Jensen Shannon distance	95-98
Biggio et al. (2008)	Low level features, text area and OCR technique	Dataset designed by Dredze et al and personal collection	SVM	94-98

5. PROPOSED IDEA

After an in-depth study of the research works related to the field of email and image spam an understanding was derived about the various features that characterizes an image spam and also of the various techniques used for exploiting such features to defeat image spam. An image spam contains text embedded in the image. Moreover spam image tend to have a rough distribution in the RGB/LAB color space [3]. Spammers ensure that each image spam is noisy and distinct. This is done by using various obscuring techniques like reposition various items, add noise arbitrarily, alter background or font colors and sizes, add indiscriminate patterns like lines or circles, borders, etc.

There have been various techniques proposed by researchers to extract such features which distinguish a spam image from a legitimate image. It has been found that simple text-categorization techniques are not sufficient to prevent image spam. Hence, it is required to assemble image features like visual features and image meta-data in order to derive an effective spam detection system.

Presently, most commercial and open-source server-side spam filters consist of different modules each aimed at detecting specific characteristics of spam e-mails[12]. The different modules can be arranged in parallel or in a hierarchical structure. For parallel structure the decision depends on combining the outputs of each module, usually given as continuous-valued scores. Modules when organized hierarchically use the simpler ones first. The more complex ones are used only if a reliable decision cannot be taken on the basis of previous ones.

A comparative study of previous research work showed a higher accuracy for the SVM classifier together with a distance measure used to determine the similarity of features with a query image or feature set [7, 3]. Such a technique has also been termed as near duplicate detection method, which can be termed as an ultimate guise for similarity search. The spam detection system proposed in this work exploits the fact that spam images are artificially generated and spammers use various randomization techniques to generate numerous distinct images from a template image [3, 10]. Moreover, these artificially generated images are sent in batches. Thus images in a batch are similar and every message sent to different users are not the same.

Analyzing the Image Spam Dataset designed by Dredze et al. (2007)[10] it has been found that spam images contain pictures apart from embedded text. Examples have been shown below in figure 2. Based on this the content analysis has been divided into two modules. Since the general definition of image spam is text embedded in images the main aim is to extract the embedded text form the image body. The similarity measure will be calculated between the text extracted from the image and the text in the body of a spam message. Whenever, the spam image contains much more than embedded text and passes undetected by the text extraction phase the low level feature extraction comes into play. The low level features considered are color features, texture features and shape features. These features after being extracted will be used to train a SVM classifier.

Stacked generalization, or stacking, is an approach for constructing classifier ensembles [8]. A classifier assemblage can be defined as a collection of classifiers. The individual decisions of each classifier are merged to classify new instances[8]. Multiple classifiers are assembled to generate a classifier with higher accuracy and improved performance. In the proposed system the outcome of the similarity measure and the SVM classifier are stacked to obtain a classifier with increased accuracy

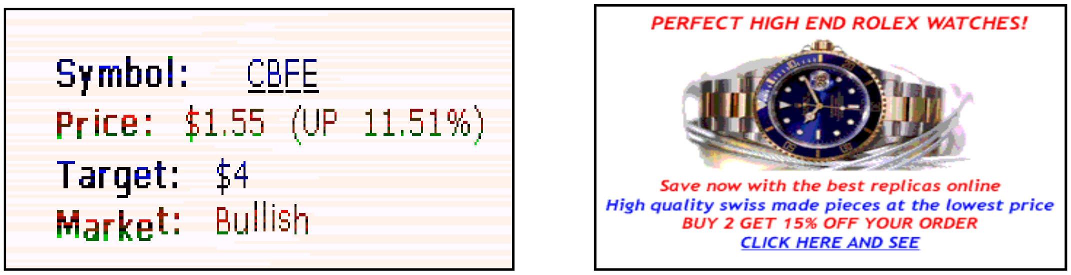
Symbol: CBFE
Price: $1.55 (UP 11.51%)
Target: $4
Market: Bullish
PERFECT HIGH END ROLEX WATCHES!
Save now with the best replicas online
High quality swiss made pieces at the lowest price
BUY 2 GET 15% OFF YOUR ORDER
CLICK HERE AND SEE

Figure 2: Examples of spam images a) image with embedded text b) image with text and picture

5.1. Idea of Implementation

The idea of implementation is to design a framework with a virtual environment created that would emulate email server/client system. When a message is received it is sent to the spam filter. It is checked if the email contains any embedded image. The filter initially segments the image and extracts the text embedded in the image. Wu et al. (1997)[13] proposed an algorithm to find text embedded in an image. This work uses an enhancement of the algorithm. The embedded text is fed into an OCR system where the extracted text is converted to text strings. The text strings are compared to the dataset of spam text. A distance measure is calculated to find the similarity between the extracted text and the spam text. The next phase of the filter is to extract the low level features like the color, texture and shape features. These features are fed to a classifier. After obtaining a classification for the text features and the low level features the two classifiers are stacked to obtain a ensemble-classifier with improved performance.

5.2. Block Diagram

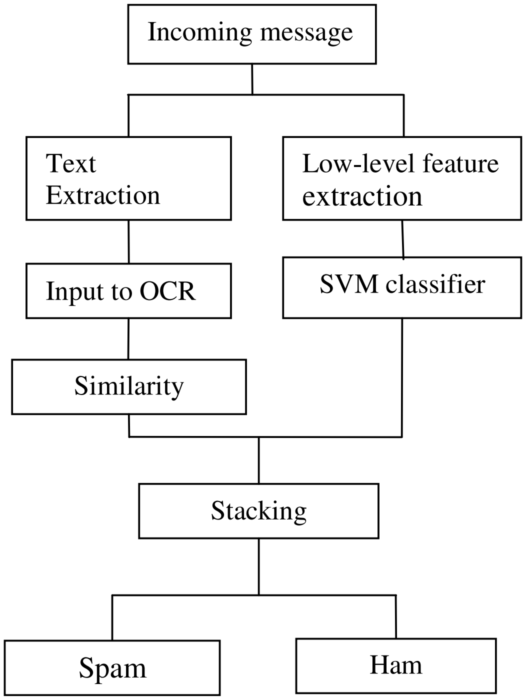

Figure 3: Block diagram of spam detection system

6. CONCLUSION

With the increasing importance of email and the incursions of Internet marketers, unsolicited commercial email (also known as spam) has become a major problem on the Internet. To detect image spam, computer vision and pattern recognition techniques are also required, and indeed several techniques have been recently proposed. The proposed framework exploits both embedded text extraction and further processing of low level features. The dataset designed by Dredze et al. (2005)[10] will be used for initial analysis. This work promises to enhance the spam filtering domain in future.

REFERENCES

[1] Godwin Caruana, Maozhen Li, “A Survey of Emerging Approaches to Spam Filtering” , ACM Computing Surveys, Vol. 44, No. 2, February 2012.

[2] Radicati S, Khmartseva M, “Email Statistics Report 2009-2013”, The Radicati Group, Inc, Palo Alto, CA, 2009

[3] Mehta, B., Nangia, S., Gupta, M., Nejdl, W.,. Detecting image spam using visual features and near duplicate detection. Proc. 17th Int. Conf. World Wide Web. ACM, pp. 497–506, 2008

[4] Blanzieri Enrico, Bryl Anton, “A Survey Of Learning-Based Techniques Of Email Spam Filtering”, Artificial Intelligence Review, Volume 29, Issue 1, Springer, pp 63-92, 2008.

[5] Gordon V. Cormack, “Email Spam Filtering: A Systematic Review”, Foundations and TrendsR in Information Retrieval Vol. 1, No. 4 335–455, 2008.

[6] Sanz EP, Hidalgo Gomez JM, Pérez Cortizo JC., “Email Spam Filtering”, Advances in Computers Volume 74, Elsevier, Pages 45–114, 2008.

[7] Battista Biggio, Giorgio Fumera, Ignazio Pillai, Fabio Roli, “A survey and experimental evaluation of image spam filtering techniques”, Pattern Recognition Letters 32, 1436–1446 ScienceDirect Pattern Recognition, 2011

[8] Sakkis Georgios, Androutsopoulos Ion, Paliouras Georgios, Karkaletsis Vangelis, Spyropoulos Constantine D., Stamatopoulos Panagiotis, "Stacking classifiers for anti-spam filtering of e-mail", In Proc. 6th Conference on Empirical Methods in Natural Language Processing, Pittsburgh, US, 2001.

[9] Wu, C.-T., Cheng, K.-T., Zhu, Q., Wu, Y.-L., "Using visual features for anti-spamfiltering", In: Proc. IEEE Int. Conf. Image Process, Vol. III. pp. 501–504., 2005.

[10] Dredze, M., Gevaryahu, R., Elias-Bachrach, A., "Learning fast classifiers for image spam." Proc. Fourth Conf. Email Anti-Spam (CEAS), 2007.

[11] Wang, Z., Josephson, W., Lv, Q., Charikar, M., Li, K., "Filtering image spam with near-duplicate detection.", In: Proc. Fourth Conf. Email Anti Spam (CEAS), 2007.

[12] Fumera Giorgio, Pillai Ignazio, Roli Fabio, "Spam Filtering Based On The Analysis Of Text Information Embedded Into Images" Journal of Machine Learning Research 7 2699-2720, 2006

[13] Wu Victor, Manmatha R, Riseman Edward M, "TextFinder: An Automatic System To Detect And Recognize Text In Images, Pattern Analysis and Machine Intelligence, IEEE Transactions Volume:21 , Issue: 11, November, 1999

Authors

1. Meghali Das
M.TECH CSE (AI)
Don Bosco College of Engineering and Technology,
Guwahati, Assam

2. Vijay Prasad
Assistant Professor,
Don Bosco College of Engineering and Technology,
Guwahati, Assam.
MTECH CSE (AI)