Clustering-based Spam Image Filtering Considering Fuzziness of the Spam Image

Master Prince
Department of Computer Science
Qasim University, College of Computer
Al Qasim, Kingdom of Saudi Arabia

***Abstract*—If there are pros, corns are always there. As email becomes a part of individual's need in our busy life with its benefits, it has negative aspect too by means of email spamming. Nowadays images with embedded text called image spamming have been used by the spammers as effective text spam filtering methods already been introduced. Tracking and stopping spam become challenge in the internet world because of versatility in the spam images. In this paper a novel model AFSIF (Autonomous Fuzzy Spam Image Filter) has been introduced. The basic idea behind AFSIF is, an spam image can combine several basic features of different spam images, so feature fusion weight of the image has been generated, which keeps combined feature of spam images and user preference as well. Here user preference has not been applied separately; it is used to calculate the fusion weight in terms of predefined topics (rule table).**

Keywords—versatility of spam image; feature fusion weight; cluster; rule table

I. INTRODUCTION

As internet comes under the reach of majority of the people the email becomes the cheapest and effective way of advertisement. Spammers are using this medium by sending unwanted email message through junk email, earlier text-based spam emails have been used but now to by-pass the conventional email filtering technique they are using image-based spam. Image spam is actually a technique of embedding text (commercial content) into image by means of penetrating the text spam filter. Most email readers spend a non-trivial amount of time regularly deleting junk email messages, even as an expanding volume of such email occupies sever storage space and consume network bandwidth [1].

To protect the inbox from image spam emails, the filter should be able to distinguish between spam and ham images. The use of computer vision and pattern recognition techniques has been investigated in recent years and several text-based spam image filtering methods have been developed. Consequently some researchers proposed techniques based on detecting the presence of embedded text, and on characterizing text areas with low level feature like their size [2, 3] or their color distribution [2, 4].

However, some realistic problems that are not dealt well by the prevalent models. Like a spam image may belongs to several categories of spam images and similarity measurement is not able to discriminate because of small difference from each of the class. And user preference we usually place at the end, an intelligent spammer can send every time new image to defeat the spam filter as the result end user spam it after seeing it.

In this paper a novel approach called AFSIF has been presented. The AFSIF comprises two steps filtration tasks. First stage is cluster based filter. The training data set is divided into clusters based on their similar features. The image is mapped with each of the cluster and declares as spam or ham depending on degree of similarity and dissimilarity respectively. If image is labeled as ham second stage comes into action, the feature fusion weight of the image has been calculated, which will be describe in subsequent sections. Finally, if image has been declared as spam by the user from inbox at the same time according to the similarity measure the spam image is associated with the closest cluster training data set. In this method user will experience negligible number of spam email because in second stage the fusion weight of the image contains user predefined topics.

The proposed AFSIF model has two advantages: 1) Incremental Learning System, 2) Features of spam image is fused with user pre-defined topics through feature fusion weight.

The rest of the paper is structured as follows. Section II gives the brief overview of literature reviewed, Section III presents the proposed AFSIF model, Section IV presents experiment and results, and finally in Section V some conclusions are drawn.

II. RELATED WORK

The wide use of image spam fetches the attention of researchers. Several attempts have been made to address filtering spam images by utilizing specific feature of image [3, 5]. For feature extraction there are various algorithms, such as principle component analysis (PCA), Independent component analysis (ICA), Partial Least squares (PLS) to transform graphical image into feature vector. In this paper, PCA has been used because of its suitability for data set in multiple dimensions.

Yih et al.[6] first address the grey mail problem and train two spam filters - the gray and b&w (ham/spam)filter on two disjoint subsets. Ming-wei Chang et al. [7] propose using the portioned logistic regression (PLR) to learn content and user model separately.

Here, Dataset accompanies combined set of gray and b&w and spam archive data provided by Giorgio Fumera's group has been considered.

III. AFSIF MODEL IMPLIMENTATION

Fig. 1 as shown is the proposed AFSIF model with two level filtering.

A. Feature Extraction & Clustering

First and very important task is to extract the feature of all the training set images; it plays a very vital role to improve the classification system. The spam archive images were taken from the spam archive data provided by Giorgio Fumera's group. In total, the images consider to this proposed work is 1204 JPG images with 964 spam images and 240 ham images and some personal data set.

Principle component analysis (PCA) is applied to convert the high dimension images into reduced set of feature vector without much loss of information. After feature extraction an image xi will be represented as a feature vector with d dimension, where d is the number of image xi features. Once all the images are represented with feature vector they clustered into group based on similarity measurement between the feature vectors [8] and user predefined topics. The weight of each image within each cluster is calculated by multiplying the difference form mean image of that cluster and feature vector of the image. The representative weight of each cluster is that whose average dissimilarity to all the images in the cluster is minimal.

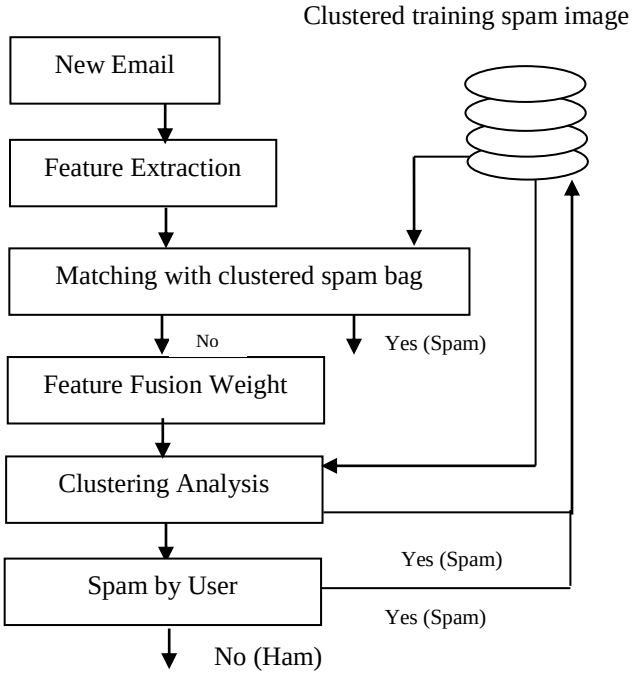

Fig. 1. AFSIF Model for spam image filtering

B. Similarity Measurement

For discrimination between spam and ham image the degree of similarity is measured. Let p = {p1, p2… p25} and q = {q1, q2… q25} are the feature vectors of two images p and q. When calculating the degree of similarity of p and q, we use Euclidean distance defined by

The smaller the distance value is, the more similar the two images. In the matching process, the similarity evaluation leads to the mapping R25 → {-1, +1}, where R25 represents the normalized 25 edge features of a new coming image, -1 and +1 denote ham, spam image respectively. So, Euclidean distance between new coming image and representative image of each cluster is compared. The new image is categorized as spam if smallest distance value is not more than a dissimilarity threshold otherwise treated as normal image.

If the image is labeled as ham the next level of filter comes into action.

C. Feature Fusion Weight Generation

The objective of FFW (Feature Fusion Weight) generation of the image is to obtain the feature vector of the given image whose feature slightly matches with more than one cluster. And smallest distance value from representative weight may be more than a dissimilarity threshold as the result spam image may bypass the filter. At the same time the user preference is also considered by means of rule table and fused with the feature vector to obtained FFW.

Fuzzy kohonen clustering network (FKCN) is employed to determine the fusion weight of the coming image corresponding to each cluster.

In Fig. 2 the input layer of the network, the feature vector of the coming image is given, the distance between the input image and cluster's representative image is calculated such that

$$dij = \| Xi - Pj \|^2 \quad (1)$$

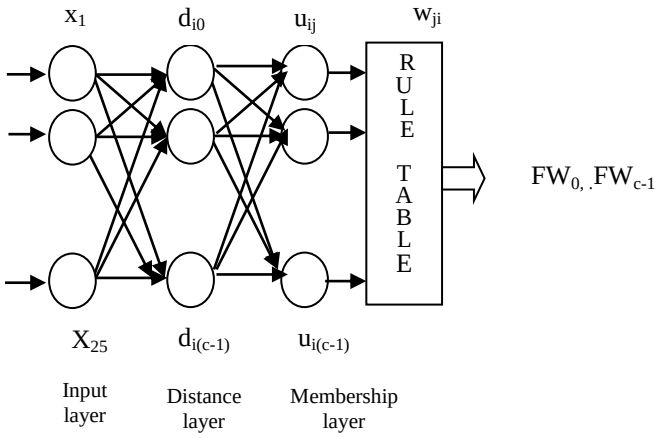

Fig. 2. Feature Fusion Weight Generation

where X_i denotes the input image and P_j denotes the jth representative image. In this layer degree of dissimilarity between coming image and the representative images is calculated. The membership layer calculates the similarity degree between coming image and representative images. If smallest distance value of the coming image from each of the representative is more than a dissimilarity threshold, then the similarity between the coming image and individual

representative image is represented by membership value from 0 to 1.

$$FW_i = \sum_{j=0}^{c-1} w_{ji}\, u_{ij} \quad (2)$$

where w_{ji} represents the representative image weight of the ith cluster. The representative image weights are designed in a rule table according to the user preference.

Table 1 describes the weight of the ith cluster representative image according to user preference. The size of the table may increase as much as clusters and user preferences increases

TABLE I. USER PREFERENCE RULE TABLE

If Similarity degree from RI							SPAM	HAM
C_1	C_2	C_3				C_n		
1	>.65	<0.34					1	
		•	•		•	•		
		•	•		•	•		
	1							1
		>0.68					1	

D. Clustering Analysis

After calculating the FFW of the incoming image feature fusion weight is calculated such as,

$$FFW_i = \sum_{j=0}^{c-1} RI_i\, FW_j \quad (3)$$

Then compare with the clustered training image set by similarity evaluation as described in section B and label accordingly. If labeled as spam, the image is added to the closest cluster and trained, so the model justify the incremental learning system.

E. Filter by User

At the end there are possibility of leaking of filter, so when marked as spam by the user from inbox, the image added to the training set and rule table may also be updated with the new user preference.

IV. EXPERIMENTAL RESULTS

A. Training dataset

In order to evaluate the performance, the experiments are carried out by using the Personal dataset in [5] for two reasons. First, the Personal dataset is one of few public corpuses containing both spam images and normal images appeared in real email exchange. Furthermore, the spam images in this dataset can reflect the property of similarity among spam images. And 1204 JPG images with 964 spam images and 240 ham images (Giorgio Fumera's group). After preprocessing through similarity measure and user preference these images are divided in to 25 clusters.

B. Performance measure

Two measurements are applied to evaluate the performance of the proposed method; True Positive Rate (TPR) and False Positive Rate (FPR) which are computed as follows:

$$\text{True Positive Rate} = \frac{\text{Correctly detected}}{\text{Total number of images}}$$

$$\text{False Positive Rate} = \frac{\text{Incorrectly detected}}{\text{Total number of images}}$$

C. Results

Table II describes the results on personal and full dataset.

TABLE II. RESULTS ON PERSONAL & FULL DATASET

	Classified	TPR	FPR
Personal (350)	341	0.9742	0.02
Full (1554)	1456	0.9399	0.063

On personal dataset the results are more efficient than the general dataset.

V. CONCULUSION AND FUTURE WORK

As user preferences are added with the weight of the incoming image through rule table very less number of spam images bounces to the inbox. So, eliminating the need of extra filter for gray/spam for user preference.

For future work, further enhancement is needed in the rule table. Machine learning can be employed to prepare standard rule table and user can set their preference. Big Five Model of Personality can also be used to automate the system.

REFERENCES

[1] Ian stuart, sung- Hyuk cha and Charles tappert, "A Neural Network Classifiers for Junk Mail", springerLink2004, pp-442-450.

[2] H. B Aradhye, G. K. Myers, and J. A Herson. Image Analysis for Efficient Categorization of image-based Spam E-mail, Eighth International Conference on Document Analysis and Recognition (ICDAR'05), August 2005, pp. 914-918.

[3] B. Biggio, G. Fumera, I. Pillai, and F. Roli. Image Spam Filtering Using Visual Information, 14th international Conference on Image Analysis and Processing (ICIAP 2007), September 2007, pp. 105-110.

[4] Y. Gao, M. Yang, X. Zhao, B. Pardo, Y. Wu, T. N. Pappas, A. Choudhary. Image Spam Hunter, IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 1765-1768, 2008.

[5] M. dredze, R. Gevaryahu, and A.E Bechrach, " Learning Fast Classifiers for Image Spam", in proc. CEAS '2007-2007.

[6] W. Yih, R. McCann, and A. Kolcz. "Improving Spam Filtering by Detecting Gray Mail", in proc CEAS'2007, 2007.

[7] Ming-wei Chang , wen Tau Yih, and Robert McCnn. "Personalized Spam Filtering for Gray Mail", in proc.'2008, 2008.

[8] Ying He, Wengang Man, and Haibo He "Incremental Clustering-based Spam Image Filtering using Representative Images", proc ICSEDMI 2011, 2011.