Improved Spam Filtering by Extraction of Information from Text Embedded Image E-mail

Seongwook Youn
Semantic Information Research Laboratory
(http://sir-lab.usc.edu)
Dept. of Computer Science
Univ. of Southern California
Los Angeles, CA 90089, USA
syoun@usc.edu

Dennis McLeod
Semantic Information Research Laboratory
(http://sir-lab.usc.edu)
Dept. of Computer Science
Univ. of Southern California
Los Angeles, CA 90089, USA
mcleod@usc.edu

ABSTRACT

The increase of image spam, a kind of spam in which the text message is embedded into an attached image to defeat spam filtering techniques, is becoming an increasingly major problem.. For nearly a decade, content based filtering using text classification or machine learning has been a major trend of anti-spam filtering systems. A Key technique being used by spammers is to embed text into image(s) in spam email. In [4], we proposed two levels of ontology spam filters: a first level global ontology filter and a second level user-customized ontology filter. However, that previous system handles only text e-mail and the percentage of attached images is increasing sharply. The contribution of the paper is that we add an image e-mail handling capability to the previous anti-spam filtering system, enhancing the effectiveness of spam filtering.

Categories and Subject Descriptors

H.4.3 [**Information Systems Applications**]: Communications Applications – *electronic mail;* H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *information Filtering*

General Terms

Algorithms, Management, Performance, Experimentation

Keywords

e-mail classification, feature selection, OCR, ontology, spam filtering

1. INTRODUCTION

Anti-spam is a very active area of research, and various forms of filters, such as white-lists, black-lists, and content-based lists are widely used to defend against spam. Many content-based filters utilize machine learning algorithms for filtering spam. The first countermeasures taken by spammers consisted of adding bogus text to their e-mails, usually taken from books or news articles, to compromise the effectiveness of statistical techniques. However, a new kind of trick introduced some years ago has rapidly spread during the past year and is now adopted in a large fraction of spam e-mails: it consists in embedding the spam message into attached images to circumvent all spam detection techniques based on the analysis of body text [3]. This kind of spam is known as image spam. A number of spammers have been evading filters recently by encoding their messages as images and including some irrelevant good words. This implies the contents are hard to retrieve from the binary image encoding. This type of image spam accounts for 40% of all global spam in 2007, compared with just 1% in late 2005 [1, 2].

2. RETRIEVAL OF TEXT FROM TEXT EMBEDDED IMAGES USING OCR

OCR (Optical Character Recognition) translates images of text, such as scanned documents, into actual text characters. OCR makes it possible to edit and reuse the text that is normally locked inside scanned images. By running a sample of 200 image e-mails, we determined that Asprise OCR was performing with an accuracy of 95%. It had the best detection rate among the approaches we analyzed; hence we decided to go with Asprise OCR for our research.

Image e-mail among the training data set is entered into OCR, and then text information is retrieved from text embedded image e-mail. Training data set is selected. Training data set is a collection of text-oriented email data. Features from the data set are selected using *tfidf*. Weka input file is created based on the selected features and the data set. Weka is a toolkit of machine learning algorithms written in Java for data mining tasks. Through Weka, classification results are generated. The classified results are converted to RDF file. The converted RDF file is fed into Jena, which is a Java framework for building Semantic Web applications. It provides a programmatic environment for RDF, RDFS, OWL, and SPARQL, and includes a rule-based inference engine. Using Jena, ontologies are created, and we can give a query to Jena. Jena will give an output for the query using ontologies created in Jena. Through these procedures, global and user-customized ontology filters are created. Incorrectly classified emails through global ontology filter are inserted into the user-customized ontology filter. The ontologies created by the

SAC'09, March 8-12, 2009, Honolulu, Hawaii, U.S.A.

implementation are modular, so those could be used in another system.

3. SPAM FILTERING

Figure 1 shows our framework to filter spam. The training data set is the set of e-mail that gives us a classification result. It is composed of both text e-mail and image e-mail. The test data is actually the e-mail will run through our system which we test to see if classified correctly as spam or not.

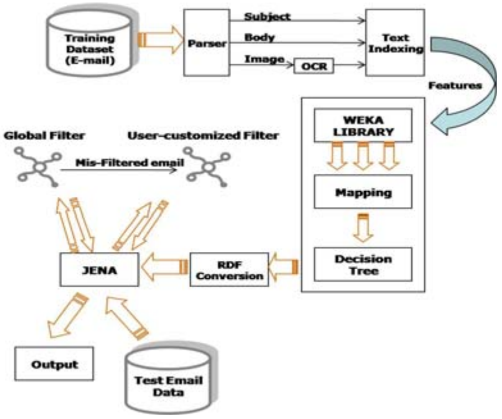

Figure 1. Spam Filtering System (SPONGY system)

4. EXPERIMENTAL RESULTS

In the experiment, we used both text e-mail and image e-mail. Data set was classified like the followings:

TS (Text Spam) -1008, TL (Text Legitimate) -1100

OCR IS (Retrieved text from Image Spam using OCR) – 131

OCR IL (Retrieved text from Image legitimate using OCR) – 177

As shown on Table 1, 2, 3, and 4, in the SPONGY (Spam ONtoloGY) system with OCR functionality, false negative rate is increased from 6.34% to 7.36%, false positive rate is increased from 2.28% to 3.16%, and accuracy (Correct classification rate) is decreased a little.

Table 1. Experimental Results of Global filter w/o OCR

	Global Filter (C4.5)		
	False Negative	False Positive	Accuracy
TS + TL	12.91%	3.67%	91.5085%

Table 2. Experimental Results of SPONGY w/o OCR

	SPONGY		
	False Negative	False Positive	Accuracy
TS + TL	6.34%	2.28%	95.4459%

5. COMPARISON WITH COMMERCIAL FILTERS

We compared Gmail, Yahoo! mail, the USC e-mail and SPONGY. In the experiment, e-mail addresses and messages of our research group members are used. We cannot specify sender's e-mail address because most of e-mail systems support authentication system. The experiment is performed with own filters of each e-mail system (Gmail, Yahoo! mail, the USC e-mail, and SPONGY) with the default setting. The experiment was performed on the real setting with different e-mail data set. In this case, the SPONGY system showed good performance in both false negative rate and false positive rate. In the SPONGY system, most balanced false negative and false positive rate values were obtained. False negative rate was 7.3610% and false positive rate was 3.1607%.

Table 3. Experimental Results of Global filter w/ OCR

	Global Filter (C4.5)		
	False Negative	**False Positive**	**Accuracy**
TS+TL+OCR IS+OCR IL	13.55%	4.74%	90.6043%

Table 4. Experimental Results of SPONGY w/ OCR

	SPONGY		
	False Negative	**False Positive**	**Accuracy**
TS+TL+OCR IS+OCR IL	7.36%	3.16%	94.6192%

Three other commercial mail systems showed low experimental results. False negative rate of the Yahoo! mail was extremely bad for us. Brightmail AntiSpam of the Symantec used in the USC e-mail system showed very low false positive rate. As a result, performance order is SPONGY, the USC e-mail system, Google Gmail, and Yahoo! mail. Experimental results are shown in Table 5.

Table 5. Comparison Result with Different E-mail Data Set

	Google Gmail	**Yahoo mail**	**USC email**	**SPONGY**
False Negative	11.2436%	19.3319%	9.6611%	7.3610%
False Positive	4.6358%	6.3830%	2.8169%	3.1607%

With some of test e-mail data set, SPONGY showed better performance at least under our experiment. By increasing image e-mail handling capability, we possibly increase the performance of the spam filtering system.

6. CONCLUSION AND DISCUSSION

The proposed framework is a hybrid two-level filter, which combines global filter and user-customized filter. Also, it is user-customized, scalable, and modularized, so that it can be embedded to many other systems for better performance. We added image spam handling capability using OCR into the text-based anti-spam filtering system. By handling of text embedded image e-mail, the proposed system can be used partially for both text e-mail and image e-mail. The experiment was somewhat restricted, but it demonstrates the potential capability of the proposed system under restricted environment. However, to cope with the image e-mail thoroughly, we need to adopt advanced image processing techniques. Then, we can face image obscuring techniques like wave, animate, deform, and rotate. In the future, we will experiment with the combination of the general corpus data set and our data set for generality.

7. REFERENCES

[1] Biggio, B., Fumera, G., Pillai, I., Roli, F. Image Spam Filtering Using Visual Information. In *Proceedings of ICIAP*, 2007, 105-110.

[2] Fumera, G., Pillai, I., and Roli, F. Spam Filtering Based On The Analysis Of Text Information Embedded Into Images. *Journal of Machine Learning Research 6,* (2006), 2699-2720.

[3] Youn, S. and McLeod, D. Spam E-mail Classification using an Adaptive Ontology, *Journal of Software (JSW), 2, 3 (2007),* 43-55