Support Vector Machines and Random Forests Modeling for Spam Senders Behavior Analysis

Yuchun Tang, Sven Krasser, Yuanchen He, Weilai Yang, Dmitri Alperovitch
Applied Research, Secure Computing Corporation
Email: {ytang, skrasser, yhe, wyang, dalperovitch}@securecomputing.com

http://www.trustedsource.org

***Abstract*— Unwanted and malicious messages dominate Email traffic and pose a great threat to the utility of email communications. Reputation systems have been getting momentum as the solution. Such systems extract Email senders behavior data based on global sending distribution, analyze them and assign a value of trust to each IP address sending email messages. We build two models for the classification purpose. One is based on Support Vector Machines(SVM) and the other is Random Forests(RF). Experimental results show that either classifier is effective. RF is slightly more accurate, but more expensive in terms of both time and space. SVM produces similar accuracy in a much faster manner if given modeling parameters. These classifiers can contribute to a reputation system as one source of analysis and increase its accuracy.**

I. Introduction

When the anti-spam war started in mid 90s, it mainly relied on detecting common words present in spam and manually blacklisting spam IP addresses. Historically, there has been much controversy around blacklists because of the subjectivity of getting placed on the list and the difficulty of being removed. Whitelists have been developed that maintain lists of legitimate senders. A problem with whitelists and blacklists is that they left a sizable set of senders in the middle of the spectrum that were not classified. This is because there was not enough credible information or feedback to make a binary decision about the sender. As spammers started using fast changing botnets, randomizing and obfuscating content in their messages, those technologies quickly became ineffective. Recently we begin to see more adversarial intelligence and data mining from spammers. Spammers continue to accumulate vast data stores on vulnerable systems and individuals. To address this problem, reputation systems for email became a topic of interest in anti-spam research around 2003.

In general, a reputation system collects and aggregates feedback from participants to provide incentives for future cooperation. Email reputation systems aim to assign a reputation value to every host that currently sends email or may send email in the future. Typically, an email reputation system uses the IP address of a computer as the identity and a reputation is built for that identity.

An example for such a reputation system is Secure Computing's TrustedSource [1]. TrustedSource operates on a wide range of data sources including proprietary and public data. The latter includes public information obtained from DNS records, WHOIS data, or Real-time Blackhole Lists (RBLs). The proprietary data is gathered from over 5000 sensors placed at email gateways global wide. It gives a unique view into global enterprise email patterns. A sensor located in the receiver mail gateway sends queries to TrustedSource about the received messages. Based on the needs of email administrator, the sensor is able to share different levels of data and query for reputation from different aspects. Currently, TrustedSource can calculate a reputation value for the IP address a message originated from, for URLs in the message, and for message fingerprints.

This paper is focused on the analysis of the performance of two commonly used data mining techniques: Support Vector Machines (SVM) [2] and Random Forests (RF) [3]. Inputs to both models are IP sending behavior collected from sensors in real-time. They return IPs reputation scores as outputs. The effectiveness and the efficiency of classification modeling are then empirically analyzed. As far as we know, this is the first time a comparison study is conducted between these two state-of-the-art classification methods on a real-world information security problem.

The rest of the paper is organized as follows. Section II provides some background regarding RBLs and reputation-based filtering approaches in the email messaging space. Behavioral data of email senders is described in Section III. Brief reviews of supervised binary classification are in Section IV. Section V empirically compares classification performance in terms of effectiveness and efficiency. A Return of Investment (ROI) analysis is given in Section VI to measure the benefits of this research. Finally, Section VII concludes the paper.

II. Background

An early approach for efficient filtering of unwanted spam email traffic is the use of RBLs. An RBL is a list of IPs (or netblocks) that are not allowed to deliver email messages to the mail server incorporating them.

Typically, RBLs can be queried over the DNS protocol. When a host on the Internet tries to send a message to an email server, the email server can query an RBL to check whether the IP of the sending host is listed. If the IP is listed then the sending host is a known source of malicious traffic. Alternatively, this information combined with a local analysis result can be used to render a final decision.

978-1-4244-2324-8/08/$25.00 © 2008 IEEE.
This full text paper was peer reviewed at the direction of IEEE Communications Society subject matter experts for publication in the IEEE "GLOBECOM" 2008 proceedings.

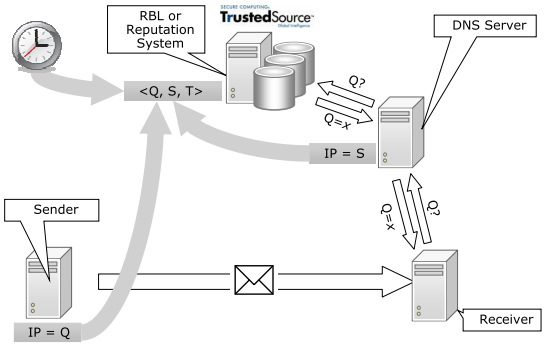


Fig. 1. Data gathered from query.

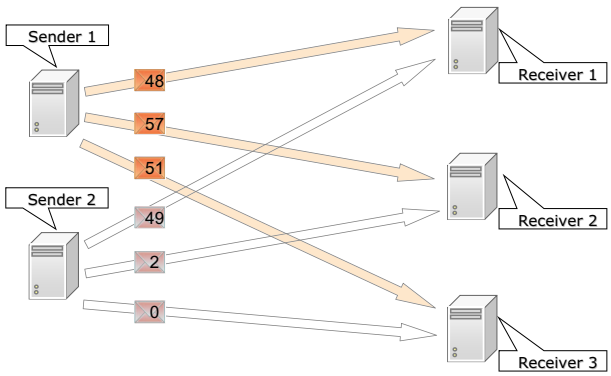


Fig. 2. Message sending distribution across receivers.

RBLs generally gather malicious senders information from spam messages delivered to spamtrap email addresses, manual listings, or user feedback. There are several drawbacks to this approach. To be automatically listed, a particular spam run needs to target a spamtrap email address. All messages that have been delivered beforehand cannot be stopped. Manual listings and user feedback require a human in the loop and are inherently non-automatic. This results in a slow reaction time. It is especially problematic to catch zombie machines with only a few hours of sending activity.

One approach to counter these shortcomings is to automate the process with more sources of feedback as inputs. For example, the real-time queries sent over DNS can yield additional insight. Fig. 1 shows how data can be collected in this manner. A sending host with IP address Q tries to send a message to a receiving email server. The email server queries an RBL for the IP address of the sending host (Q). We therefore call it the *queried IP*. The query is relayed over DNS to an RBL server. The server sees the query packet coming from the IP address S of the DNS server. S is called the *source IP*. In addition, time T when the query was received is stored. This results in a tuple $< Q, S, T >$ for each query.

III. Sensor Distribution Feature Extraction

In previous work, we verified that classification on features extracted from $< Q, S, T >$ records is effective for spam senders detection [4]. In this paper, we investigate whether global message sending distribution is informative for spam senders detection. Ramachandran *et al.* describe a similar feature extraction idea from email senders behavior [5]. Their work focuses on unsupervised clustering of IP addresses based on the set of target domains to which an IP attempts to send messages. Their data are based on traces retrieved from a mail hosting company hosting about 115 domains. In contrast, our work is based on data retrieved from large-scale sensor distribution data. They are collected by the TrustedSource reputation system globally. This includes over 5000 TrustedSource-enabled sensors worldwide. Also, our focus is to use supervised machine learning techniques on this data. One advantage of supervised classification is that we can precisely forecast the accuracy of a model before it is released into a production system by Cross Validation (CV) or Out-of-Bag (OB) estimation. We can even control the aggressiveness of spam senders detection by conducting a Receiver Operating Characteristic (ROC) analysis.

Fig. 2 explains the idea of sensor (or receiver) distribution. There are two senders and three receivers. Sender 1 and Sender 2 have different sending distribution patterns. Sender 1 sends approximately the same amount of messages to each of the receivers while Sender 2 mainly sends messages to Receiver 1. Assume Receiver 1 is in the US, Receiver 2 in Europe, and Receiver 3 in Asia, and they all represent different businesses. Intuitively Sender 2 appears more legitimate if it is mainly doing business in the US while Sender 1 is more likely to be a spammer because legitmate businesses tend to have a stronger geographical bias in the traffic distribution while spammers hit email addresses more randomly. In addition, the set of targeted recipients can be assumed to be more stable among multiple IPs operated by the same spammer [5]. This makes it attractive to investigate this distribution as an input feature to determine the legitimacy of a sender.

IV. Supervised Binary Classification

In this section, we provide background knowledge on the classification techniques used, how they are deployed, and how the results can be analyzed and compared.

A. Preliminary

Given a training dataset Tr with n samples $(\mathbf{x_1}, y_1), (\mathbf{x_2}, y_2), ..., (\mathbf{x_n}, y_n)$, where $\mathbf{x_i}$ is a feature vector in a d-dimensional feature space R^d and $y_i \in \{-1, +1\}$ is the corresponding class label, $1 \leq i \leq n$, the task is to find a classifier with a decision function $f(\mathbf{x}, \theta)$ such that $y = f(\mathbf{x}, \theta)$, where y is the class label for $\mathbf{x}$, θ is a vector of unknown parameters in the function.

B. Support Vector Machines

SVM is a learning system based on recent advances in statistical learning theory [2]. Geometrically, the SVM modeling algorithm finds an optimal hyperplane with the maximal margin to separate two classes, which requires to solve the

978-1-4244-2324-8/08/$25.00 © 2008 IEEE.
This full text paper was peer reviewed at the direction of IEEE Communications Society subject matter experts for publication in the IEEE "GLOBECOM" 2008 proceedings.

following optimization problem.

$$\begin{aligned} &\text{maximize} \\ &\quad \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j K(\mathbf{x_i}, \mathbf{x_j}) \\ &\text{subject to} \\ &\quad \sum_{i=1}^{n} \alpha_i y_i = 0 \\ &\quad 0 \leq \alpha_i \leq C, i = 1, 2, ...n \end{aligned} \tag{1}$$

where α_i is the weight assigned to the training sample $\mathbf{x_i}$. If $\alpha_i > 0$, $\mathbf{x_i}$ is called a support vector. C is a "regulation parameter" used to trade-off the training accuracy and the model complexity so that a superior generalization capability can be achieved. K is a kernel function, which is used to measure the similarity between two samples. A popular RBF kernel function, as shown in (2), is used in this research.

$$K(\mathbf{x_i}, \mathbf{x_j}) = \exp(-\gamma \|\mathbf{x_i} - \mathbf{x_j}\|^2), \gamma > 0 \tag{2}$$

After the weights are determined, a test sample $\mathbf{x}$ is classified by

$$y = \text{sign}\left(\sum_{i=1}^{n} \alpha_i y_i K(\mathbf{x_i}, \mathbf{x}) \right), \tag{3}$$

$$\text{sign}(a) = \begin{cases} +1, & \text{if } a > 0 \\ -1, & \text{otherwise} \end{cases}$$

To determine the values of $< \gamma, C >$, a Cross Validation (CV) process is usually conducted on the training dataset. CV is also used to estimate the generalization capability on new samples that are not in the training dataset. A k-fold CV randomly splits the training dataset into k approximately equal-sized subsets, leaves out one subset, builds a classifier on the remaining samples, and then evaluates classification performance on the unused subset. This process is repeated k times for each subset to obtain the CV performance over the whole training dataset. If the training dataset is large, a small subset can be used for CV to decrease computing costs.

C. Random Forests

RF is a tree based classifiers combination algorithm [3]. It utilizes a bootstrapping method to randomly generate an In-Bag (IB) subset Tr_IB and an Out-of-Bag (OB) subset Tr_OB from Tr. A bootstrapping process generates Tr_IB of size $n'(n' <= n)$ by uniformly sampling from Tr with replacement. By sampling with replacement it is likely that some samples are repeated in Tr_IB. If $n' = n$ (100% bootstrapping), with large n the set Tr_IB is expected to have 63.2% samples of Tr, the rest being duplicates. The remaining 36.8% samples form Tr_OB in that case. If $n' < n$, less samples are expected to be in Tr_IB and more in Tr_OB. We can do bootstrapping multiple times on Tr to generate multiple IB subsets and OB subsets.

Next, an unpruned decision tree is modeled on each in-bag dataset. During the modeling of a decision tree, a small fraction of input features are randomly selected to determine the split at each node of the tree. A test sample is classified by majority vote from the ensemble of decision trees.

The OB samples can be utilized for accuracy estimation. For example, if we run 100 times 100% bootstrapping, a particular sample is not used for modeling 36.8 decision trees on average. The 36.8 decision tree predictions can be aggregated by major-voting as the OB prediction. Finally, the OB predictions over the whole training dataset can be used to estimate the generalization capability on unseen new samples.

By aggregation on an ensemble of weak but low-correlation tree classifiers, RF usually can significantly improve classification accuracy over one single decision tree classifier that is modeled on the original training dataset.

D. ROC analysis

Receiver Operating Characteristic (ROC) analysis and Area under ROC curve (AUC) have been designed to evaluate classification effectiveness/accuracy [6]. ROC and AUC can indicate a classifier's balance ability between its true positive (TP) rate and its false positve (FP) rate as a function of varying a classification threshold. An area of 1 represents a perfect classification, while an area of 0.5 represents a worthless model. By drawing the corresponding ROC curve, one classifier's TP rates under different FP rates can be visualized to help the administrator to control the balance between aggressive versus conservative spam senders detection.

SVM and RF are well known as superior classification algorithms in a high-dimensional feature space ($d > 1000$) and hence have been widely applied in many application domains. However, as far as we know, they are seldom used in information security domain, especially for email senders behavioral analysis.

V. Experiments

In this section, we present the experiments conducted and discuss the results. All classification modeling is carried out on a workstation with a Intel Xeon CPU at 1.86 GHz and 16 GB of memory.

A. Experiment Design

We extracted a records of about half million sending IPs from the TrustedSource dataset. These IPs have contacted 3,158 unique sensors. Since this dataset is still huge and contains far more spam IPs than non-spam IPs, we randomly undersample the set of spam IPs such that the number of spam IPs is twice the number of non-spam IPs. Table I lists the characteristic of the dataset for modeling before and after undersampling. We divide the dataset randomly into three equal-sized subsets such that each subset also has a two-to-one ratio between spam and non-spam IPs. Two subsets are used for training and the remaining subset is used as the testing dataset. The task is to build a classifier in this 3,158-dimensional space to discriminate spam IPs (labeled $+1$) from non-spam IPs (labeled -1).

We choose LIBSVM (available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`) for SVM modeling with

978-1-4244-2324-8/08/$25.00 © 2008 IEEE.

TABLE I

SENSOR DISTRIBUTION DATA ON NOV/08/2007

	original	after undersampling
#sensors	3,158	3,158
#IPs	491,566	54,393
#non-spam IPs	18,131	18,131
#spam IPs	473,435	36,262

TABLE II

EFFECTIVENESS/EFFICIENCY WITH 2/3 TRAINING AND 1/3 TESTING

Method	AUC-Testing	Time	Space
C4.5 Decision Tree	0.95714	243 mins	≈2 G
Random Forest	0.98919	339 mins	≈4 G
SVM $\gamma = 0.125$, $C = 128$	0.98466	91 mins	≈190 M

the RBF kernel. Because there are more spam IPs than non-spam ones, we use a false negative (FN) cost of 1 and the false positive (FP) cost of 2.

For SVM modeling, 5-fold CV is conducted for both parameter tuning and generalization capability estimation. Because CV is time-consuming, only a subset of 10% randomly selected training samples is used. As suggested by Hsu *et al.* [7], a grid search heuristic is used to select the best (γ, C) parameters from $\gamma = 2^{-15}, 2^{-13}, ..., 2^3$ and $C = 2^{-5}, 2^{-3}, ..., 2^{15}$. Therefore, there are 110 groups of parameters. For each group, a 5-fold CV is conducted. The group with the highest CV AUC value is used to build an SVM on the whole training dataset and to predict on the testing dataset.

We choose the R randomForest package (available at `http://cran.r-project.org/src/contrib/Descriptions/randomForest.html`) for RF modeling. 100 bootstrapping datasets are generated, on each of which a random tree is modeled. As the dataset is huge, 10% bootstrapping is used. For each tree modeling, 6% of features are randomly selected to decide the best feature for splitting at each node. Hence, each tree is a weak classifier and correlation among trees is low. We expect bootstrapping aggregation can boost the 100 weak tree classifiers into a highly accurate classifier.

B. Result Analysis

The effectiveness/efficiency analysis is presented in Table II. Notice that the modeling time includes both training time and testing time. We use the popular C4.5 decision tree classifier [8] as the baseline for comparison. Both RF and SVM are much more effective than C4.5. Between them, RF is slightly more accurate, but it is expensive in terms of both time and space. After modeling parameters are fixed, SVM can achieve a similarly accurate classification performance in far less time.

AUC values alone cannot justify the effectiveness. In a real-world spam filtering system, a classifier with $FP\_rate > 1\%$ is not acceptable. Fig. 3 depicts ROC curves with a cut-off at $FP\_rate \leq 10\%$ (left) and $FP\_rate \leq 1\%$ (right). Once again, these curves show that both SVM and RF are effective.

From the ROC curves, we can see that at $FP\_rate = 0.2\%$, SVM can catch 39.2% spam IPs while RF can catch 31.2% spam IPs based on features extracted from global email patterns. At $FP\_rate = 0.3\%$, SVM and RF can catch 45.8% and 51.7% spam IPs, respectively. In comparison, Ramachandran *et al.* reported a classification accuracy of $TP\_rate = 10\%$ and $FP\_rate = 5\%$ based on clustering of local email data [5]. The increased classification performance can be attributed to both the application of supervised learning techniques and the higher breadth of features using the TrustedSource dataset.

In TrustedSource, the sensor distribution features are combined with other informative features to generate a classifier with $TP\_rate \geq 90\%$ at $FP\_rate \leq 0.01\%$. The features incorporate additional data feeds and are out of the scope of this paper. Examples of other potential features are discussed in our previous work on behavioral spam filtering [4].

Also note that a classifier at the reputation-level does not replace a solution incorporating local analysis of messages. A reputation-based filtering solution aims at sifting out as much traffic as possible at the connection level to decrease the computational load for an in-depth local analysis.

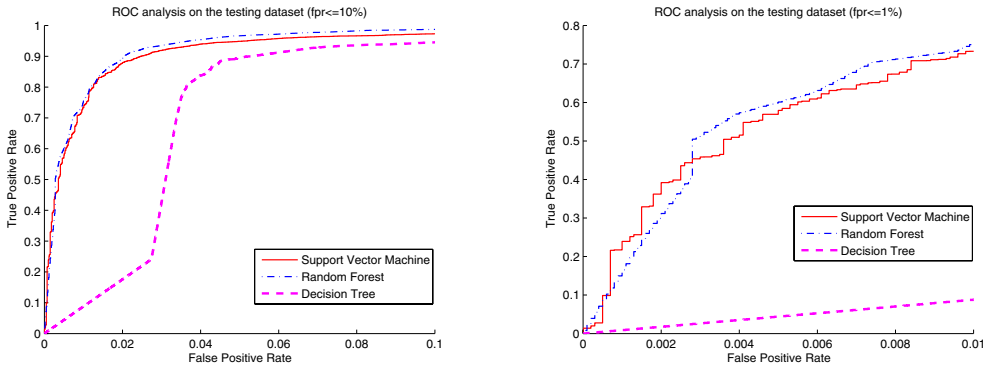
VI. RETURN OF INVESTMENT ANALYSIS

Based on the email traffic seen by TrustedSource, we present an estimation of the Return Of Investment (ROI) of the outlined classification techniques. With spam levels comprising over 90% of all email by early 2008, the hardship and costs endured by organizations and end-users due to this problem can be immense. They are caused by 4 factors: bandwidth, storage, processing, and productivity loss.

The bandwidth component of the overall cost of spam has been on the increase due to the rise in image-based spam [9]. The average size of a spam message in early 2008 is 9.7 kB. At the same time, spammers have been able to dramatically increase the volume of spam they are able to send on a daily basis by optimizing the performance of their mail deployment software and utilizing greater number of zombies to do the sending. Secure Computing has detected a 100% increase in mail volumes in 2007 and a 65% growth in new zombie machines in the same period. A mid-size organization consisting of 10,000 users can now expect to receive an average of 3 million spam messages each day, resulting in 29 GB of daily malicious traffic entering their network over the SMTP port. Assuming a conservative price of bandwidth of $100 per terabyte, this would cause a $1000 annual loss.

The second component of the spam cost is storage. Since many organizations do not immediately delete but instead quarantine identified spam due to the risk of misclassification of legitimate email, as well as regulations, such as Sarbanes-Oxley Act (SOX) in the United States, that demand archival of all records, including email, for up to 7 years (not every organization is required to be compliant with SOX and similar regulations, but in our experience most do try to archive their email for at least a number of years to decrease their legal liability risk). Even assuming a conservative 30% yearly growth in spam, the amount of storage that such an organization would need in 7 years just to keep all that spam would exceed 35

978-1-4244-2324-8/08/$25.00 © 2008 IEEE.



(a) ROC curves with $fpr \leq 10\%$

(b) ROC curves with $fpr \leq 1\%$

Fig. 3. ROC analysis on Nov/08/2007 sensor distribution data.

petabytes. With a relatively inexpensive storage cost of $2 per gigabyte, it would cost over $10 million annually.

In addition to the bandwidth and storage concerns, many organizations have had to purchase additional hardware and software processing capacity to cope with the sharp rise in incoming mail volumes attributed to spam. With the average cost of enterprise-level email security solutions being around $10,000-$50,000, the processing concern is major component of the overall spam cost. With most organizations purchase additional systems every 3 years, they can expect their average annual costs to be around $25,000.

Finally, the productivity loss associated with end-users having to read and delete the trickle spam that inevitably comes through their anti-spam solutions and arrives in their inbox, as well as the legitimate emails that may be misclassified and never received by the recipient are immeasurable.

When calculating the Return of Investment, inevitably one has to also consider the risk of false positives. While the costs associated with a critical email that is misidentified can be substantial, they are much harder to calculate. If an email is rejected while the SMTP connection to the sender is still open, the sender will typically get an instant notification that their message did not reach the destination and they can try other avenues to contact their recipient, which minimizes the cost of the false positives. If other actions are taken, such as a silent drop or quarantine where neither the sender or receiver are notified, the cost to organization can potentially be enormous depending on the importance of the email.

In our implementation of classification modeling, it is able to identify about 39.2% of all spam IPs at $FP\_rate = 0.2\%$ that we detect targeting our customers across the world. With a uniform distribution of those IPs, an organization of 10,000 users can expect an annual cost saving of up to $3.92 million from the introduction of this classifier in the reputation system they may be using by rejecting all connections originating from those IPs at their mail gateways.

VII. Conclusion

In this paper, we introduce two state-of-the-art classification algorithms, SVM and RF to detect malicious email servers. These methods utilize low-informative and high-dimensional sensor distribution data. Global sending behavior features are extracted from simple sensor query records that do not appear to be informative locally. We also show the classification results produced from real global Email traffic data. As far as we know, this is the first time a comparison study is conducted between these two well-known classification methods on a real world information security problem. Our study demonstrates that these features combined with advanced classification techniques can be reliable for spam detection. Both SVM and RF are effective resources for building accurate classifiers. Between them, RF is marginally more accurate but more expensive in terms of time and space. SVM generates similar accuracy in a much more efficient way if modeling parameters are fixed. The resulting classifiers contribute to TrustedSource as one source of input and prevent unwanted and malicious messages from being delivered to email users.

References

[1] "Secure Computing Corporation, TrustedSource website, http://www.trustedsource.org."

[2] V. N. Vapnik, *Statistical Learning Theory*. New York: John Wiley and Sons, 1998.

[3] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[4] Y. C. Tang, S. Krasser, P. Judge, and Y.-Q. Zhang, "Fast and effective spam IP detection with granular SVM for spam filtering on highly imbalanced spectral mail server behavior data," in *Proc. of The 2nd International Conference on Collaborative Computing*, 2006.

[5] A. Ramachandran, N. Feamster, and S. Vempala, "Filtering spam with behavioral blacklisting," in *CCS '07: Proc. of the 14th ACM conference on Computer and communications security*, pp. 342–351, 2007.

[6] A. P. Bradley, "The use of the area under the roc curve in the evaluation of machine learning algorithms.," *Pattern Recognition*, vol. 30, no. 7, pp. 1145–1159, 1997.

[7] C.-W. Hsu, C.-C. Chang, and C.-J. Lin, "A practical guide to support vector classification," tech. rep., Department of Computer Science and Information Engineering, National Taiwan University, Taipei, 2003.

[8] R. J. Quinlan, *C4.5: Programs for Machine Learning*. San Mateo, CA: Morgan Kaufmann, 1993.

[9] S. Krasser, Y. C. Tang, J. Gould, D. Alperovitch, and P. Judge, "Identifying image spam based on header and file properties using C4.5 decision trees and support vector machine learning," in *Proc. of the 2007 IEEE Systems, Man and Cybernetics Information Assurance Workshop (IAW 2007)*, pp. 255–261, 2007.

978-1-4244-2324-8/08/$25.00 © 2008 IEEE.
This full text paper was peer reviewed at the direction of IEEE Communications Society subject matter experts for publication in the IEEE "GLOBECOM" 2008 proceedings.