Identifying Potentially Useful Email Header Features for Email Spam Filtering

Omar Al-Jarrah*, Ismail Khater‡ and Basheer Al-Duwairi†
*Department of Computer Engineering
†Department of Network Engineering & Security
Jordan University of Science & Technology, Irbid, Jordan 22110
‡Department of Computer Systems Engineering
Birzeit University, Birzeit, West Bank, Palestine
Email: aljarrah@just.edu.jo, ikhater@birzeit.edu, basheer@just.edu.jo

***Abstract*—Email spam continues to be a major problem in the Internet. With the spread of malware combined with the power of botnets, spammers are now able to launch large scale spam campaigns causing major traffic increase and leading to enormous economical loss. In this paper, we identify potentially useful email header features for email spam filtering by analyzing publicly available datasets. Then, we use these features as input to several machine learning-based classifiers and compare their performance in filtering email spam. These classifiers are: C4.5 Decision Tree (DT), Support Vector Machine (SVM), Multilayer Perception (MP), Nave Bays (NB), Bayesian Network (BN), and Random Forest (RF). Experimental studies based on publicly available datasets show that RF classifier has the best performance with an average accuracy, precision, recall, F-Measure, ROC area of 98.5%, 98.4%, 98.5%, and 98.5%, respectively.**

***Index Terms*—Email Spam, Machine Learning**

I. Introduction

Email spam, defined as unsolicited bulk email, continues to be a major problem in the Internet. Spammers are now able to launch large scale spam campaigns, malware and botnets helped spammers to spread spam widely. Email spam cause many problems, increase traffic and leading to enormous economical loss. Recent studies [1], [2] revealed that spam traffic constitute more than 89% of Internet traffic. According to Symantec [3], in March 2011 the global Spam rate was 79.3%. The cost of managing spam is huge compared with the cost of sending spam which is negligible. It includes the waste of network resources and network storage, the cost of traffic and the congestion over the network, in addition to the cost associated with the waste in employees' productivity. It was estimated that an employee spends 10 minutes a day on average sorting through unsolicited messages [4]. Other studies [5], [6], [7] reported that spam costs billions of dollars. Ferris Research Analyzer Information Services estimated the total worldwide financial losses caused by spam in 2009 as $130 billion; $42 billion in the U.S. alone [8].

Spammers are increasingly employing sophisticated methods to spread their spam emails. In addition, they employ advanced techniques to evade spam detection. A typical spam campaign involves using thousands of spam agents to send spam to a targeted list of recipients. In such campaigns, standard spam templates are used as the base for all email messages. However, each spam agent substitutes different set of attributes to obtain messages that do not look similar. Moreover, spammers are increasingly adopting image-based spam wherein the body of the spam email is converted to an image, which renders text-based and statistical spam filters useless.

Blocking spam email is considered a priority for network administrators and security researchers. There have been tremendous research efforts in this field that resulted in a lot of commercial spam filtering products. Header-based email spam filtering is considered as one of the main approaches in this field. In this approach, a machine learning classifier is applied on features extracted from email header information to distinguish ham from spam, and the accuracy of the header-based email spam filter depends greatly on the email header fields used for feature selection. In this paper, we identify potentially useful email header features based on analyzing large publicly available dtatsets to determine the most distinctive features. Also, we include most of the mandatory and optional email header fields in order to fill any gab or missing information that is required for email classification.

This paper presents a performance evaluation of several machine learning-based classifiers and compare their performance in filtering email spam based on email header information. It also proposes including important email header features for this purpose. The rest of this paper is organized as follows: Section II reviews related work. Section III discusses the main features of email header considered in our work. Section IV evaluates the performance of different machine learning-based classifiers in filtering header-based email spam. Finally, Section V concludes the paper.

II. Related work

An email message typically consists of header and body. The header is a necessary component of any email message. The Simple Mail Transfer Protocol (SMTP) [15] defines a set of fields to be contained in the email message header to achieve successful delivery of email messages and to provide important information for the recipient. These fields include:

 ISBN: 978-1-61208-176-2

email history, email date, time, sender of the email, receiver(s) of the email, email ID, email subject, etc. Header-based email spam filtering represents an efficient and lightweight approach to achieve filtering of spam messages by inspecting email message header information. Typically, a machine learning classifier is applied on features extracted from email header information to distinguish ham from spam. For example, Sheu [10] categorized emails into four categories based on the title: sexual, finance and job-hunting, marketing and advertising, and total category. Then he classified them according to the attributes from email message header. He proposed a new filtering method based on categorized Decision Tree (DT), namely, applying the Decision Tree technique for each of the categories based on attributes (features) extracted from the email header. The extracted features are from the sender field, email's title, sending date, and the email's size. Sheu applied his filter on a Chinese emails and obtained accuracy, precision, and recall of 96.5%, 96.67%, 96.3%, respectively.

Wu [11] proposed a rule-based processing that identifies and digitizes the spamming behaviors observed from the headers and syslogs of emails by comparing the most frequent header fields of these emails with their syslog at the server. Wu noticed the differences in the header filed of the sent email from what is recorded in the syslog, and he utilized that spamming behavior as features for describing emails. A rule-based processing and back-propagation neural networks were applied on the extracted features. He achieved an accuracy of 99.6% with ham misclassification of 0.63%. Ye et al. [12] proposed a spam discrimination model based on SVM to sort out emails according to the features of email headers. The extracted features from email header fields are the return-path, received, message-id, from, to, date and x-mailer; They used the SVM classifier to achieve a recall ratio of 96.9%, a precision ratio of 99.28%, and an accuracy ratio of 98.1%.

Wang [13], presented a statistical analysis of the header session message of junk and normal emails and the possibility of utilizing these messages to perform spam filtering. A statistical analysis was performed on the contents of $10,024$ junk emails collected from a spam archive database. The results demonstrated that up to 92.5% of junk emails are filtered out when utilizing mail user agent, message-id, sender and receiver addresses as features.

Recently, Hu et al. [9] proposed an intelligent hybrid spam-filtering framework to detect spam by analyzing only email headers. This framework is suitable for extremely large email servers because of its scalability and efficiency. Their filter can be deployed alone or in conjunction with other filters. The extracted features from the email header are the originator field, destination field, x-mailer field, sender server IP address, and email subject. Five popular classifiers were applied on the extracted features: Random Forest (RF), C4.5 Decision Tree (DT), Nave Bayes (NB), Bayesian Network (BN), and Support Vector Machine (SVM). The best performance was obtained by the RF classifier with accuracy, precision, recall , and F-measure of 96.7%, 92.99%, 92.99%, 93.3%, respectively. These results were obtained when applying the classifiers on a dataset of $33,209$ emails and another dataset of 21,725 emails. The work presented in this paper focuses mainly on potentially useful header features for email spam filtering. These features were selected by analyzing publicly available dtatsets (described in Subsection IV-B). Table I provides a summary of the main email header features considered by different spam filtering techniques as reported in the literature. It also shows the main features that we consider in our work.

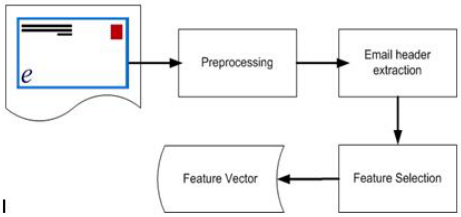

Fig. 1. The process of building feature vector of an email

III. Feature Selection

Feature selection represents the most important step of Header-based email spam filtering technique. In this step, we study information available in the email message header and carefully select some of them to be among the features used for classification. It is important to mention that the selection of email header features is based on analyzing large publicly available dtatsets (described in Subsection IV-B) to determine the most distinctive features. It is also important to point out that we include most of the mandatory and optional email header fields in order to fill any gab or missing information that is required for email classification. Figure 1 shows the process of building a feature vector of an email. This process starts by preprocessing of email messages to convert them into a standard format as described in RFC 2822. After that, we extract the header of the email to select the required features and build the feature vector which summarizes all the needed information from an email. This feature vector is then used to build the feature space for all emails that are needed for the classification phase.

The following subsections describe the fields of email message header that we consider in our work which turn to be of important value to classify email messages.

A. Received Field

Each email can contain more than one "Received" field. This field is typically used for email tracking by reading it from bottom to top. The bottom represents the first mail server that got involved in transporting the message, and the top represents the most recent one, where each received line represents a handoff between machines. Hence, a new received field will be added on the top of the stack for each host received the email and transport it, and to which host the message will be delivered, in addition to the time and date of passing. The following are the features that we extract from this field:

 ISBN: 978-1-61208-176-2

TABLE I
EMAIL HEADER FEATURES CONSIDERED BY DIFFERENT MACHINE LEARNING SPAM FILTERING TECHNIQUES

Sheu, 2009 [10]	Ye et. al., 2008 [12]	Wu, 2009 [11]	Hu et. al., 2010 [9]	Wang & Chen, 2007 [13]	Our Approach
Length of sender field , Sender field, Title (more than one category) , Time, Size of email	Received field (domain add., IP add., relay servers, date, time), From field, To field, Date field, Message-ID, X-Mailer	Comparing header fields with syslog	Originator fields, Destination fields, X-Mailer field, Sender IP, Email subject	Sender address validity, Receiver address (To, CC, BCC), Mail User Agent, Message-ID	Received field # of hops, Span Time, Domain add. Legality, Date & Time Legality, IP add. Legality, sender add. legality, # of Receivers (To, CC, BCC), Mail User Agent, Message-ID, Email subject Date of reception

1. *The number of hops*. This feature represents the number of the relay servers used to deliver the message from its origin to it its final destination. It was noticed based on different datasets that most of spam messages have a small number of hops. That means the spammers have exploited a predefined relay servers for delivering their spam, so the number of hops is limited, while in the normal case the number of relay servers may vary according to the paths the message follow to reach its final destination.
2. *Span time*. Span time represents the total time of the email through its journey from its origin to its final destination. This feature is considered as one of the most important features in our work. It is noticed that most of the spam emails have a large span time as compared to legitimate emails and some of them is negative in value.
3. *Domain address existence*. Domain address existence feature expresses whether the domain address of the host that delivers the message exists or not. This could be of little value to discriminate the spam emails from ham emails, but we keep it as a supporting feature.
4. *Date and time legality*. The purpose of this feature is to discover illegal date and time of email messages. The idea here is to check the date and time of email messages as they travel from one relay server to another. We believe this is an important feature because typically the date and time of legitimate email servers would be adjusted correctly. However, this is not necessarily the case for compromised machines that are used as email relays as we have discovered in the spam dataset.
5. *IP address legality*. This feature checks the legality of the host IP address, because spammers tend to hide or obfuscate IP addresses of their spam messages in order to avoid being blacklisted. We just check the format and the existence of the IP address.

B. Sender Address Legality

This feature is a conventional feature that is mentioned in most of the header based filters. The "From" field is one of the mandatory fields that every email must include, so the absence of this field is a cue for spamming behavior, the spammers tend to hide or use fake email addresses in order to avoid being blacklisted.

C. Number of Receivers

The recipients addresses of an email message are listed in one or more of the "To", "CC", and "BCC" fields. The "To" field contains the addresses of the primary recipients and the carbon copy "CC" field contains the addresses of the secondary recipients of the email, while the blind carbon copy "BCC" field contains the addresses of the recipients that are not included in copies of the email sent to the "To" and "CC" recipients. Many studies (e.g., [9], [13]) showed that spammers prefer to use the "BCC" field in order to send spam emails to a large number of recipients, at the same time no one of the recipients can obtain the list of the addresses that are collected by the spammers, because the SMTP server send a separate email to each one of the recipients listed in the "BCC" field, and every recipient has no information about the other recipients. In fact, most of the spam emails usually have small number of addresses in the "To" field which suggests that these emails were originally sent to many recipients using the "BCC" field such that individual recipients would not be able to identify other recipients of the same email.

D. Date of Reception

The "Date" field is a mandatory field that represents the date and time of the email when it is sent by the sender at the Mail User Agent (MUA). It is to be mentioned that the time recorded in this field is based on the location of the mail server of the sender which could belong to a time zone different from that of the recipient. Therefore, we convert all timing information into Universal Time Coordination (UTC) to have a common base for comparison. Basically, we compare the date of sending the email with the date of reception as recorded at the final hop in the "Received" field. We noticed that most spam emails do not have valid date of reception which suggests that this feature could be very helpful in our study.

E. Mail User Agent (MUA)

This is an optional field in the email header, appears as "X-Mailer" field which contains the email program used for the generation of the email. In this field, the email client or MUA name and version is recorded. Spammers usually tend to leave this field empty or fill it with random text. Based on

 ISBN: 978-1-61208-176-2

that, we take this field into consideration by checking whether it is existing or it is missing from the email message header.

F. Message-ID

This is a globally unique ID for each generated message. The "Message-ID" field is a machine readable ID which takes the name of the machine and the date and time of the email when it is sent. This field consists of two parts separated by @ sign. The right side part specifies the domain name or the machine name. This could be of a particular interest, because we noticed that most of spammers tend to hide this part or even fake the domain name to avoid being blacklisted. Therefore, it is required to make sure that the domain name in the "Message-ID" field is the same as the domain name in the "From" field. Inconsistency of this information would indicate a spamming behavior. It is important to mention here, that some mail user agents append the machine name to the domain name to the right of the @ sign. To overcome this issue, we used the partial matching with the domain name in the "From" field, and we noticed mismatches in most of the spam emails.

G. Email Subject

The subject contains a limited number of characters as described in RFC 822 and RFC 2822 [15]. It contains the topic and a summary of the email. Spammers may exploit the subject and use some special characters or words (e.g., "Try it for free!", "$ Money Maker $", "* URGENT ASSISTANT NEEDED *", etc.) to attract the user to open the email. Therefore, having special characters/phrases in the subject line may strongly indicate that the email is spam.

IV. Performance Evaluation

In this section, we evaluate the performance of several machine learning-based classifiers and compare their performance in filtering email spam based on email header information mentioned in Section III. In particular, we consider C4.5 Decision Tree (DT), Support Vector Machine (SVM), Multilayer Perception (MP), Nave Byays (NB), Bayesian Network (BN), and Random Forest (RF). Basically, our experiments involve evaluating the performance of these classifiers in terms of accuracy, precession, recall, and F-measure as defined Subsection IV-A using publicly available datasets. Email spam datatsets have been divided into a train and test sets according to the cross validation technique, where we used 10-fold cross validation. Weka tool [14] has been used for applying the machine learning techniques. Weka requires that the used features must conform to the input format of Weka. Therefore, the used features were ordered in a CSV file in the following format:

feature 1, feature 2, , feature n, class label

By default the class labels are located at the end of each row. In our experiments, we have two class labels used to categorize the image in the email, a legitimate email is marked as *Ham*, while the spam email is marked as *Spam*.

Prediction	Actual	
	Spam	Ham
Spam	TP	FN
Ham	FP	TN

Fig. 2. Confusion Matrix

A. Performance Metrics

We use the following standard performance metrics to evaluate the proposed technique: accuracy, precision, recall, F-measure, which are defined as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

$$F - measure = \frac{2Precision.Recall}{Precision + Recall} \quad (4)$$

where *FP, FN, TP, TN* are defined as follows:

- *False Positive (FP):* The number of misclassified legitimate emails.
- *False Negative (FN):* The number of misclassified spam emails.
- *True Positive (TP):* The number of spam messages that are correctly classified.
- *True Negative (TN):* The number of legitimate emails that are correctly classified.

Precision is the percentage of correct prediction (for spam email), while spam Recall examines the probability of true positive examples being retrieved (completeness of the retrieval process), which means that there is no relation between precision and recall. On the other hand, F-measure combines these two metrics in one equation which can be interpreted as a weighted average of precision and recall. In addition, we use Receiver Operating Characteristics (ROC) curves which are commonly used to evaluate machine learning-based systems. These curves are basically a two-dimensional graphs where TP rate is plotted on y-axis and FP rate is plotted on x-axis. Therefore, depicting the tradeoffs between benefits TP and costs FP [19]. A common method to compare between classifiers is to calculate the Area Under ROC Curve (AUC).

It is important to mention that our definition of the performance metrics is mainly based on the confusion matrix shown in Figure 2.

B. Datasets

Our experiments are based on the following two publicly available recent datasets.

- CEAS2008 live spam challenge laboratory corpus [16] which contains 32703 labeled emails. Among these emails there are 26180 spam emails and 6523 ham

 ISBN: 978-1-61208-176-2

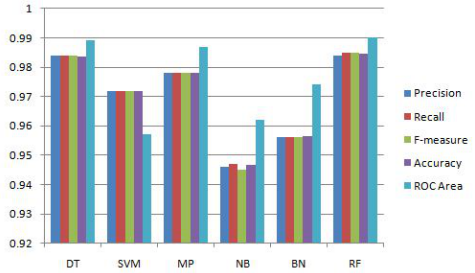

Fig. 3. The Performance of different machine learning techniques applied on CEAS2008 dataset in terms of Accuracy, precession, recall, F-measure, and ROC area

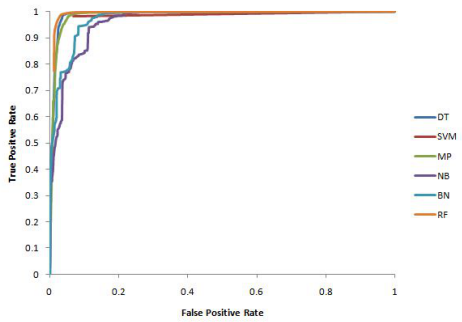

Fig. 4. ROC curves for the six classifiers applied on CEASE2008 dataset

email, this dataset was collected during the CEAS 2008 conference and it is considered as one of the TREC public spam corpus.

- CSDMC2010 spam corpus [18]. This dataset contains 4327 emails out of which there are 2949 non-spam (ham) emails and 1378 spam emails.

It is important to mention that these datasets were used for training and testing.

C. Experimental Results

1) Results based on CEAS2008 dataset: Figure 3 depicts the performance of the different classifiers in terms of accuracy, precision, recall, F-measure and the area under ROC. This figure shows the disparity among the classifiers in terms of precision, recall, F-measure and accuracy. It can be seen that RF classifier outperform all the other classifiers with an average accuracy, precision, recall, F-Measure, ROC area of 98.5%, 98.4%, 98.5%, 98.5%, and 99%, respectively. The ROC curves for all classifiers considered in this study are shown in Figure 4. This figure confirms that the RF classifier has the best performance compared to other classifiers as it maintains the best balance between false positive rate and true positive rate. DT classifier comes after RF classifier, then MP and SVM classifiers, while the BN and NB classifiers comes last. NB classifier was the worst in this group.

It is important to be mentioned that the results of other classifiers were as follows: DT classifier achieved an average precision and recall of 98.4%, which indicates that DT classifier succeeds in classifying most of the emails based on their header information. For the SVM classifier, it can be seen that it achieved good results for this dataset. However, the results were not that good in case of small size dataset as described in Subsection IV-C2. The other issue is the trade-off between FP and FN, which can be described by the ROC area. In the case of the MP classifier, the datasets were divided using the cross validation technique. Having the trained network; we can use it in recognizing spam emails of the testing set by invoking the simulation function, which takes the input feature vector and the trained network as inputs and computes the outputs according to the weights of the neurons, then it finds the output of the maximum weight. This classifier achieved an average precision and recall of 97.8%.

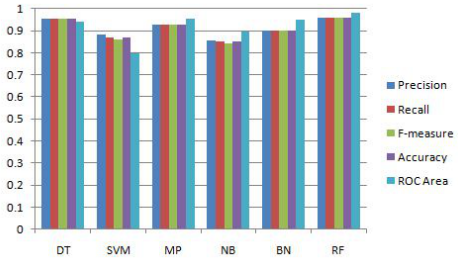

Fig. 5. The Performance of different machine learning techniques applied on CSDMC2010 dataset in terms of Accuracy, precession, recall, F-measure, and ROC area

2) Results Based on the CSDMC2010 dataset : In order to confirm the results obtained using CEAS2008 dataset, we repeated our experiments using another recent dataset (however, with smaller size). Figure 5 depicts the performance of the different classifiers using this dataset in terms of accuracy, precision, recall, F-measure and the area under ROC. It can be seen that RF classifier outperform all the other classifiers with an average accuracy, precision, recall, F-Measure, ROC area of 95.8%, 95.8%, 95.8%, 95.8% and 98.1%, respectively. It is to be noted that all classifiers achieved comparable performance this time indicating that the performance of some classifiers depends on the dataset used for testing and training.The MP classifier was very successful in recognizing 99In both cases, RF classifiers was on top of thlist in terms of performance. The ROC curves for all classifiers considered in this study are shown in Figure 6. This figure confirms that the RF classifier has the best performance compared to other classifiers as it maintains the best balance between false positive rate and true positive rate.

D. Comparison with Previous Work

In this subsection, we compare the performance of the proposed scheme with other header-based email spam filtering techniques ([9], [10], [11], [12], [13]) based on the results reported in the literature for these techniques. Table II shows the best performance obtained them and compare it to the results obtained using the proposed work. It can be seen that applying RF classifier to the email header features described in Section III results in better performance as compared to

TABLE II
PERFORMANCE OF THE PROPOSED WORK COMPARED TO OTHER HEADER-BASED EMAIL SPAM FILTERS. A: ACCURACY, P: PRECISION, R: RECALL, F: F-MEASURE

Spam Filter	Sheu, 2009 [10]	Ye et al., 2008 [12]	Wu, 2009 [11]	Hu et al., 2010 [9]	Wang & Chen, 2007 [13]	Our Approach
Classifier(s) used	DT	SVM	Rule-based & back-propagation NN	RF, DT, NB, BN, SVM	Statistical analysis	DT, SVM, MP, NB, BN, RF
Best performance obtained	A=96.5%, P=96.67%, R=96.3%	A=98.1%, P=99.28%, R=96.9%	A=99.6 % (ham misclassification = 0.63%)	RF (A=96.7%, P=93.5%, R=92.3%, F=93.3%)	92.5% of junk emails are filtered out	RF (A=98.5% , P=98.9%, R=99.2%, F=99%)

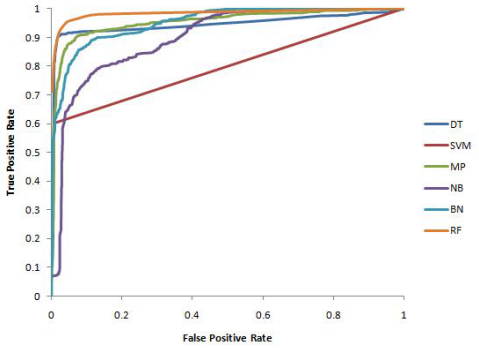

Fig. 6. ROC curves for the six classifiers applied on CSDMC2010 dataset

other header-based spam filters.

V. CONCLUSION

Spammers are increasingly employing sophisticated methods to spread their spam emails. Also, they employ advanced techniques to evade spam detection. A typical spam campaign involves using thousands of spam agents to send spam to a targeted list of recipients. In such campaigns, standard spam templates are used as the base of all email messages. However, each spam agent substitutes different set of attributes to obtain messages that do not look similar. In this paper, we evaluated the performance of several machine learning-based classifiers and compared their performance in filtering email spam based on email header information. These classifiers are: C4.5 Decision Tree (DT), Support Vector Machine (SVM), Multilayer Perception (MP), Nave Bays (NB), Bayesian Network (BN), and Random Forest (RF). We adopted header-based email spam filtering by including additional header information features that found to be of a great importance to improve the performance of this technique. We evaluate the proposed work through experimental studies based on publicly available datasets. Our studies show that RF classifier outperform all the other classifiers with an average accuracy, precision, recall, F-Measure, ROC area of 98.5%, 98.4%, 98.5%, 98.5%, and 99%, respectively.

REFERENCES

[1] C. Kreibichy, et al., "Spamcraft: An Inside Look At Spam Campaign Orchestration," Proceedings of the Second USENIX Workshop on Large-Scale Exploits and Emergent Threats (LEET '09), Boston, Massachusetts, April 2009.

[2] M. Intelligence, "MessageLabs Intelligence: 2010 Annual Security Report," 2010. Retrieved: July, 2011. Available at: http://www.clearnorthtech.com/images/MessageLabsIntelligence_2010_Annual_Report.pdf

[3] Symantec. March 2011 Intelligence Report. Retrieved: July, 2011. Available at: http://www.symantec.com/about/news/release/article.jspprid=20110329_01

[4] S. Hinde, "Spam, scams, chains, hoaxes and other junk mail," Computers & Security, vol. 21, pp. 592 - 606, 2002.

[5] A. R. B. Blog. October, 2010, The Dangers of SPAM. Retrieved: June, 2011. Available: http://www.anthonyricigliano.info/the-dangers-of-spam/

[6] A. C. Solutions. January 7, 2011 Statistics and Facts About Spam. Retrieved: July, 2011. Available: http://www.acsl.ca/2011/01/07/statistics-and-facts-about-spam/

[7] H. R. Cournane A, "An analysis of the tools used for the generation and prevention of spam," Computers & Security, vol. 23, pp. 154-66, 2004.

[8] R. JENNINGS. JANUARY 28, 2009, Cost of Spam is Flattening - Our 2009 Predictions. Retrieved: July, 2011. Available at: http://ferris.com/2009/01/28/cost-of-spam-is-flattening-our-2009-predictions/

[9] Y. Hu, et al., "A scalable intelligent non-content-based spam-filtering framework.," Expert Syst. Appl., vol. 37, pp. 8557-8565, 2010.

[10] J.-J. Sheu, "An Efficient Two-phase Spam Filtering Method Based on E-mails Categorization " International Journal of Network Security, vol. 9, pp. 34-43, July 2009.

[11] C.-H. Wu, "Behavior-based spam detection using a hybrid method of rule-based techniques and neural networks," Expert Systems with Applications, vol. 36, pp. 4321-4330, April, 2009

[12] M. Ye, et al., "A Spam Discrimination Based on Mail Header Feature and SVM," In Proc. Wireless Communications, Networking and Mobile Computing, 2008. WiCOM '08. 4th International Conference on Dalian Oct. 2008.

[13] C.-C. Wang and S.-Y. Chena, "Using header session messages to anti-spamming," Computers & Security, vol. 26, pp. 381-390, January 2007.

[14] Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. "The WEKA Data Mining Software: An Update. SIGKDD Explorations", 2009.

[15] P. R. Network Working Group, Editor. "Request for Comments RFC 2822," Retrieved: July, 2011. Available: http://tools.ietf.org/html/rfc2822.html

[16] CEAS 2008 Live Spam Challenge Laboratory corpus. Retrieved: March, 2011. Available at: http://plg1.uwaterloo.ca/cgi-bin/cgiwrap/gvcormac/fooceas.

[17] R. Beverly and K. Sollins, "Exploiting Transport-Level Characteristics of Spam," presented at the CEAS, Mountain View, CA, August 2008.

[18] C. GROUP. (2010, Spam email datasets, CSDMC2010 SPAM corpus. Retrieved: March, 2011. Available at: http://csmining.org/index.php/spam-email-datasets-.html

[19] T. Fawcett, "An introduction to ROC analysis," Pattern Recognition Letters - Special issue: ROC analysis in pattern recognition, vol. 27, pp. 861-874, June 2006

 ISBN: 978-1-61208-176-2