Spam Decisions on Gray E-mail using Personalized Ontologies

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

E-mail is one of the most common communication methods among people on the Internet. However, the increase of e-mail misuse/abuse has resulted in an increasing volume of spam e-mail over recent years. As spammers always try to find a way to evade existing spam filters, new filters need to be developed to catch spam. A statistical learning filter is at the core of many commercial anti-spam filters. It can either be trained globally for all users, or personally for each user. Generally, globally-trained filters outperform personally-trained filters for both small and large collections of users under a real environment. However, globally-trained filters sometimes ignore personal data. Globally-trained filters cannot retain personal preferences and contexts as to whether a feature should be treated as an indicator of legitimate e-mail or spam. Gray e-mail is a message that could reasonably be considered either legitimate or spam. In this paper, a personalized ontology spam filter was implemented to make decisions for gray e-mail. In the future, by considering both global and personal ontology-based filters, we can show a significant improvement in overall performance.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *information filtering*; I.2.6 [**Artificial Intelligence**]: Learning – *concept learning.*

Keywords

e-mail classification, spam filtering, ontologies, gray e-mail, feature selection.

1. INTRODUCTION

E-mail is an efficient and popular communication mechanism, and the amount of e-mail traffic is now huge. E-mail management has become an important and growing problem for individuals and organizations because it is prone to misuse.

The blind posting of unsolicited e-mail messages, known as spam, is an example of misuse. Spam is commonly defined as the sending of unsolicited bulk e-mail. A further common definition of spam is restricted to unsolicited commercial e-mail, a definition that does not include non-commercial solicitations such as political or religious pitches, scams, etc., as spam. We take the broader definition of spam as e-mail that was not requested or appropriate for a user, given his/her requests and privacy context.

Statistical text classification is at the core of many commercial and open-source anti-spam solutions. Statistical classifiers can either be trained globally with one classifier learned for all users, or personally where a separate classifier is learned for each user. Personally-trained classifiers have the advantage of allowing each user to provide their own personal definition of spam. A user actively refinancing his home can train a personal filter to delete unsolicited stock advice as spam but deliver unsolicited refinancing offers to his inbox. Another user might train a personal filter to block all unsolicited offers. Personal classifiers can quickly identify terms that are unique to an individual and use them as strong indicators of legitimate e-mail.

Spam filters are faced with the challenge of distinguishing messages that users wish to receive from those they do not. At first glance this seems like a clear objective, but in practice this is not straightforward. For example, it has been estimated that two-thirds of e-mail users prefer to receive unsolicited commercial e-mail from senders with whom they have already done business, while one-third consider it spam. Spam not only causes loss of time and computational resources, leading to financial losses, but it is also often used to advertise illegal goods and services or to promote online fraud. As suggested in recent reports by Spamhaus [1], spam is increasingly being used to distribute viruses, worms, spyware, links to phishing web sites, etc. The problem of spam is not only an annoyance, but is also becoming a security threat.

A great number of learning-based spam filters are proposed in the literature. Some of them use the knowledge about the structure of the message header, retrieving particular kinds of technical information and classifying messages according to it, for example the method based on SMTP path analysis [4]. Other methods use human language technologies to analyze the message content, for example, the approach based on smooth n-gram language modeling [5]. However, there is a large group of learning-based

SAC'09, March 8-12, 2009, Honolulu, Hawaii, U.S.A.
Copyright 2009 ACM 978-1-60558-166-8/09/03…$5.00.

filters that observe a message just as a set of tokens. The most popular method in this group is Naive Bayes [7].

Much work on spam e-mail filtering has been done using techniques such as decision trees, Naive Bayesian classifiers, neural networks, etc. To address the problem of growing volumes of unsolicited e-mail, many different methods for e-mail filtering are being deployed in many commercial products. We have experimentally constructed a framework for efficient e-mail filtering using ontologies. Ontologies allow for machine-understandable semantics of data, so they can be used in any system [9, 12, 13]. It is important to share the information with each other for more effective spam filtering. Thus, it is necessary to build ontologies and a framework for efficient e-mail filtering. Using ontologies that are specially designed to filter spam, most of unsolicited bulk e-mail can be filtered out on the system. This paper proposes an efficient spam e-mail filtering method using ontologies. As a part of spam filtering, we need to do some research about how to decide gray e-mail as spam or legitimate. It is very important to improve the performance of spam filtering. In our initial experiments, we used Waikato Environment for Knowledge Analysis (Weka) explorer, and Jena to make ontologies based on a sample data set.

The remainder of the paper is organized as follows: Section 2 describes gray e-mail and a personalized ontology, and introduces current spam e-mail trends; Section 3 introduces the personalized ontology spam filtering system to filter spam using user profile ontologies, and discusses experimental results; Section 4 concludes the paper with possible directions for future work.

2. MOTIVATION AND SPAM E-MAIL TRENDS

Gray e-mail is a message that could reasonably be considered either legitimate or spam. However, we are getting many e-mails that cannot be decided clearly every day, so handling of gray e-mail is very important issue in spam filtering system. Our approach is to resolve the problem by considering both advantages of global filter and personalized filter. Also, to cope with the potential new spam e-mail, we introduce the recent spam e-mail trends here.

2.1 DECISION ON GRAY E-MAIL

Gray e-mail is a message that could reasonably be considered either legitimate or spam. For example, unsolicited commercial e-mail, or newsletters that do not respect unsubscribe requests, could sometimes be useful. Message users prefer e-mail for personal communications or business transactions. Unsolicited e-mail like messages advertising illegal products, or phishing message, is spam. E-mail that we cannot agree on unanimously is gray e-mail. Gray e-mail can be considered as good e-mail for some people, or as bad e-mail for some other people. Hence, a personalized filter is required to handle the gray e-mail decisions by considering different user preferences or providing learning filters combining different training/testing policies.

For example, if a customer buys one pair of speakers in the last month, then advertising e-mail about amplifiers, home theater receivers, or speaker cables, can be good e-mail. However, if after buying the speakers, they receive advertising e-mail about speakers of another brand name, then the advertising e-mail is spam to this user

The gray mail problem can be treated as a special kind of label noise. Instead of accidentally flipping the label from spam to good or vice versa by mistake, different users may simply have different e-mail contexts and preferences. Another reason is that individual users change their own preferences over time. For example, it is common for a user who tires of a particular newsletter to begin reporting it as spam rather than unsubscribing [8, 11].

Some companies also do not respect unsubscribe requests and continue sending mail that some users then consider spam. In all cases the effect is the same. Senders send mail that is not clearly spam or good and spam filters are faced with the challenge of determining which subset of this mail should be delivered. There are two major problems in global anti-spam systems because of the presence of gray e-mail. First, when personalization or user preference and context are ignored, because gray e-mail is not clearly good or spam by definition, it makes accurate evaluation of a filter performance a challenge. Thus it is important that we are able to detect this mail and handle it appropriately in the context of anti-spam systems

In the research, a user profile ontology is created for each user or class of users to handle gray e-mail. Figure 1 shows the user profile ontology creation procedure. A structured taxonomy is created to serve as a global ontology filter, and user profile ontologies are created based on users' preferences and contexts. A user profile ontology creates a blacklist of contacts and topic words. If a user wants to block some contact persons, they add their e-mail addresses to the blacklist. Then e-mail from those addresses will be classified as spam by the filter. Also, if a user wants to repel a certain topic, then they can add the terms related to the topic. In this case, added terms are included into a feature set to classify the data set, so although a term is added to the topic list, not all e-mail including that term is classified as spam. However, in case of "do not receive" blacklist, e-mail with addresses in that list must be classified as spam.

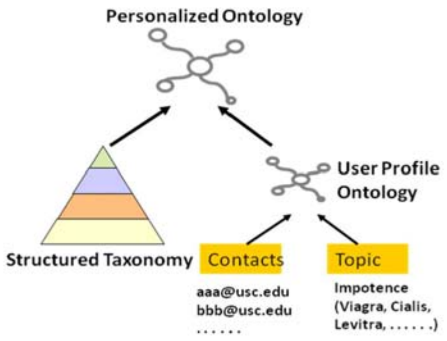


Figure 1. Personalized Ontology

2.2 SPAM E-MAIL TRENDS

New spam techniques appear again and again. We will look at several new spam techniques that have appeared since 2007.

2.2.1 PDF SPAM and Malware

A new technique of attaching popular Portable Document Format (PDF) files is designed to bypass many traditional spam filters.

This new spam technique using the PDF format prevents the adoption of blocking policies. The PDF spam has a randomized content, similar to the image-based stock spam, and randomly altered to fetter Optical Character Recognition (OCR) technology. Sometimes, the PDF spam is combined with a Malware URL.

Malware is software designed to infiltrate or damage a computer system without the owner's informed consent. It is a portmanteau of the words "malicious" and "software".

Generally, Malware means a variety of forms of hostile, intrusive, or annoying software or program code. Sometimes, the PDF spam is combined with a Malware URL. By clicking the Malware URL, an unauthorized program can be downloaded and installed automatically on user's machine.

2.2.2 Blended Spam with Malware Website Links

Zombies link text of a URL that hyperlinks to a website containing malicious software or commercial product instead of sending messages with the usual virus attachments. Malware distributor needs to hack into the legitimate site's web server to place the malware page there. Generally, e-mail filter often will assume that the URL within the site is also legitimate if the e-mail filter identifies the site as legitimate. Sometimes, spammers still use image spam to redirect to enhancement website. If user clicks the image, then it will redirect to the commercial website.

2.2.3 Discussion

Topics of spam of 2007 are shown on Figure 2.

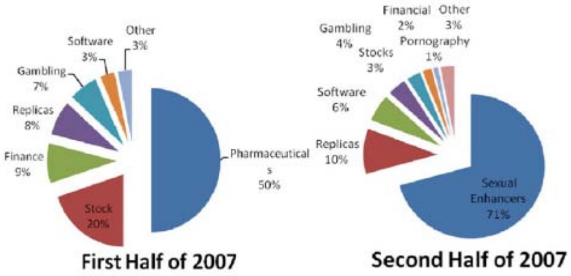


Figure 2. Topics of Spam E-mail

As we explained here, spamming techniques are also evolving to evade existing spam filtering techniques. New spamming technique appears continuously and traditional spamming technique is also prevailing. Topics of spam e-mail are also changing continuously. Spam filters are destined to modify and evolve to face various spamming techniques [2].

3. SPAM FILTERING SYSTEM

We have built the global spam e-mail filtering system in the previous paper [13]. Based on the previous spam filtering system, we created a personalized ontology filter. Specifically, the system works as follows:

1. A training data set is selected; this is a collection of text-oriented e-mail data.
2. Features from the data set are selected using the both *tfidf* and user profile ontology.
3. A 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.) [10]
4. Through Weka, classification results are generated.
5. The classified results are converted to an RDF file.
6. 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 [3].
7. 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, the personalized ontology filter is created. Details of the personalized ontology filter are shown in Figure 4.

In contrast to other approaches, ontologies were used in our approach. In addition, the C4.5 algorithm was used to classify the training data set [6]. The ontologies created by the implementation are modular, so those could be used in another system. In our previous classification experiments, the C4.5 showed better results than Naïve Bayesian, Neural Network, or Support Vector Machine (SVM) classifiers [12].

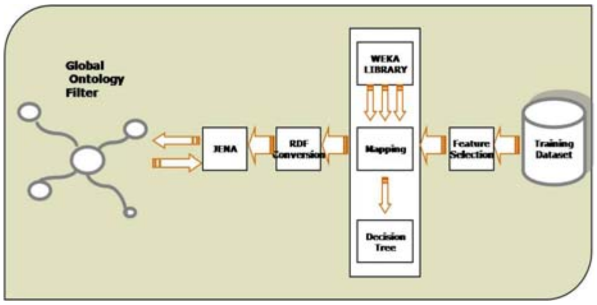


Figure 3. Global Spam Filter

Figure 3 shows the architecture of the global spam e-mail system provided in the previous paper [13].

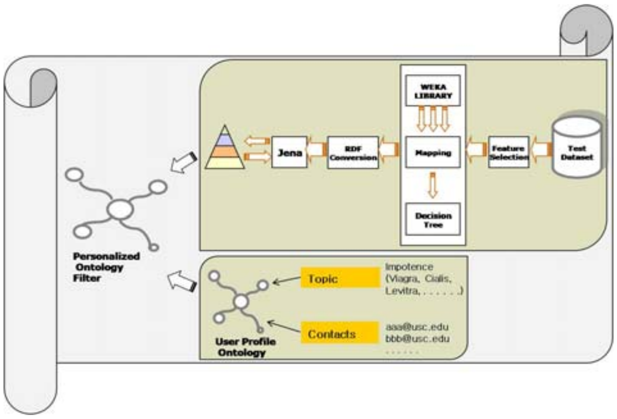


Figure 4. Personalized Ontology Spam Filter (POSF)

The training data set is the set of e-mail which gives us a classification result. The test data is actually the e-mail we will run through our system which we test to see if it is classified correctly as spam or not. This will be an ongoing test process and since the test data is not finite because of the learning procedure, the test data will sometimes merge with the training data. The training data set was used as input to the C4.5 classification. To do that, the training data set should be modified to a compatible input format. The proposed spam filtering system gives us the classification result using the C4.5 classifier.

To query the test e-mail in Jena, an ontology is created based on the classification result. To create the ontology, an ontology language was required. RDF was used to create an ontology. The classification result of the RDF format was input to Jena, and input RDF was deployed through Jena; finally, an ontology was created. An ontology generated in the form of an RDF data model is the base on which the incoming mail is checked for its legitimacy. Depending upon the assertions that we can conclude from the outputs of Jena, the e-mail can be defined as either spam or legitimate. The e-mail is actually the e-mail in the format that Jena will take in (i.e. in a CSV format) and will run through the ontology that will result in spam or legitimate.

Both spam filtering systems (global spam filter and personalized ontology spam filter) periodically update the data set with the e-mails classified as spam when user spam report is requested. Then, a modified training data set is input to Weka to get a new classification result. Based on the classification result, we can get a new ontology, which can be used as a second spam filter (that is, a user profile ontology). Through this procedure, the number of ontologies will be increased. Finally, these spam filtering ontologies will be customized for each user. User profile ontology filter would be different from the other depending on each user's background, preference, hobby, etc. That means one e-mail might be spam for person A, but not for person B. User profile ontologies evolve dynamically. The personalized ontology spam filter system provides an evolving spam filter based on users' preferences, so users can get a better spam filtering result.

The input to the system is mainly the training data set and then the test e-mail. The test e-mail is the first set of e-mail that the system will classify and learn and after a certain time, the system will take a variety of e-mail as input to be filtered as a spam or legitimate. The classification results through Weka need to be converted to an ontology. The classification result which we obtained through the C4.5 decision tree was mapped into the RDF format. This was given as an input to Jena which then mapped the ontology for us. This ontology enabled us to decide the way different headers and the data inside the e-mail are linked based upon the word frequencies of each word or character in the data set. The mapping also enabled us to obtain assertions about the legitimacy and non-legitimacy of the e-mail. The next part was using this ontology to decide whether a new e-mail is a spam or legitimate. This required querying of the obtained ontology which was again done through Jena. The output obtained after querying was the decision whether the new e-mail is a spam or legitimate. In summary, test e-mail is checked whether it is spam or legitimate based on global ontology created with training data set. In the personalized ontology spam filter, most of the procedure of spam filtering are the same as the global ontology spam filter. Additionally, it uses a user profile ontology created with user's spam report. With the help of adaptive user profile ontology, total spam filtering rate (the correct classification percentage) will be increased.

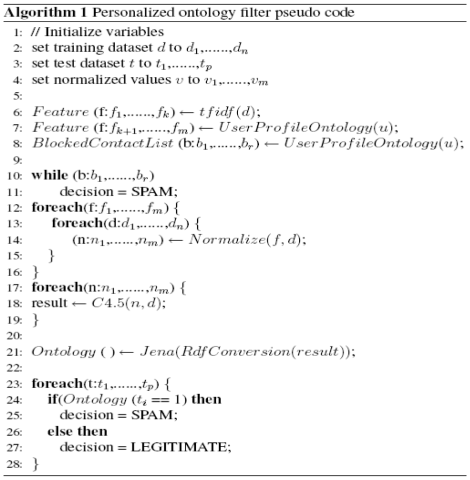
The primary way in which a user can provide the necessary feedback to the system would be through a GUI or a command line input with a simple 'yes' or 'no'. This would all be a part of a full-fledged working system as opposed to our prototype, which is a basic research, experimental system.

3.1 POSF Implementation

Personalized ontology spam filter is created as shown on Algorithm 1. Using a user profile ontology, the context and preferences of a specific user would be adapted in the feature selection procedure. A user profile ontology includes a list of people to block their e-mail and a list of words to block the e-mails related with some topic that is disliked by user. These blacklists and words will be combined with the words that were selected from the *tfidf* as shown in the algorithm.

3.2 Experimental Results

In our experiment, traditional C4.5 decision tree filter (We call it as a global ontology spam filter here) with the features selected by *tfidf* compared with the personalized ontology spam filter created using 200, 400, and 600 user data set. 2108 test data set (1008 spam and 1100 legitimate e-mail) was used.



Algorithm 1 Personalized ontology filter pseudo code

```
// Initialize variables
set training dataset $d$ to $d_1$,......,$d_n$
set test dataset $t$ to $t_1$,......,$t_p$
set normalized values $v$ to $v_1$,......,$v_m$

$Feature$ (f:$f_1$,......,$f_k$) $\leftarrow tfidf(d)$;
$Feature$ (f:$f_{k+1}$,......,$f_m$) $\leftarrow UserProfileOntology(u)$;
$BlockedContactList$ (b:$b_1$,......,$b_r$) $\leftarrow UserProfileOntology(u)$;

while (b:$b_1$,......,$b_r$)
      decision = SPAM;
foreach(f:$f_1$,......,$f_m$) {
    foreach(d:$d_1$,......,$d_n$) {
        (n:$n_1$,......,$n_m$) $\leftarrow Normalize(f, d)$;
    }
}
foreach(n:$n_1$,......,$n_m$) {
result $\leftarrow C4.5(n, d)$;
}

$Ontology$ ( ) $\leftarrow Jena(RdfConversion(result))$;

foreach(t:$t_1$,......,$t_p$) {
   if($Ontology$ ($t_i$ == 1) then
      decision = SPAM;
   else then
      decision = LEGITIMATE;
}
```

All the experimental results are summarized in Figure 6. Figure 5 shows an ROC curve of experimental results of the personalized ontology spam filter with 200, 400, and 600 user data set respectively. Receiver Operating Characteristic (ROC), or simply ROC curve, is a graphical plot of the sensitivity vs. (1 - specificity) for a binary classifier system as its discrimination threshold is varied. As we expected, a personalized filter created with a user profile ontology learns with the increase of the training data set. A personalized filter that was learned with 600 training data set shows better performance. By using a personalized filter, we can improve the performance of the spam filtering as you can see in Figure 5 and 6. We measured precision, recall, and correct classification rate through the experiment. ROC curve in Figure 5 shown that the filter is learning with more training data set, so experimental result with 600 training data set is better than those with 200 or 400 training data set. We compared the personalized ontology spam filter with traditional C4.5 decision tree filter (global ontology spam filter). As shown on Figure 6, the personalized ontology spam filter shows better experimental results than the global ontology spam filter (on spam recall, spam precision, legitimate recall, legitimate precision, and correct classification rate). From this result, by adding gray e-mail decision mechanism, we increased the spam filtering performance. When we increase the user data set for the personalized ontology spam filter, there was no more big improvement at some point. Also, we can see that the system evolves with the personalized ontology filter created with each user training data set.

We will use the followings for the Figure 5 and 6.

- **POSF 200 - Personalized Ontology Spam Filter with 200 user training data set**
- **POSF 400 - Personalized Ontology Spam Filter with 400 user training data set**
- **POSF 600 - Personalized Ontology Spam Filter with 600 user training data set**
- **Global – C4.5 decision tree filter**

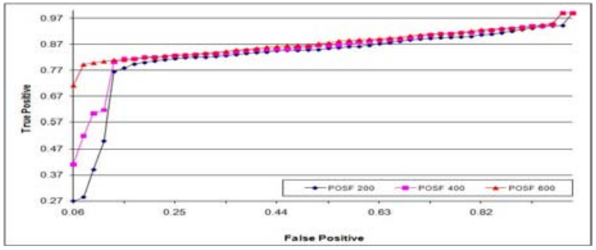


Figure 5. Experimental Results (ROC)

Precision and Recall were used as the metrics for evaluating the performance of each email classification approach.

$$\mathrm{Re}\,call = \frac{N_{ii}}{N}, \mathrm{Pr}\,ecision = \frac{N_{ii}}{N_i}$$

- **N = # of Total Interesting Email**
- **Ni = #of Email Classified as Interesting**
- **Nii = #of Interesting Email Classified as Interesting**

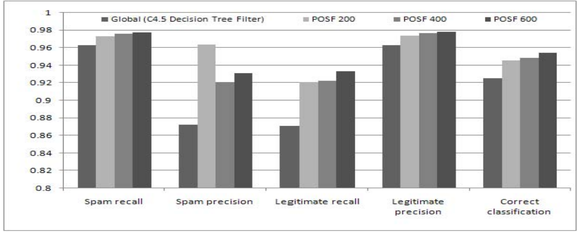


Figure 6. Experimental Results

4. CONCLUSION

In this paper, we suggested the method to detect spam e-mail from gray e-mail using personalized spam ontology filters. Global classifiers learned for a large population of users can leverage the data provided by each individual user across thousands of users. Proponents of a personalized classifier insist that statistical text learning is effective because it can consider the unique aspects of each individual's e-mail. There is a trade-off between globally- and personally- trained anti-spam classifiers. It is believed that globally-trained filters outperform personally-trained filters for both small and large collections of users under a real environment. However, globally-trained filters sometimes ignore personal data. Globally-trained filters cannot retain personal preferences and contexts as to whether a feature should be treated as an indicator of legitimate e-mail or spam. Hence, we used a personalized filter to make a decision based on personal preferences and context.

Gray e-mail can be considered as good e-mail for some people or as bad e-mail for some other people. Hence, a personalized filter is required to handle the gray e-mail decisions by considering different user preferences or providing learning filters combining different training/testing policies.

In the experiment, the personalized ontology spam filter could improve the spam filtering performance by including gray e-mail detection mechanism. As we explained, new spamming techniques appear continuously and traditional spamming techniques are also prevailing. Spamming techniques are advancing yet further, hence the spam filtering techniques also must catch up with the new spamming techniques. Even though some spam e-mails are classified as legitimate e-mails, we don't have to avoid the situation that valuable legitimate e-mail is classified as spam e-mail by considering personal data.

In the current system, *tfidf* as a feature selection algorithm and C4.5 decision tree as a classifier, were implemented. Various feature selection and classification algorithms should be compared to find the best system environment in the future. Also, we need to find ways how to make a user profile ontology more conveniently. In the future, we will experiment with the combination of the general corpus data set and our data set for generality.

5. ACKNOWLEDGEMENT

This work was funded in part by a grant from the Department of Homeland Security, ONR Grant number N00014-07-1-0149.

6. REFERENCES

[1] Burns, E. *The deadly duo: Spam and viruses*. Jun. 2006. http://www.clickz.com/stats/sectors/e-mail/print.php/3614491.

[2] Commtouch. http:// www.commtouch.com.

[3] An Introduction to RDF and the Jena RDF API. http://jena.sourceforge.net/tutorial/RDF_API/index.html.

[4] Liu, R. Dynamic Category Profiling for Text Filtering and Classification. In *Proceedings of the 10th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD '06)*, 2006, 255-264.

[5] Medlock, B. An adaptive approach to spam filtering on a new corpus. In *Proceedings of the 3rd Conference on E-mail and Anti-Spam (CEAS '06)*, 2006.

[6] Quinlan, J. Bagging, Boosting, and C4.5. In *proceedings of AAAI/IAAI, Vol. 1*, 1996, 725-730.

[7] Sahami, M., Dumais, S., Heckerman, D., and Horvitz, E. A Bayesian Approach to Filtering Junk E-Mail. In *Proceedings of the AAAI Workshop on Learning for Text Categorization*, 1998, 55-62.

[8] Segal, R. Combining global and personal anti-spam. In *Proceedings of 4th Conference on E-mail and Anti-Spam, (CEAS '07), 2007.*

[9] Taghva, K., Borsack, J., Coombs, J., Condit, A., Lumos, S., and Nartker, T. Ontology-based Classification of E-mail. In *Proceedings of the International Symposium on Information Technology (ITCC '03*), 2003, 194-198.

[10] *Weka: the Waikato Environment for Knowledge Analysis*. http://www.cs.waikato.ac.nz/~ml/publications/1995/Garner95-WEKA.pdf.

[11] Yih, W., McCann, R., and Kolcz, A. Improving Spam Filtering by Detecting Gray Mail. *In Proceedings of the 4th Conference on E-mail and Anti-Spam (CEAS '07), 2007.*

[12] Youn, S., and McLeod, D. A Comparative Study for E-mail Classification," In *Proceedings of International Joint Conferences on Computer, Information, System Sciences, and Engineering (CISSE '06)*, 2006, 387-391.

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