INTEGRATION OF FEATURE SETS WITH MACHINE LEARNING TECHNIQUES FOR SPAM FILTERING

C.R. Cyril Anthoni           Dr. A. Christy
Lecturer     Professor & Principal
Management Development Institute St. Mary’s School of Management Studies
Singapore     Rajiv Gandhi Road, Chennai – 119

ABSTRACT

In this era of Information Technology, Internet is emerging at a rapid pace and it is playing a major role in the field of communication, business, economy, entertainment and social groups all over the globe. E-mail is the most preferred tool for communication as it is easy to use, easy to prioritize, easier for reference, secure and reliable. As emails are so important for all the computer savvy people, email privacy has become an equally important concern. This paper proposes a novel approach for spam filtering using selective feature sets combined with machine learning techniques and the results shows that approach has produced less number of false positives.

Keywords: False positive, False negative, HMM, TF and IDF

INTRODUCTION

E-mail operates on the Simple Mail Transfer Protocol (SMTP). The protocol SMTP was written in 1982 without taking the problem of spam into consideration. The only way the end users used to differentiate the spam is from the sender’s address.

The task of spam filtering is to rule out unsolicited e-mails automatically from a user’s mail stream. Mails contain headers, which contain the source of the message and the examination of the email header can reveal the sender machine’s IP address. Previous studies have used the spam filtering techniques like Blacklist, Real-Time Blackhole List, Whitelist, Greylist, Content-Based Filters, Word-Based Filters, Heuristic Filters, Bayesian Filters, Challenge/Response System, Collaborative Filters and DNS Lookup Systems and Feature sets in filtering the e-mails containing spam. Naïve Bayes, a commonly used classification spam filtering algorithm is found to be sensitive to feature selection methods on small feature sets (Le Zhang, Jingbo Zhu and Tianshun Yao, 2004).
SECURITY ISSUES

The *Unsolicited Commercial E-mail (UCE)* messages are often used to force the users to reveal their personal information. Spam mails are commonly used to ask for information that can be used by the attackers. The safeguarding of email messages from unauthorized access and inspection is known as electronic privacy. Much of the Internet is susceptible to attacks. Emails travel across the unprotected paths. E-mail spams started growing from 1990s. Spammers collect e-mail addresses from chatrooms, websites, customer lists, newsgroups, and viruses which harvest user’s address books, and are sold to other spammers. According to a survey conducted by Message Anti-Abuse Working Group, the amount of spam email was between 88 to 92%. According to a Commtouch report in the first quarter of 2010, there are about 183 billion spam messages sent every day. The most popular spam topic is "pharmacy ads" which make up 81% of e-mail spam messages, Replica about 5.40%, Enhancers 2.30%, Phishing 2.30%, Degrees 1.30% and Weight loss 0.40%, etc. When grouped by continents, spam comes mostly from Asia (37.8%, down from 39.8%) , North America (23.6%, up from 21.8%) , Europe (23.4%, down from 23.9%) and South America (12.9%, down from 13.2%).

A study of International Data Corporation (IDC) ranked spams in the second position of the ISPs’ problems. Similarly technology has offered solutions to the email privacy issues. The security measures include cryptography, digital signatures and the use of secure protocols, which can ensure email privacy to a certain extent. PGP (Pretty Good Privacy) is one of the encryption standards that encrypts and decrypts the email message. This standard of encryption is a part of today's operating systems. Email encryption is another method that achieves email privacy. It is achieved by means of public key cryptography. The popularly used protocols for email encryption are S/MIME, TLS and OpenPGP.

Spam filtering can be recast as Text Categorization task where the classes to be predicted are spam and legitimate. A variety of supervised machine-learning algorithms have been successfully applied to mail filtering task. Traditionally, filters of this type have so far been based mostly on manually constructed keyword patterns. An alternative method is to use a naïve Bayesian classifier trained automatically to detect spam messages. The studies in Text Categorization have revealed that building classifiers automatically by applying some machine-learning algorithms (Support vector machine, Ada Boost and Maximum Entropy Model) to a set of pre-classified documents can produce good performance across diverse data sets.

SYSTEM ARCHITECTURE

To protect email privacy, it is obvious that the messages have to be encrypted. However, in order for the filtering process to be effective, the architecture is designed to reduce the false positives. We have used the standard metrics to measure the spam filtering accuracy. A ham email that is classified as spam by the filtering scheme is termed as a *false positive*. The false positive percentage is defined as the ratio of the number of false
positive emails to the total number of actual ham emails in the dataset used during the testing phase. A supervised learning algorithm is used, which retrieves the selected feature sets and finds is threshold value. The general architecture is depicted in shown in Figure 1.

![Figure 1 General architecture](image)

The feature preserving technique is based upon the concept of Shingles, which has been used in a wide variety of web and Internet data management problems, such as redundancy elimination in web caches and search engines, and template and fragment detection in web pages. Shingles are essentially a set of numbers that act as a fingerprint of a document. Shingles have the unique property that if two documents vary by a small amount their shingle sets also differs by a small amount.

Machine Learning learns through examples, domain knowledge and feedback. As Information Extraction systems are domain specific, machine learning plays a vital role in classification and prediction. The algorithm presented in Table 1 illustrates the architecture. A set of frequently identified spam words are selected as the training sets and its accuracy in identifying as spam is evaluated by various metrics. If the values are very close to 0, the feature sets resemble close to the training data sets.

<table>
<thead>
<tr>
<th>Input</th>
<th>D(F₀,F₁,…,Fₙ₋₁) // a training data set with N features</th>
</tr>
</thead>
<tbody>
<tr>
<td>S₀</td>
<td>// a subset from which to start the search</td>
</tr>
<tr>
<td>δ</td>
<td>// a stopping criterion</td>
</tr>
<tr>
<td>Output</td>
<td>S₀ is the best // an optimal subset</td>
</tr>
</tbody>
</table>

begin
initialize : S₀ = S₀;
γ₀ = eval(S₀, D,A);
// evaluate S₀ by a set of measures

do begin
S = generate(D); //generate a subset for evaluation
γ = eval(S, D,A); // evaluate the current subset S by A
if (γ is better than γ₀)
    γ₀ = γ;
    S₀ = S;
end until (γ is reached);
return S₀;
end;

Table 1 Algorithm
IDENTIFIED METRICS

The efficiency of the feature sets collected are evaluated by various metrics. The Baye’s rule is adopted for Hidden Markov Model which relates the conditional probability of an event

A given B to the conditional probability of B given A:

$$\text{PROB} (B|A) = \text{PROB} (A|B)*\text{PROB} (A)/\text{PROB} (B) \quad (1)$$

Each keyword can be represented as a term vector of the form $\bar{a} = (a_1, a_2, ..., a_n)$, each term $a_i$ has a weight $w_i$ associated with it and $w_i$ denotes the normalized frequency of word in the vector space, where

$$w_i = tf_i \times \log \left( \frac{D}{df_i} \right) \quad (2)$$

where $tf_i$ is the term frequency of $a_i$

$idf_i$ is inverse document frequency denoted as

$$IDF(w_i) = \log \left( \frac{|D|}{DF(w_i)} \right) \quad (3)$$

where $D$ is the total number of mails and $DF$ is the number of mails in which a term has appeared in a collection.

EXPERIMENTS

A list of keywords under the category Filter #1, Filter #2, Filter #3 and Filter #4 from the website ‘http://www.vaughns-1-pagers.com/internet/spam-word-list.htm’ are collected as feature sets and trained. The proposed method is compared with two popular spam filtering approaches, namely Bayesian filtering and simple hash-based collaborative filtering. In Bayesian filtering, the message is classified as a spam if its spamminess is greater than or equal to a preset threshold ($\mu$), and vice-versa. On the other hand, using hash based technique, its decision is based on the number of times the email corresponding to a particular hash value have been reported as spam. If this spam count of the hash value corresponding to in-coming email exceeds a threshold, the email is classified as spam, and otherwise it is classified as ham.

The datasets used in the experiments are derived from two publicly available email corpus, namely TREC email corpus and the SpamAssassin email corpus. One third of email set including ham and spam are used for training, and the remainder is used for testing.
The good words are selected randomly from a good word database created from the labeled ham data. We vary the amount of appended words in the range of 0% to 50% of the original emails’ word count and we call it degree of attack. The experimental setup consists of 12 agents with each agent employing 50% of the messages in its email set during the training phase. The initial experiments show that the supervised learning algorithm combined with HMM model is effective in identifying the feature sets, which reduces the amount of false positives.

CONCLUSION

Bayesian schemes are threshold-based approaches, so finding the appropriate threshold to achieve both low false positive and false negative rates is the key to the success of these approaches. The results from previous experiment are obtained when 50% of the emails in its email set are used during the training phase. We vary the threshold parameters of the two schemes and collect the false positive and false negative percentages. In Figure 2 and Figure 3 we plot the results of the experiment with false positive percentages on the X-axis and the false negatives on the Y-axis. The results show that neither of the approaches outperforms the other at all false positive percentage values. Users have a much lower tolerance of false positives than false negatives, and anything more than 1% percent false positives is usually considered unacceptable. Our initial experiments show that HMM approach is very effective in filtering spam, has high resilience towards various attacks, and it provides strong privacy protection to the participating entities.
REFERENCES


