Email classification for Spam Detection using Word Stemming

Ms.D.Karthika Renuka
Lecturer
Department of IT
PSG College of Technology

Dr.T.Hamsapriya
Assistant Professor
Department of IT
PSG College of Technology
Coimbatore, Tamilnadu-641004.

ABSTRACT

Unsolicited emails, known as spam, are one of the fast growing and costly problems associated with the Internet today. Among the many proposed solutions, a technique using Bayesian filtering is considered as the most effective weapon against spam. Bayesian filtering works by evaluating the probability of different words appearing in legitimate and spam mails and then classifying them based on that probabilities.Most of the current spam email detection systems use keywords to detect spam emails.These keywords can be written as misspellings eg: baank or bannk instead of bank. Misspellings are changed from time to time and hence spam email detection system needs to constantly update the blacklist to detect spam emails containing misspellings. It's impossible to predict all possible misspellings for a given keyword and add those to the blacklist. In this paper a better and more successful approach for improving E-mail content classification for spam control is proposed. It used the Word Stemming or Word Hashing Technique for improving the efficiency of the content based spam filter.The proposed system extract the base or stem of a misspelled or modified word, to detect spam emails. It considers every misspelled keyword applies a word stemming technique and passes the base word to the content based filter. Using a proposed if-then rule, we can decide whether or not this unknown mail is spam [1].This paper also provides an Email archiving solution which classifies the E-mail relating to a person, family, corporation, association, community, or nation.

Keywords

Spam, Filters, Bayesian, content based spam filter, Word Stemming, Email, Email archiving.

1. INTRODUCTION

Generally a content based spam filter works on words and phrases of email text and if it finds offensive content it gives that email a numerical value (depending on the content). After crossing a certain threshold, that email may be considered as SPAM. This technique works well only if the offensive words are lexically correct. That means the words must be valid words with correct spelling. Otherwise most content based spam filters will be unable to detect offensive words. In this paper, we showed that if we use some sort of word stemming or word hashing technique that can extract the base or stem of a misspelled or modified word, the efficiency of any content based spam filter can be significantly improved. Here we presented a simple word stemming algorithm specifically designed for spam detection.

Content based spam filters are useless if they cannot 'understand' the 'meaning' of the words or phrases in an email. Nowadays, spammers change one or more characters of offensive words in their spam in order to foil content based filters. But the important thing to observe is that the spammers change the words in such a way that a human being can understand the meaning the words without any difficulty. Spammers do not make any drastic change in the words so that it can be easily recognized by humans. Based on the above mentioned observations, we developed a rule based word stemming [3] technique that can match words those both look alike and sound alike. For example, the versions of the word 'Viagra', 'Via*gra', 'Vi\gra!', 'V.i-a.g*r.a' etc. cannot be detected by conventional spam filters.

2. USERS ATTITUDES TOWARD SPAM

Internet technologies such as electronic mail, web sites and digital media offer companies, the abilities to expand their customer reach, target specific communities and communicate as well as interact with customers in a highly customized manner. In the last few years, electronic mail has emerged as an important marketing tool to build and maintain closer relationships with customers as well as prospects. E-mail marketing has become a popular choice for several companies as it greatly minimizes the costs associated with other conventional methods such as direct mailing, cataloging and telecommunication marketing. The growth in the use of e-mail marketing has been accompanied by an enormous increase in the amount of Unsolicited Commercial e-mail (UCE), popularly known as Spam [5]. The unprecedented amount of unsolicited messages is now recognized as a serious problem, costing the community billions of dollars every year. The problem of Spam extends beyond household Internet users to the realm of companies, as many precious employee hours are being wasted due to spam messages. Internet users have reported that they trust email less, and 29% of users even say they use e-mail less because of Spam[13]. They complain that it uncontrollably clutters their inboxes and imposes uninvited, deceptive, and often disgustingly offensive messages.

- 63% of email users say spam has made them less trusting of e-mail in general.
- 77% of email users say spam has made being online unpleasant or annoying.
- 30% of e-mail users are concerned that their filtering devices may block incoming e-mail.
- 23% of e-mail users are concerned that their e-mails to others may be blocked by filtering devices
- 73% of e-mail users avoid giving out their e-mail addresses; 69% avoid posting their email addresses on the Web.
- 86% of e-mail users report that usually they "immediately click to delete" their incoming spam.

- 5% of e-mail users report that they have ordered a product or service that was offered in an unsolicited email.

3. RELATED WORK

To effectively combat Spam, an adaptive new technique that must be familiar with spammers' tactics is needed as they change over time. The proposed Spam Filter works based on two techniques:

Content based spam filter which works on words and phrases of email text and if it finds offensive content it gives that email a numerical value (depending on the content). After crossing a certain threshold, that email may be considered as Spam [6]. This technique works well only if the offensive words are lexically correct. That means the words must be valid words with correct spelling. Otherwise most content based spam filters will be unable to detect offensive words.

Word stemming or word hashing technique that can extract the base or stem of a misspelled or modified word, so the efficiency of any content based spam filter can be significantly improved. A simple rule-based word stemming algorithm has been specifically designed for spam detection.

4. CONTENT BASED SPAM FILTER

Historical information about the messages sent and received by the cluster is obtained and this information is used to reclassify messages by structure i.e. the contact list of the user. The key idea is that the set of distinct recipients that spammers and the legitimate users sent messages to, as well as the set of distinct senders from which users receive messages from, can be used as identifiers of senders and recipients in e-mail traffic. The Content Based spam filter technique exploits the structural similarity in groups of senders and recipients of e-mail. The basic assumption is that, users have the list of peers they often have contact with (i.e. they send / receive a email to / from). The contact list certainly changes over time however it is expected to be much less variable than other identifiers commonly used for spam reduction [3].Content based spam filter uses Bayes theorem for detecting spam mails. Bayesian inference uses aspects of the scientific method, which involves collecting evidence that is meant to be consistent or inconsistent for a given hypothesis. Bayesian inference uses a numerical estimate of the degree of belief in a hypothesis before evidence has been observed and calculates a numerical estimate of the degree of the belief in the hypothesis after evidence has been observed [1].

Bayes theorem [2] as shown in Eq.1, relates the conditional and marginal probabilities of stochastic events A and B:

$$P(A/B) = \frac{P(B/A)\,P(A)}{P(B)} \qquad (1)$$

P (A) is the prior probability or marginal probability of A. ‘prior’ in the sense that it doesn’t take into account any information about B.

- P(A/B) is conditional probability of A, given B
- P(B/A) is conditional probability of B, given A
- P(B) is the prior or marginal probability of B

```
for all arriving message m do
    Class =classification of m by auxiliary detection method;
    sc =find cluster for m.sender;
    Update spam probability for sc using mClass;
    $P_s(m)$ =spam probability for sc;
    $P_r(m) = 0$;
    for all recipient r in m. recipients do
        rc =find cluster for r;
        Update spam probability for rc using   mClass;
        $P_r(m) = P_r(m)$ +spam probability for rc;
    end for
    $P_r(m) = P_r(m)$/size (m. recipients)
    SP (m) = compute spam rank based on $P_s(m)$ & $P_r(m)$;
    if SP (m) > w then
        classify m as spam;
    else if $SP(m) < 1 - w$ then
        classify m as legitimate;
    else
        classify m as mClass;
    end if
end for
```

Figure 1. Algorithm for Email Classification

Content based classification is based on the algorithm as specified in Fig. 1.These filters are useless if they cannot ‘understand’ the ‘meaning’ of the words or phrases in an email. Nowadays, spammers change one or more characters of offensive words in their spam in order to foil content based filters. But the important thing to observe is that the spammers change the words in such a way that a human being can understand the meaning of the words without any difficulty. For example, the versions of the word ‘sex’, ‘s*e$x’, ‘s\e..x!’, ‘S.e-x.’ etc. cannot be detected by conventional spam filters.

5. WORD STEMMING TECHNIQUE

“Stemming”, is used to conflate the morphological variants thereby broadening the results. A stemming algorithm is an algorithm that converts a word to a related form. One of the simplest such transformations is conversion of plurals to singulars.

- Affix removal algorithms
- Successor Variety
- Table Lookup
- N-gram

The features that have been considered for word stemming are as follows:

- Replace consecutive repeated characters by a single character.
- Use phonetic algorithms on the resultant string.
- Give it a numeric value depending on the operations performed over it.
- Use this resultant string (numeric value) to look up a table (that contains a list of offending words where each word has a range of acceptable values)
- Replace original word with that of the table.

Figure 2.Algorithm for word stemming/hashing

6. MAIL ARCHIVING

E-mail archiving is a systematic approach to saving and protecting the data contained in e-mail messages so it can be accessed quickly at a later date. In the past, companies often relied on end-users to maintain their own individual e-mail archives. The IT department would back up e-mail, but not in a manner that made messages searchable. If a specific e-mail needed to be traced, it often took weeks to find it. With today's compliance legislation and legal discovery rules, it has become necessary for many IT departments to manage the entire company's e-mail archiving in bulk so specific messages can be located in minutes, not weeks [6]. Email archiving solution which classifies the E-mail relating to a person, family, corporation, association or community.

7. EXPERIMENTAL RESULTS

The proposed anti spam solution uses two PC's as SMTP client and one PC as the back-end SMTP server. The configurations of the PC's are listed below:

SMTP Client's: These machines are equipped with an Intel Pentium(R) D IV (2.80 GHz) CPU with 504 MB RAM, running Microsoft Windows XP.

SMTP Server: Intel Pentium(R) D IV (2.80 GHz) CPU with 504 MB RAM, running Microsoft Windows XP.

This program is implemented in JAVA for receiving the SMTP connections. The proposed Spam filter algorithm which is been employed in the SMTP server made a correct classification of the ham and spam emails. The efficiency of the stemming technique is been plotted in the Fig.3 which concludes that with word stemming technique the classification between the ham and spam email is made accuracy of 96%.These results are more efficient when compared with the previous techniques.

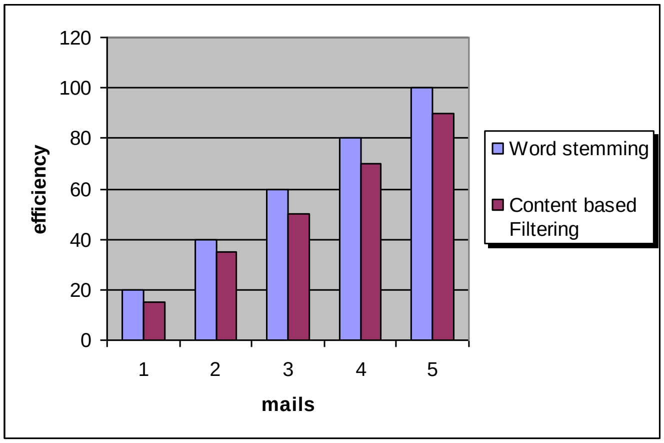

Figure 3.Efficiency of Spam Filter

The fact that spammers modify words in such a way so that the words can be easily recognizable by a human being was the key to build this word stemming technique.

8. REFERENCES

[1] Leonard and Hsu, 2001. Bayesian methods: an analysis for statisticians and interdisciplinary researchers. Cambridge University Press, Cambridge.

[2] Bernardo and Smith, 1994. Bayesian theory, John Wiley and Sons, Chi Chester.

[3] Clayton, R. (2004). Stopping spam by extrusion detection. Proceedings of the First Conference on Email and Anti-Spam (CEAS).

[4] Orwant J. et al. *Mastering Algorithms with Perl*. O'Reilly and Associates, ISBN: 1-56592-398-7, 1999.

[5] Amavisd-new Home Page, http://www.ijs.si/software/amavisd, Accessed 01 July 2004.

[6] Send mail Home Page, http://www.sendmail.org, Accessed 01, July 2004.

[7] Spam Assassin Home Page, http://www.spamassassin.org, Accessed 01, July 2004.

[8] Proc mail Home Page, http://www.procmail.org, Accessed 03, Mar 2004.

[9] Graham, P. *Better Baysian Filtering*. In Proceedings of Spam Conference, 2003.

[10] http://www.Blog Spam Database.com

[11] http://www.Email Spam Filter Word List.com

[12] http://www.ceas.cc/papers-2004/172.pdf.

[13] Internet Users and Spam: What the attitudes and behavior of Internet users can tell us about fighting spam ,Deborah Fallows Pew Internet & American Life Project, Washington, DC, 20036 USA.