International Journal of Engineering Research and Development
***e**-ISSN: 2278-067X, **p**-ISSN: 2278-800X, www.ijerd.com*
Volume 3, Issue 4 (August 2012), PP. 05-09

Diagnosis of Email Spams – Some Statistical Considerations

K. Srikanth[1], S.Ramakrishna[2]
Department of Computer Science, S.V.University, Tirupati, 517502, India

Abstract—While email is one of the fastest form of communication, the user is frequently faced with receiving unsolicited emails called spams. Non- spam mails are known as hams which are legitimate mails. It is practically very difficult to perfectly classify a mail into spam or ham basing on the content or subject of the mail. Several statistical methods are available which classify mails with some chance of misclassification. The most popular is the Bayesian approach which use the conditional probability of occurrence of given words in the spam/ham groups of the training data. Most of the content-based classifiers are based on word tokenization leading to large corpus of words along with their probabilities of occurrence. In this paper, we discuss some statistical properties of data sets used as corpora for training classifiers.

I. THE NATURE OF SPAM MAILS

Email has been an efficient and popular communication medium as the number of internet users increase during the recent ten years. Therefore email management is an important and growing problem for individuals and organizations because it is prone to misuse. Email spam is an *unsolicited*, unwanted email that is sent indiscriminately, directly or indirectly by a sender having no relationship with recipient. Email spam has steadily grown since 1990's. According to Rebecca Lieb (2002), Botnets networks of Virus-infected computers used to send about 80% of spam. A significant amount of time and resources is wasted by examining the spams and deleting them and the cost is borne by the recipient. Spammers are the people who send unsolicited mails to different users. Spammers collect email address from chat rooms, websites and customer lists, which harvest users address books and sell to other spammers

Characteristics of Spam mails

Filtering of spam emails is an important task for email providers as well as individuals. We need to be able to distinguish spam from legitimate mails. To do this, we need to identify typical spam characteristics and filters, to block these spam messages. Individuals can define their own filters to block unsolicited mails.

Spammers continuously improve their spam tactics and create disturbances to the internet users. So it is important to keep up to date on new spam filters from time to time to make spam blocked. Spam characteristics generally appear in two parts of a message, Email headers and message contents.

Email Headers indicates the way the mail reaches the destination. It has other information about sender and recipient address, message ID, date and time of transmission, subject and other email characteristics. Most of the spammers try to hide their identity by forging email headers to hide the real source of message. Spammers use mass mailing method to send mails to large number of recipients.

Message contents use certain language in their email message where companies use to distinguish spam messages from others. Words and phrases like free, click here, act now, risk free, lose weight, earn money, exclamation marks (!) and capital letters in their messages to attract the attention of the recipient.

Many spam emails are mainly from web ads. According to the commtouch report (2010), there were 183 billion of spam mails sent daily to internet users. Among them, the most popular is Pharmacy ads with (81%) followed by Replica (5.40%), Enhancers (2.30%), Phishing (2.30%), Degress (1.30%) and Casino (1%).

II. STATISTICS OF EMAIL SPAM

Spam mails are constantly increasing day by day, in which the amount of spam for internet users in their mailboxes is only a portion of total spam sent. According to Josh Halliday (2011), the quantum of spam messages was estimated to be around 200 billion sent per day. A survey by European email users and US (2010) showed that despite knowing the risks of opening spam mails, 46% of users still opened them and putting their computers into risk.

As per the year wise report, from the starting of 2002 there were 2.4 billion spam mails per day and in the year 2004 it reached to 11 billion per day. In the mid of June 2007 it goes to 100 billion spam mails per day. Spam mails rate increased a lot for the year 2010 January reached to 183 billion per day. According to Steve Ballmer (2004), Microsoft founder Bill Gates receives four million emails per year, most of them spam, at the same time Jef Poskanzer (2006), owner of the domain name "acme.com", was receiving over one million spam e-mails per day. Sophos (2008), has reported the countries which are major sources of spam as given in Table1.

Table-1: country wise spam statistics

S.No	Country	Status
1	The united states	Hiked from 18.9% to 19.8%

2	China	Hiked from 5.4% to 9.9%
3	Russia	Fallen from 8.3% to 6.4%
4	Brazil	Hiked from 4.5% to 6.3%
5	Turkey	Fallen from 8.2% to 4.4%

2.1 Methods of Detecting spam Mails

Classification is the main tool for email management. A classifier is a method of arranging the data into two groups “Spam” and “Ham”. There are different methods used in classification of spam mails among which Bayesian and Paul Graham methods are known to be best probabilistic classifiers

2.2 Bayesian Classification Method

Bayesian classification is a statistical method to filter spam mails. It makes use of a Naïve Bayesian classifier to identify spam email. Bayesian method works by correlating the use of tokens, with spam and Non-spam emails and then using Bayesian statistics to calculate the probability whether an email is spam or not. The scholarly publication on Bayesian spam filtering was by Sahami (1998).

Mathematical foundations for Bayesian email filter take the advantage of Bayesian theorem. Bayesian method use conditional probability to filter spam mails given as

$$\Pr(S/W) = \frac{\Pr(W|S).\Pr(S)}{\Pr(w|S).\Pr(s) + \Pr(W|H).\Pr(H)}$$

where

- Pr(S/W) is the probability that a message is a spam, knowing that the word is in it.
- Pr(S) is the overall probability that message is a spam
- Pr(W/S) is the probability that the word appears in the message
- Pr(H) is the overall probability that any given message is ham
- Pr(H/W) is the probability that the word appears in ham messages

2.3 Paul Graham Method

Paul graham is different way of implementing Naïve Bayes classifier on spam classification. Paul graham in his datasets contains groups of emails spam and ham. The size of his training set is almost same. For better spam filtering Paul graham focused on the domain of tokens by considering alphanumeric characters, dashes, apostrophes and dollar signs as tokens and all other token separators.

Paul graham in his method wants to reduce false positives in the classification. He uses two ways; one is doubling the number of appearances in ham for each token in email. By doing this we find out which token rarely used in legitimate email and consider some of tokens not appearing in ham at all. The other way is slight bias Graham used to fight against false positives using number of emails in each individual email group, rather than total number of emails in both groups.

Modified Paul graham method defined as

$$\Pr(S/W) = \frac{\Pr(W|S)}{\Pr(W|S) + 2\Pr(W|H)}$$

III. TOKENIZATION – A Basic Need For Spam Detection

Tokenization is the process of reducing a message to its colloquial components, like individual words, word pairs and other small chunks of text. It is an important component in statistical filtering of emails. The data generated by the tokenizer will passed on for further analysis and interpretation by statistical tools.

Tokenization is a type of heuristic process that is usually defined once at the time of building the tokenizer and rarely requires further maintenance. The fundamental goal of tokenization is to separate and identify specific features of a text. Tokenization of text starts with separating the message into smaller components, which are usually plain words. The primary delimiter is a white space, since a space separates the words in most of the languages. For instance, consider the message ‘*Are you unique enough? Find out from 30th August. www.areyouunique.co.uk’*. This message has two delimiters-white space and ‘?’. There are nine distinct words called *tokens*. Several computer codes are available for tokenization and word count.

The statistical approach for filtering is based on discrimination between spams and hams basing on *token probabilities*. The basic question is “ what is the chance of getting a specific token say T_i in spam(or in ham)”. A large

number of messages whose nature is already known (spam/ham) shall be tokenized to estimate the proportion of each token in all the tokens. This proportion is a estimate of the probability of the token T_i. If N denotes the total number of distinct tokens then $P_i = N_i/N$ is the estimated probability. Interestingly this probability is based on (a) the number of available tokens (Corpus) and (b) the method used for tokenization.
In the following section some statistical methods of comparing various filters are outlined.

IV. STATISTICAL EVALUATION OF FILTERS

There are several measures of performance to assess the discriminating ability of a filter between spam and ham mails. These are

1. Percentage of misclassification
2. Diagnostic Odds Ratio
3. Receiver operating Characteristic (ROC) analysis

Some of the popular statistical tools used for filtering are

1. Bayesian filter which is based on conditional probability with a cutoff probability of 0.5.(several improvements of simple Bayesian filter are reported in the recent times.)
2. Fisher's Linear Discriminant Analysis (LDA) which is Multivariate statistical procedure. A filter by this method provides a linear combination of token probabilities with suitable weights (Estimated form data). The output of this function is a number between 0 and 1 and treated as the probability of a message being spam called *Group Membership Probability.* When this probability is greater than 0.5 the message is classified as spam.
3. Logistic regression which is another multivariate tool that works similar to LDA and considered more efficient than the Bayesian filter.
4. Among all these Bayesian is popular due its simplicity in implementation.

V. CONCLUSIONS

Diagnosis of email spam has an important role in assessing the quality of email services. Various email providers use different filters to ensure that spam mails are properly detected. However, no single rule can correctly classify spam mails because of its probabilistic nature and spammers adopt probabilistic strategies (like in Game theory). The efficiency of the filter also changes with time because the spammers strategies are dynamic. Instead of simple tokenization of mail-text it is time to explore other tokens like images, special characters and suspicious words as identifiers of spam. There is ample scope for program development in this area.

ACKNOWLEDGMENTS

The authors express their thanks to Prof.K.V.S.Sarma of Statistics S.V.University for his valuable suggestions on statistical aspects in this paper.

REFERENCES

[1]. Juan Carlos Gomez and Marie-Francine Moens (2010) "Using Biased Discriminant Analysis for Email Filtering"

[2]. SpamBayes, (2002) http://spambayes.sourceforge.net/.

[3]. Nikila Arkalgud (2008) " Logistic Regression for Spam Filtering"

[4]. Rebecca Lieb (2002), "Make Spammers Ply Before You Do". The ClickZ Network.

[5]. "Q1 2010 Internet Threats Trend Report". Commtouch Software Ltd.

[6]. Josh Halliday (2011). "Email spam level bounces back after record low".

[7]. 2010 MAAWG Email security Awareness and usage Report, Messing Anti-Abuse working Groups/IPOs Public Affairs.

[8]. Jef Poskanzer (2006) "Mail Filtering". ACME Laboratories.

[9]. A " Sophos Press Relase" (2008)

[10]. Staff (18 November 2004). "Bill Gates most spammed person" , BBC News.

[11]. Sahami, M.Dumais, S., Heckerman, D., and Horvitz, E. (1998), A Bayesian Approach to Filtering Junk Email. In learning for Text Classification – Papers from AAAI workshop,pp.55-62. Madision Wisconsin. AAAI Technical Report WS-98-05,1998.

Appendix- Vb code for Tokenization

```
Private Sub tokenize()
t1 = content                                  ‘ Reads the content of a message from the database’
L = Len(t1)                                   " LENGTH OF THE TEXT"
t = " " + t1 + " "
```

```
dim u, j, k as integer
u = 1
j = 1
k = 1                                         ' indicator for the number of words found
Dim w(500) As String
Dim d(500) As String
Dim fre(500) As Integer
Do While j <= L
'Inner loop
    i = 1
    ww = Left(t, i)                           " READ FIRST CHARACTER"
    If ww <> "" Then
        u = u + 1
        k = k + 1
    End If
    ww1 = Mid(t, i + 1, 1)                    " READ SECOND CHARACTER"
    Do While i <= L And ww1 <> " "
     ww1 = Mid(t, i + 1, 1)
     If ww1 = " " Or ww1 = "." Or ww1 = "," Or ww1 = "!" Or ww1 = "?" Then
         w(u) = ww
      Else
         ww = ww + ww1
         w(u) = ww
     End If
     i = i + 1
    Loop
t = Mid(t, Len(ww) + 1)
j = j + 1
Loop
' sorting the words
For i = 1 To k
 For j = 1 To k
    If w(j) <= w(j + 1) Then
       temp = w(j)
       w(j) = w(j + 1)
       w(j + 1) = temp
    End If
  Next j
Next i

c = 0
For i = 0 To k
If w(i) <> "" Then
    If w(i) <> w(i + 1) Then
        c = c + 1
        d(c) = w(i)
    End If
End If
Next i

h = 0
For j = 1 To c
   fre(j) = 0
      For i = 1 To k
        If d(j) = w(i) Then
         fre(j) = fre(j) + 1
      End If
     Next
  If fre(j) >= h Then
  h1 = j
  h = fre(j)
  End If
Next
'Print "Number of distinct words = ", c
Print
'Dim tot As Integer
```

```
tot = 0
'Print "Frequency table" ' finding distinct words

db1.Execute "delete * from SPAMS"
For i = c To 1 Step -1
rs2.Open "select * from SPAMS ", db1, adOpenStatic, adLockOptimistic
    rs2.AddNew
    rs2!SNO = MN
    rs2!token = d(i)
    rs2!inspam = fre(i)
    rs2.Update
 rs2.Close
 Next
Call corpus1  'this function creates a new table enron3 by adding new tokens and their counts

End Sub
```