Available online at www.elixirpublishers.com (Elixir International Journal)

Computer Engineering

Elixir Comp. Engg. 82 (2015) 32469-32473

ISSN: 2229-712X

Designing Spam Filtering by Analyzing User and Email Behaviour

Faisal Bin Al Abid
Department of CSE, International Islamic University Chittagong.

ARTICLE INFO

Article history:
Received: 5 February 2014;
Received in revised form:
30 April 2015;
Accepted: 8 May 2015;

Keywords
Email,
Ham,
Spam,
False positive,
Time,
Accuracy.

ABSTRACT
Electronic Mail is the "killer network application". It is ubiquitous and pervasive. This paper presents an implemented framework for data mining behavior models from email data. There are different methods for detection of spam through email. The main goal is to develop a method that outperforms the existing methods in terms of detection of spam, ham and wrongly classified spam, that is, the accuracy of the proposed method is higher compared to the other existing methods. The other goal is to implement the proposed algorithm for reducing the time. So, to recapitulate, this paper deals with the accuracy and process timing based on prioritization of detecting email messages.

Introduction

Electronic mail is one of the most popular forms of communications today. The surprisingly fast acceptance of this communication medium is best exemplified by the sheer number of current users, estimated to be as close to three quarters of a billion individuals, and growing [1]. This form of communication has the simple advantage of being almost instantaneous, intuitive to use, and costing virtually nothing per message. The current email system is based on the SMTP protocol RFC 821 and 822 developed in 1982 and extended in RFC 2821 in 2001[2]. This system defines a common standard to unite the different messaging protocols in existence prior to 1982. It allowed users the ability to exchange messages with one another using a system based on the SMTP protocol and email addresses. These protocols allowed messages to pass from one user to another, making it practical and easy for different users to communicate independent of the service-provider or the client application. In 1982, Denning [3] wrote about the problem of working with email, asking" Who will save the receivers from drowning in the rising tide of information so generated?"

Emails for the most part are held in data files or folders with no structured relationship (at files), making anything more than a keyword search very slow. Users may choose to move messages into time-ordered sub-folders of related messages. Finding a particular past message across these sub-folders can easily turn into a daunting task. Not only is the email the subject of search, but also the folder in which it might have been placed. Within these at file folders, attachments are encoded in MIME format making analysis of anything other than simple filename close to impossible. Recent tools have been released which allow indexing and searching local data including emails and parts of attachments. Above and beyond simply sending messages, studies have shown that many users have quickly adopted email to a variety of tasks including task delegation, document archiving, personal contact list, and reminder and scheduling [4]. In addition to these organization issues, the Achilles heel of the current email system is its relative ease of abuse. The protocols were based on the assumption that email users would not abuse the privilege of sending messages to each other. The misuse and abuse of the email system has taken on many forms over the years. Typical misuse includes forged emails, unwanted emails (spam), fraudulent schemes, and identity theft and fraud through "Phishing" emails. Abuse includes virus and worm attachments, and email DOS attacks. The common denominator among all these categories is they exploit the email system's lack of controls and authentication of sender and recipient.

Review literature of Spam filtering method

Spam filtering methods

We will discuss about the various email classifications of the existing methodologies. The main methodologies used for spam filtering are Bayesian spam filtering, improved Bayesian filtering, A Naive Bayes classifier, Meta spam filtering, and Greylist. We will discuss about these methodologies in the next section.

Bayesian spam filtering

The first known mail-filtering program to use a Bayes classifier was Jason Rennie'sifile program, released in 1996. The program was used to sort mail into folders. The first scholarly publication on Bayesian spam filtering was by Sahami et al. in 1998[5]. That work was soon thereafter deployed in commercial spam filters. However, in 2002, Paul Graham was able to greatly improve the false positive rate, so that it could be used on its own as a single spam filter. It is known as statistical spam filtering method.

Process of Bayesian Spam filtering

Particular words have particular probabilities of occurring in spam email and in legitimate email. The filter doesn't know these probabilities in advance, and must first be trained so it can build them up. To train the filter, the user must manually indicate whether a new email is spam or not. Each word in the email contributes to the email's spam probability, or only the

Tele:
E-mail addresses: faisaliut42@yahoo.com

most interesting words. This contribution is called the posterior probability and is computed using Bayes' theorem. If the spam probability of the number of words exceeds more than 95%, the email is considered to be a spam. The Bayes theorem is used in various context of spam, at first taking into consideration that a single word appears in a spam, secondly, taking into consideration that all of the words in email , thirdly taking into consideration of very uncommon words in email.

Advantages of Bayesian spam filtering

1. Can be trained on a per-user basis. It will eventually assign a higher probability based on the user's specific patterns.
2. Bayesian spam filtering accuracy after training is often superior to pre-defined rules.
3. It can perform particularly well in avoiding false positives, that is, wrongly identified spam as it takes into consideration all the words used in email.

Disadvantage of Bayesian spam filtering

1. Bayesian spam filtering may be susceptible to Bayesian poisoning that uses legitimate words in spam email.
2. Words that normally appear in large extent in spam may also be transformed by spammers.
3. The email may contain a link or picture that contains the illegitimate words.

Improved Bayesian spam filtering

Final decision is made based on the weighted score of the attributes of both attitude analysis phase and relevancy analysis phase. The attitude analysis holds 0.5 weight ages for both e-mail id and subject trusted. Similarly, relevancy analysis phase holds 0.5 weight age for relevant content. If the weighted value is greater than 0.5 then the email is moved to Inbox and the pre-processed root words which are not already exist are added to positive dictionary. If the weighted value is less than 0.5 then the email is moved to spam and the pre-processed root words which are not already exist are added to negative dictionary. If the weighted value is equal to 0.5 then the e-mail is hold. The number of normal e-mail that are classified as spam and the reverse will be significantly trim down since there are a two levels of validating a e-mail in the system. Also user can classify spam and ham e-mail according to his personal interest on a particular e-mail rather than going for a generalized spam filter.

Assumed Ham classified as *C0*, Spam classified as *C1*, decision-making text messages as legitimate risk conditions,

$R(HAM|D)=P(C_1|D)$ ……(1)

$R(SPAM|D)=1-P(C_1|D)$…….....(2)

After calculating a probability the e-mail is spam, one need to compare with the critical value to determine whether it is a spam.

Suppose *D* is spam e-mail the probability of $P(C1|D)$, the probability of the normal messages $P(C0|D)=1-P(C1|D)$. Threshold in two forms: [6]

a. Set the critical probability *t*, if $P(C1|D) > t$, then that e-mail is spam;

b. Set the critical ratio *k*, if the $(P(C1|D) / P(C0|D)) > k$, then that e-mail is spam

It is easy to get the relationship between *t* and *k* is:

$t = \frac{k}{1+k}$ ……(3)

$k = \frac{t}{1-t}$ ……(4)

Therefore the text D decision-making of risk as spam R (SPAM | D) =k (1- (P (C1 | D))

Advantages of improved Bayesian spam filtering

1. The risk of loss factor of *k* thatis weight factor of the ham emails recognized wrongly are reduced.

Disadvantages of improved Bayesian spam filtering

1. Dependency on threshold value, if the threshold value is not chosen properly, then the Ham and spam are not detected correctly.

The Naive Bayes probabilistic model

Abstractly, the probability model for a classifier is a conditional model p (C|F1, …,Fn) over a dependent class variable C with a small number of outcomes or classes, conditional on several feature variables F1 through Fn . The problem is that if the number of features n is large or when a feature can take on a large number of values, then basing such a model on probability tables is infeasible. We therefore reformulate the model to make it more tractable.

Using Bayes' theorem, we write

$p(C|F_1,\dots,F_n) = \frac{p(C)p(F_1,\dots,F_n|C)}{p(F_1,\dots,F_n)}$ …….. … (5)

In plain English the above equation can be written as

$posterior = \frac{\text{prior} \times \text{likelihood}}{\text{evidence}}$ …… (6)

In practice we are only interested in the numerator of that fraction, since the denominator does not depend on C and the values of the features Fi are given, so that the denominator is effectively constant. The numerator is equivalent to the joint probability model p(C, Fi, …,Fn) which can be rewritten as follows, using the chain rule for repeated applications of the definition of conditional probability:

p(C, F1, … , Fn)

α p(C) p(F1, … , Fn|C)

α p(C) p(F1|C) p(F2, … , Fn|C,F1)

α p(C) p(F1|C) p(F2|C, F1) p(F3, … , Fn|C, F1, F2)

α p(C) p(F1|C) p(F2|C, F1) p(F3|C, F1, F2) p(F4, … , Fn|C, F1, F2, F3)

α p(C) p(F1|C) p(F2|C, F1) p(F3|C, F1, F2) …… p(Fn|C, F1, F2, F3, … , Fn-1) …. ……..…….. (7)

Now the "naive" conditional independence assumptions come into play: assume that each feature Fi is conditionally independent of every other feature Fj for $j \neq i$. This means that p (Fi|C, Fj) = p(Fi|C) for $i \neq j$, and so the joint model can be expressed as

$p(C,F_1,\dots,F_n) \alpha p(C) p(F_1|C) p(F_2|C) p(F_3|C) \dots \alpha p(C) \prod_{i=1}^{n} p(F_i|C) \dots$ …… (8)

This means that under the above independence assumptions, the conditional distribution over the class variable *C*can be expressed like this:

$p(C|F_1,\dots,F_n) = \frac{1}{Z} p(C) \prod_{i=1}^{n} p(F_i|C)$ …… (9)

Where, *Z*(the evidence) is a leveling factor dependent only on F_1, …,F_n, , i.e., a constant if the values of the feature variables are known.

Advantages of Naive Bayes probabilistic model

1. Easy to implement.
2. Requires a small amount of training data in order to estimate the parameters.
3. Good results are obtained in most of the cases.

Disadvantages of Naive Bayes probabilistic model

1. In this method, there is class conditional independence; therefore, it provides loss of accuracy.
2. Practical dependencies among the variables cannot be modeled correctly.

Meta spam filtering technique

Given the significance of the spam blight and the competitive nature of the spam-blocking vendor landscape, most organizations are diligently evaluating suppliers, and in many cases bringing in products for hands-on testing. In addition, many trade publications are doing on-site bake-offs to determine the effectiveness of various solutions, including on-premises

software, appliances, and managed services. In some cases, the testing methodology is flawed, and the results do not represent the actual effectiveness of the product or service. The root cause of the invalid testing is that testers typically take a corpus of mail and forward it to the spam-blocking service or product. In such cases, because of the message forwarding, the vendor is not unable to perform a series of sender IP validation tests, nor is it able to glean intelligence from the SMTP setup. In some cases, these real-time tests can contribute up to 20% of spam being blocked.

Grey list Spam filtering technique

A relatively new spam-filtering technique, greylists take advantage of the fact that many spammers only attempt to send a batch of junk mail once [7].Under the grey list system, the receiving mail server initially rejects messages from unknown users and sends a failure message to the originating server. If the mail server attempts to send the message a second time — a step most legitimate servers will take - the greylist assumes the message is not spam and lets it proceed to the recipient's inbox. At this point, the grey list filter will add the recipient's email or IP address to a list of allowed senders.

Advantages of Grey list

1. From user's end, Grey listing requires no additional configuration
2. From a mail administrator's point of view the benefit is twofold , first it takes minimal configuration to get up and running with occasional modifications of any local white lists. Secondly, benefit is that rejecting email with a temporary 451 error (actual error code is implementation dependent).
3. Grey listing is particularly effective in many cases at weeding out miss configured message transfer agents.
4. Some grey listing packages support a SQL backend which allows for a distributed multiple-server frontend to be organized with the same grey listing data on all frontends.

Disadvantages of Grey list

1. The biggest disadvantage of grey listing is that for unrecognized servers, it destroys the near-instantaneous nature of email that users have come to expect.
2.Send mail, one of (if not the most) prolific internet message transport agent has a default retry interval of 15 minutes and the biggest delays from grey listing systems are incurred when communicating with poorly configured sending systems with retry intervals left set at several hours or more.
3. When a mail server is Grey listed, the duration of time between the initial delay and the re-transmission is variable.
4. Grey listing delays much of the mail from non-white listed mail servers - not just spam - until typical patterns of communication are recorded by the grey listing system.
5. Grey listing can be a particular annoyance with websites that require an account to be created and the email address confirmed before they can be used.

Proposed method

A pictorial representation of the proposed spam filtering method is given in figure 1. This figure depicts the proposed spam filtering model. This specifically demonstrate the updating method of the spam filtering process based on white listed and black listed region, analyzing pre-filtering based on sender behavior, spam filtering based on email message body, and post filtering based on receiver behavior. We use four separate spam filtering engine to connect with the central knowledge base. It will also increase the performance of the email server and will reduce the process time. The data are stored in the knowledge base that it uses support and confidence rule in order to find out the spam and ham emails. The confidence rule used as {Upper and lower case letters in email subjects, length of the subject is between 70 to 80 characters}⇒ {spam email} has a confidence of 0.7 which is heuristically found. This means that if these two criteria are included, there is 70% likeliness for the email to be a spam. The Apriori algorithm is used for this purpose containing two steps such as finding all frequent item sets, and then using frequent item sets to generate rule.

The post filtering method is based on the detection of spam and ham on the choice of receivers. Let single malicious email has come to 100 receivers. Among them, 60 receivers considered that email to be malicious whereas (100-60)= 40 of the receivers do not take any action. In this case, the email is tagged as spam by the email server and that email will go to the receivers as spam in future by adding the email address to be black listed. Here the post filtering can be tagged as PF. The PF is detected on the ratio of the total number of users considering the emails to be spam divided by the total number of receivers receiving the same email. If PF is greater than 0.5, then the email is considered to be malicious for the rest of the users in future. The vice versa case applies in the case of spam email where the spam is considered as ham for the receivers. So, Post Filtering,

$$PF = \frac{N_A(S_A|H_A)}{N}$$

Where, N_A=Number of receivers taking action

N= Total number of receivers from a single sender email

S_A= Action taken to include the email in spam inbox

$H_{A=}$ Action taken to include the email in ham inbox

Thus post filtering can be used in order to detect the spam and improve the accuracy for the proposed spam filtering method.

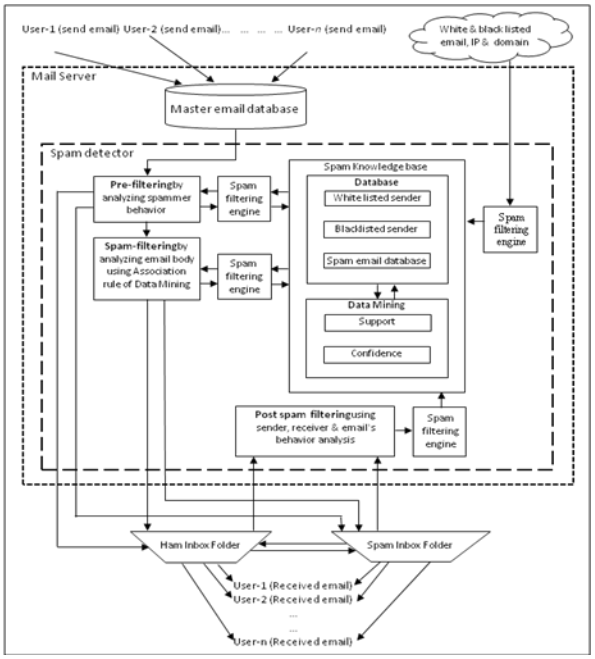

Figure 1: Proposed Spam filtering method

Algorithm of proposed method

The algorithm for the spam filtering method is subcategorized into the following subsections: process prioritization and post filtering method.

Subalgorithm: Process Prioritization

The algorithm of the process prioritization that is responsible for reducing the process time from others:

SetUTYPE = Process update sequence type in database

Set SYSTIME = Current System Time, UTIME = Auto update time in database

If UTYPE = Manual then

Input sequence for each process
Update process priority database
Else
If SYSTIME = UTIME then
[Load process list, current priority, total spam detection]
PROCESS<- All process
PSEQ<- Process sequences
PSPAM<- No of spam detection after last sequence update by the processes
WHILE N = 0 to *COUNT*(*PROCESS*)
INDX = index of *MAX*(*PSPAM*)
Set PSPAM[*INDX*] = 0 [Spam count reset]
Set PSEQ[*N*] = *INDX*
END WHILE
Update priority database by the array PSEQ
End if
End if

Sub algorithm: Post Filtering Method

The algorithm of the post filtering method that will improve the method is as below:

Procedure Spam_By_Post_Filtering (EMAIL, EMAILTYPE, SENDER, RECEIVER)
If EMAILTYPE = HAM then
R_ACTION = Get Receiver's Response
If R_ACTION =1 then [1: Receiver Marked as SPAM, 0: No Action by receiver]
Move EMAIL to SPAM inbox
Add SENDER address to BLACK_LIST for RECEIVER
RCOUNT = Number of receivers of the EMAIL
MOVECOUNT = Number of receivers marked EMAIL as SPAM

*If (MOVECOUNT*100)/ RCOUNT >50 Then*
Add SENDER address to BLACK_LIST for all receivers under this email server
End If
Else
COUNT = Count EMAIL in HAM inbox
If COUNT=3 then
Add SENDER address to WHITE _LIST for RECEIVER
Else
End if
Else
R_ACTION = Get Receiver's Response
If R_ACTION =1 then [1: Receiver Marked as HAM, 0: No Action by receiver]
Move EMAIL to HAM inbox
Add SENDER address to WHITE_LIST for RECEIVER
RCOUNT = Number of receivers of the EMAIL
MOVECOUNT = Number of receivers marked EMAIL as HAM
*If (MOVECOUNT*100)/ RCOUNT >50 Then*
Add SENDER address to WHITE _LIST for all receivers under this email server
End If
Else
COUNT = Count EMAIL in SPAM inbox
If COUNT=3 then
Add SENDER address to BLACK_LIST for RECEIVER
Else
End if
End if
End Procedure

Performance analysis

In this section the comparison between the proposed and existing methods are shown. Here we will see the performance analysis among the existing and the proposed method using a large number data set. Also the comparison using the same data set among presently used well known software and the proposed method has been made in this section.

Comparison with the existing methods

The outcome of the proposed method is compared with the existing Bayesian and Naïve Bayesian approach and the following result was found. The accuracy is computed based on 70,053 emails.

Table 1: Performance analysis between existing and proposed method

Features	Bayseian spam filter	Improved Bayseian approach	Naïve Bayseian approach	Meta spam filter	Greylist approach	Proposed method
Spam detected accuracy	98.00%	99.10%	97.30%	98.60%	96.00%	99.92%
False positive	1.16%	0.46%	1.20%	1.63%	3.50%	0.10%

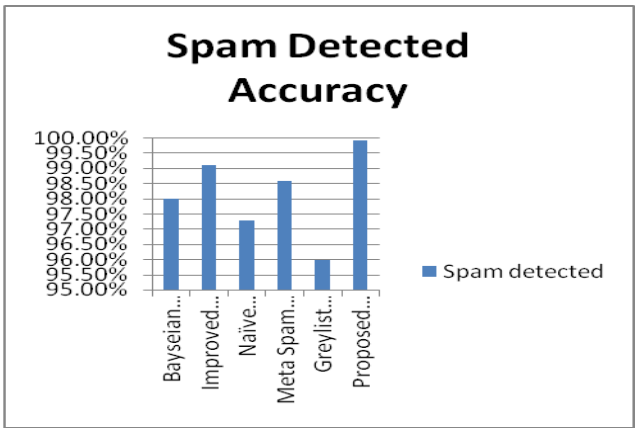

Figure 2: Performance analysis of spam detection

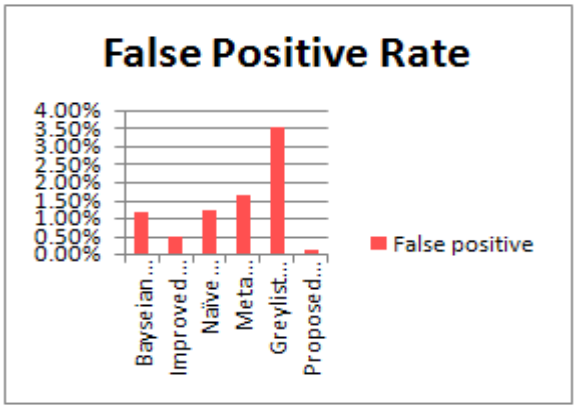

Figure 3: False positive rate using common data set

Comparison with the existing software

The proposed method is named as MAN method. The outcome of the proposed method is compared with the current version of Windows Live Mail 2011 (Build 15.4.3555.0308) & Gmail and following result was found. The accuracy is computed based on 8000 same data set.

Observation from the output

From all the figures above, it is seen that the proposed spam filtering method works better and overwhelms the performance

of the existing method. The features and the parameters used in order to detect the performance analysis of the methods are spam detected, hams classified and the false positive, i.e wrongly detected spam by the method. It is observed that the spam detected by the proposed method is higher and performs better than the existing one. The checking mechanism of detecting the spam is carried out through 10,000 to 70,000 email messages and the proposed method was able to detect almost 99.92% of the spam. It is also observed that, the proposed method detects hams correctly, finds out the spam detected and has almost zero false positive (wrong detection of spam) which indicates the authority of the method over the other four spam detection.

So, to recapitulate, it can be said that the proposed method is able to detect spam better and able to provide user comfort.

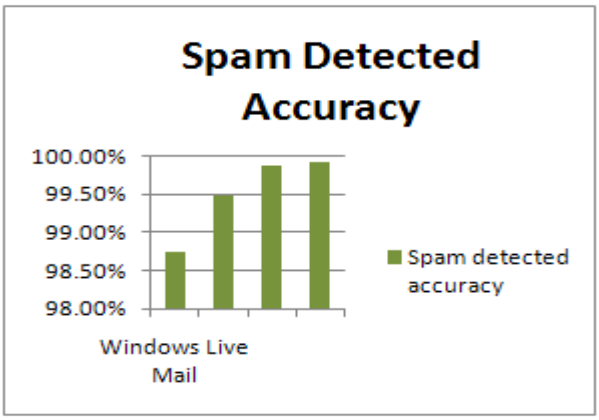

Figure 4: MAN compared to other existing software

Table 2: Performance analysis among the existing and implemented software using common data set

Features	Windows Live Mail	Gmail	Proposed method	
			Before Post Filtering	After Post Filtering
Spam detected accuracy	98.75%	99.47%	99.87%	99.92%
False positive rate	2.5%	0.26%	0.46%	0.1%

Conclusion and future work

Email has become parts and parcel of our everyday life. Making it efficient saves significant amount of time from each of our lives. Due to it critical role in saving our time we selected the topic and came out with the idea of introducing the proposed method. We have successfully demonstrated the better capability of proposed method in comparison to other methods. The concept of sender authentication with confidentiality, availability and integrity can be added to ensure the security to the receiver. Moreover, an appropriate algorithm can be used for this purpose. The knowledge base can be used to derive the age, gender, preference, area of the receiver. Based on the age, gender, preference, area of the receiver, clustering can be used in order to find out the emails that are considered to be valid to the same age group, gender, preference and area of the receiver.

Reference

[1] *Radicati Sara, "Email Statistics Report, 2009-2013", The Radicati Group, Inc ., 2009*

[2] *Klensin J., "Technical Report RFC 2821, IETF, Simple mail transfer protocol", Network Working Group, October 2008*

[3] *Denning, P. J. "Electronic junk. Communication of the ACM", Purdue University, India, 1982, 25(3):163–165.*

[4] Ducheneaut, N. and Bellotti, V. "E-mail as habitat: an exploration of embedded personal information management. Interactions", 2001, 8(5):30–38.

[5] *Sangeetha C., Amudha P., Dr. Sivakumari S., "Feature Extraction Approach For Spam Filtering", 2012, ISSN NO: 6602 3127, IJART, Vol. 2 Issue 3, pp 89-93.*

[6] Hu Yin, Zhang Chaoyang, Hubei, China, on "An improved Bayesian Algorithm for Filtering Spam E-mail", Network Center Huanggang Normal University Huangzhou, International Symposium on Intelligence Information Processing and Trusted Computing, IEEE, 2011.

[7] M. Kucherawy, D. Crocker, Brandenburg Internet Working, Internet Engineering Task Force (IETF), ISSN: 2070-1721, June 2012