International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 3 Issue 5, May - 2014

A Hybrid Spam Filtering Technique Using Bayesian Spam Filters and Artificial Immunity Spam Filters

Smera Rockey
P.G Scholar
SCMS School of Technology and Management
Cochin, Kerala, India

Rekha Sunny T
Assistant Professor
SCMS School of Technology and Management
Cochin, Kerala, India

***Abstract*— Spam emails are one of the crucial problems faced by most of the email users. There are a lot of algorithms to filter spam mails from ham mails. In this paper two efficient filters-Bayesian filters and Artificial Immunity filters are studied and compared. Bayesian classifier has been suggested as an effective method to construct anti-spam filters for its strong categorization and high precision. Artificial immune system has become a new embranchment in computing intelligence for its good self-learning, self-adaptability and robustness. This paper proposes a new hybrid filtering system based on Bayesian, AIS and the concept of blacklist and whitelist .**

***Keywords*— Artificial Immune System, Bayesian Algorithm, Gene Library, Spam mails, Antigens**

I. INTRODUCTION

Today millions of people use emails to communicate around the globe and is a critical application for many businesses. These emails provide a perfect way to send millions of advertisement at no cost. So as a result the mailboxes of millions of people were cluttered with these unsolicited bulk email which is also known as spam or junk email. Emails being cheap to send it cause a lot of trouble to the internet community. It not only frustrates email user , it strains infrastructure of a organization but large amount of spam traffic between servers delays legitimate email and bandwidth was send to download these unnecessary mails. There is also pornographic spam that should not be exposed to children. If there are no anti spam activities spam will flood the network system, kill the employee productivity and steal the bandwidth.

An electronic message is a spam if it meets the following criteria.

1.)Anonymity The identity and address of the sender are concealed.

2.)Mass Mailing: If the email is send to the large group of people.

*3.)Unsolicited :*The email is not requested by the recipients .

In short spam email is any mail that was sent to the user and many other (not always) with malicious intent. The source and identity of the sender is anonymous and there is no option to cease receiving future emails.

Most of us can easily identify spam messages in our inbox and ignore it. . But there are individuals who will respond to spam messages. Even if only a small fraction receives respond to email spam out of millions of users/receivers, it will result into significant turnover for the spammer. So more the spam mails a spammer send, the more likely he or she is to get recipient to respond to the mail. And the cost in money and time to send spam emails is quite low even if the sender uses his or her email server or proxy server.

The entire paper is divided into five sections. Section II provides the literature survey. In III comparison of the AIS and Bayesian filter is done . The hybrid system is explained in section IV followed by conclusion in section V .

II. LITERATURE SURVEY

A. *Spam Filtering*

There are many ways of fighting spam. [1]There are social methods like legal measures and by plain personal involvement. The example of former is anti spam laws introduced in US and of latter is that never respond to spam , never publish your email on web pages etc. There are technological ways like blocking spammer IP address and at last there is email filtering. Automatic email filtering is one of the most effective methods for countering spam. There is a tight competition between the spammer and spam filtering technique. The two general approaches to mail filtering are Knowledge engineering and Machine learning
In knowledge engineering set of rules is created according to which messages are categorized as spam or legitimate email. These rules are created either by the cross of the filter or by some authority. The main drawback of this method is that we have to constantly update these set of rules and maintaining it is not convenient for most of the users. But the maintainer of spam filtering tool can update the rules in centralized manner and there is even a peer to peer knowledge base solution .When these rules are publically available spammer can adjust his message accordingly so that it would pass through the filter. The machine language approach does not require specifying any rules explicitly. A set of classified document is used .Then we use a specific algorithm to "learn" the classification rules from this data. There are lot of algorithm suitable for this work.

B. *Bayesian Filters*

[4]Bayesian filters are the most advanced form of content based filtering. Content based filter determines whether an email is spam or legitimate by evaluating words or phrases found individual message. It employees the laws of mathematical probability to determine which message are legitimate and which are spam. In this, particular words will have particular probability of occurring in spam mail and legitimate mail. These probabilities are not known in advance. The end user must initially "train" it by manually flagging each message either as a junk or legitimate. Then it generates database with words or phrases and tokens collected from these samples of spam email and valid email. Then the filter scans the contents of the email and then compares the cost against its two word lists to calculate that the message is spam. Bayesians filter builds its word list based on the messages that an individual user receives. So it becomes more effective the longer it is used.

We will compute the probability that a message containing a particular word is spam.

Let us consider a particular message with a particular word say ‘x’. The person who receives the mail will know that this message could be a spam. Spam filter don’t know to identify whether the message is spam or not without computing probabilities.
Let S be the event that the message is spam
Let W be the event that we have a particular word
Let H be the event that the mail is legitimate mail.

$$P\ (S|W) = \frac{P(W|S)*P(S)}{P(W|S)*P(S)+P(W|H)*P(S)} \tag{1}$$

P (S|W): Probability that message is spam given that word is say “x”
P (W|S): Probability that the word “x” appears in the spam message
P (W|L): Probability that the word appears in the legitimate message
P(S): overall probability that any given message is a spam
P (L): Probability that the given message is not spam
There is no a-priori reason for any incoming message to be spam rather than legitimate mail which have equal probability of P(S)=P(L)=0.5
The filter that use this hypothesis is said to be not biased so the formula can be simplified as

$$P\ (S|W) = \frac{P(W|S)}{P(W|S)+P(W|H)} \tag{2}$$

This calculates the percentage of occurrence of a particular word in the spam message. This quantity is called spamicity. P (W|S) gives the frequency of messages containing the word "x” that is contained in message identified as spam. Similarly P (W|H) gives the frequency of message containing the word that is contained in the message identified as legitimate.

Most of the Bayesian formula is strictly based on formulas that are strictly valid only if the words present in the message are independent events. Usually statistical correlation between individual words is not known. On this basis ,

$$p = \frac{p1.p2....pn}{p1.p2...pn+(1-p1)(1-p2)..(1-pn)} \tag{3}$$

p1is the probability that the mail is spam given word1 P2 is the probability that the mail is spam given the second word and so on. Spam filters based on this formula are called naive baye’s formula. After computing the probability based on some threshold value of p classification is done on the basis whether the mail is spam or not.

C. *Artificial Immunity based Spam filtering*

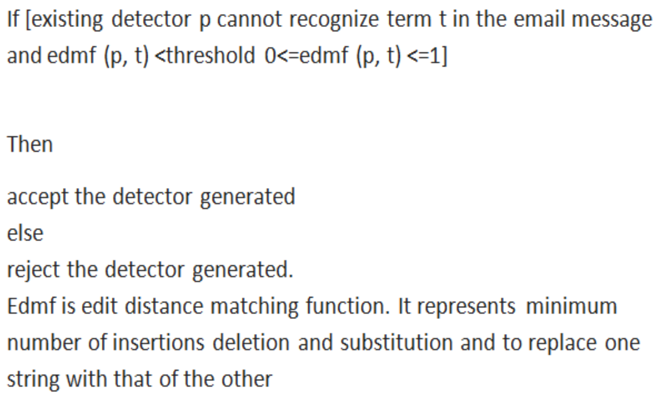
[6]The vertebrate immune system is one of the intelligent biological systems. It distinguishes harmless from harmful substances such as pathogens which find new mechanisms to attack the body and its immune system. Similarly spammers change one or more characters of offensive words is their spam in order to foil the content based filters. So the problems of spam and immunity are conceptually similar so to detect the spam artificial immunity system is used. Diverse responses and theories of natural immune system inspired some of the AIS models.

1) *The immune System*: [6]The immune system distinguishes self and potentially harmful non self elements. These harmful non self elements are called pathogens. We have antigens which are cell surface and soluble proteins called antigens. These antigens are used to identify the pathogens. The acquired immune system comprises mainly lymphocites.These are white blood cells that detect and destroy pathogens by binding them. After binding, immune system can respond to new pathogens as well as the pathogens that are similar to those already encountered. Lymphocytes perform the recognition. The overall immune response involves three evolutionary methods. Gene library evolution generating effective antibodies, negative selection eliminating inappropriate antibodies and cloned selection cloning well performed antibodies. In this article we see how the existing system implements the gene library evolution method and compare with Bayesian filter. In gene library evolution antibodies recognizes antigens by the properties that belong only to the antigens. So we require the properties of antigen to generate competent antibodies. So here gene libraries acts as information source on commonly observed antigens. Comparing immune system with spam filters spams can be considered as a pathogen which tries to attack the body or the system. Initially we should extract message features from a training set of spam and non spam Messages. Then these features are used as input to the classification model to take decision. We have a collection of detectors that is created from genes and combined randomly into new detectors. These gene libraries can be taken from variety of sources including words or phrases in one or more language.

The detectors generated from this gene library are trained with messages and any detector with matches is collected. Some of detectors will have receptors that can bind to more than one email through regular expression.

2) *Feature Extraction and Learning features:* [6]Initially, initial set of detectors are generated from the training set of emails. We use mime parsers to parse the body and the subject part. These parts are identified based on some delimiters’ like whitespace colon etc. The remaining terms are taken as base string. This base string will be used to generate detectors. We combine regular expression quantifiers and character

generalization rules to produce detectors. The generated detector is combined against existing detectors and matched detector is eliminated. The basic rule for generate detector are :



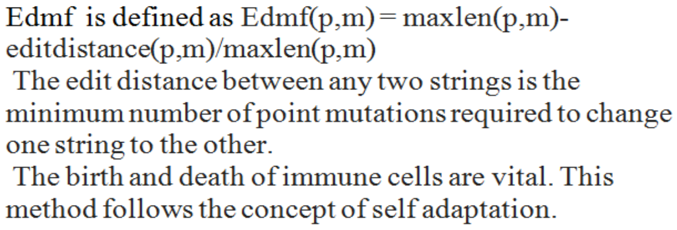
If [existing detector p cannot recognize term t in the email message and edmf (p, t) <threshold 0<=edmf (p, t) <=1]

Then

accept the detector generated
else
reject the detector generated.
Edmf is edit distance matching function. It represents minimum number of insertions deletion and substitution and to replace one string with that of the other

Figure 1 Basic rule for detector generator

3) *Gene Library:* A collection of base strings and a collection of character generalization rule. The base string is a word that is found in the spam message. The fitness value f rate of this string is greater than a threshold value . The format for character generalization rule is character~[a,à,á,@]. The detector is a regular expression. This is formed by replacing each character in the base string by right side of the character generalization rule. Regular expression quantifiers are also inserted to base string. Initially these base strings are empty. They are dynamically updated with each and every message processed. The next step is to learn these features. Consider srate, nsrate and arate. A detector that is matched with terms in spam messages has its spam term frequency of its corresponding base strings increased by the spam rate. And those matched with terms in legitimate messages have their corresponding non spam term frequency increased by non spam rate. In addition each base string has its age decreased by the aging rate for every message processed. The detector and a string from an email are matched by edit distance minimum function.



Edmf is defined as Edmf(p,m) = maxlen(p,m)-editdistance(p,m)/maxlen(p,m)
The edit distance between any two strings is the minimum number of point mutations required to change one string to the other.
The birth and death of immune cells are vital. This method follows the concept of self adaptation.

Figure2 Edit distance funtion

III. COMPARISON OF BAYESIAN AND AIS FILTERS

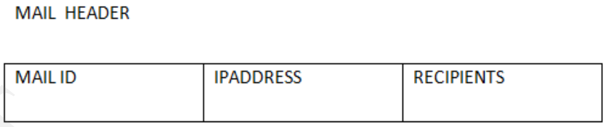
In baye's algorithm the ability to detect the spam drift is very weak. They rely training features and rule. The AIS model can easily identify the spam drifts.Where as in spam immune system it has the ability to learn and unlearn things. It can cop up with the fact that self and non self i.e. spam n legitimate messages change all the time. This system takes only less time to train themselves because the individual genes can be taken from variety of sources like a list of words already chosen by using trained or partially trained .

TABLE I
COMPARISON OF AIS AND BAYESIAN FILTERS

Filters	**Comparison of Bayesian and AIS Filters**		
	Algorithm	*Spam Drifts*	*Time to train*
AIS Filter	Flexible	Have ability to detect	Easily trained
Bayesian Filter	Less Flexible	Very weak in detecting drifts	Take time to train

IV. A HYBRID FILTERING SYSTEM

In this system there are two phases. Divide entire mail into two parts as Mail header and mail body before stepping onto these phases .The mail header contains following parts.Mail address of the sender , ip address of the sender and total number of recipients.

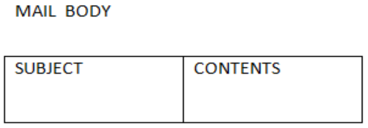


MAIL HEADER

MAIL ID	IPADDRESS	RECIPIENTS

Figure 3 Mail Header structure

The mail body can be divided into mail subject and contents.



MAIL BODY

SUBJECT	CONTENTS

Figure 4 Mail Body Structure

The entire filtering process is divided into two phases, Initial filtering phase and hybrid filtering phase.

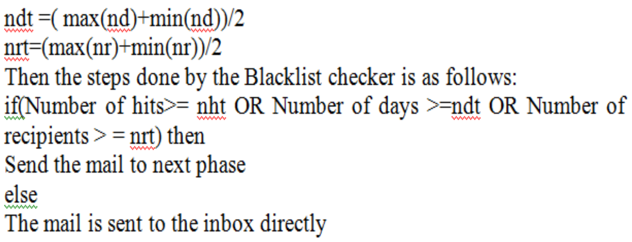
A. *Initial Phase*

In this phase , check the mail headers to identify the spamicity of the entire mail and in hybrid filtering we does the actual filtering of the mail contents using AIS and Bayesian filters, filter the message header using blacklist, whitelist and number of recipients. This system have a blacklist and whitelist repository which contains the list of whitelist and blacklists. Blacklist and whitelist checkers do the filtering based on some rules.

1.) *WhiteList/Blacklist:* A whitelist is a list, which includes all addresses from which we always wish to receive mail. A blacklist works similarly to competitive alternatives: this is a list of addresses from which we never want to receive mail.

2.) *Whitelist checker*: Mails have mail header which contains email and IP address of the sender .Initially the mail will be passed through whitelist checker. It will check whether the sender mail id/IP address is already there in the list or not. If the mail id is already listed in the whitelist the email can be considered to be send from known person and the mail may be send to the or classified as the ham mail. If the mail doesn't belong to the whitelist then it may be send to the blacklist checker.

3) *BL checker:* The mail header is now passed through the blacklist checker. Blacklist contains list of IP addresses and email addresses which were marked as spam and such mail ids or IP addresses is said to have got spam hits or simply hits. Consider number of hits(nh) and the number of days in which the hit value for the particular mail header has not changed. Now compare the Number of hits compared with threshold value. The threshold value of nh, nht can be considered as (max hit value in repository + min hit value in the repository) divided by 2.Similarly the threshold value for number of days in which particular mail ID/IP address had been remaining same (nd) and (nr) the number of recipient is considered



```
ndt =( max(nd)+min(nd))/2
nrt=(max(nr)+min(nr))/2
Then the steps done by the Blacklist checker is as follows:
if(Number of hits>= nht OR Number of days >=ndt OR Number of
recipients > = nrt) then
Send the mail to next phase
else
The mail is sent to the inbox directly
```

Figure 5 BLChecker

B. *Hybrid Phase*

The mail passed by the Blacklist checker is processed here. Initially the mail body of the corresponding mail header is filtered by Byesian Filters.In this phase we use two subroutines Threshold generator and Comparator.

1.) *Threshold generator :* If the number of spam word is less than a particular threshold value , continue the filtering using AIS for more efficient result. If the number of spam word is greater than the threshold value , the mail will be considered as spam and no need to pass the mail through AIS filters.This is because if the number of spam word is greater means the chances for the mail to be spam is very high so we no need to pass through it through AIS filters for further checking.

The main idea behind passing the message through Bayesian filter is to count the number of spam words and the AIS will do more detailed checking which can't be done by the Bayesian, thereby getting more efficient result. The threshold value can be calculated base d on number of hits (nh) ,number of spam words (nsw), number of non spam words (nnw) and total number of words in the mail (tw).The threshold value is calculated as follows:

$\Delta S = |nnw - nsw|$
if tw is greater than nh
$thi=(\Delta S /tw)*100$
else
$th=thi/nh(mod(tw))$

This is just a algorithm to generate a random threshold value.Both th and thi can be considered as the threshold value based on the conditions.As the number of spam words can be more that number of non spam words and vice versa we use modulus to calculate ΔS. The number of hits(nh) can be too high that is greater than the number of words.So we take nh(mod(tw)) so that the number of hits can be considered in calculating the threshold value and at the same time it should be consistent with the total number of words too.

2.) *Comparator :* The comparator routine compares number of spam words with the threshold value .
If it is greater than threshold value the mail is considered to be spam and if lesser the mail is considered to be the intermeiate spam because the chances that the mail is spam can be more or less and therefore it is send to the AIS for further filtering.
AIS filters again scans the message for the purpose of advanced filtering so that it can identify whether the mail body contains spam words that was not identified by the basyesian filters.It checks for all the possible combinations of alternate words that can be used by the spammers to foil the filters.After the AIS filter scans the mail content they are send to either inbox or spam folder based on whether the mail contains more spam or less spam
The following figure 6 shows the flow chart of the entire system.

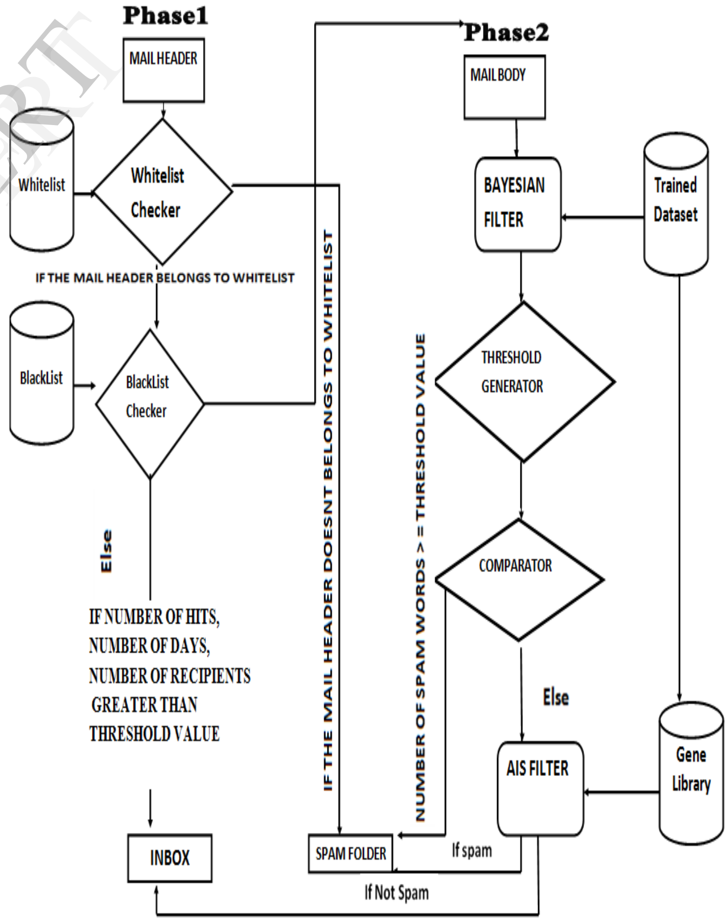

Phase1
MAILHEADER
Whitelist
Whitelist Checker
IF THE MAIL HEADER BELONGS TO WHITELIST
BlackList
BlackList Checker
Else
IF NUMBER OF HITS, NUMBER OF DAYS, NUMBER OF RECIPIENTS GREATER THAN THRESHOLD VALUE
INBOX
IF THE MAIL HEADER DOESNT BELONGS TO WHITELIST
SPAM FOLDER
NUMBER OF SPAM WORDS > = THRESHOLD VALUE
If spam
If Not Spam
Phase2
MAIL BODY
BAYESIAN FILTER
Trained Dataset
THRESHOLD GENERATOR
COMPARATOR
Else
AIS FILTER
Gene Library

Figure 6 The flow chart of Hybrid filter

V. Conclusion

When comparing AIS and BAYESIAN filters , AIS is much more flexible than Bayesian filters as it strongly accept spam drifts. Still both are equally efficient in their own way.

The algorithm is simple but comparatively it may cause delay in delivering mails. A model is used that can implement both the heuristics n AIS model. They can be combined so that the processing time can be reduced. The addresses in the white list need not be passed through these filters. It can use more than one gene library. Those genes may also contain white list and blacklists. They just need to be compared with the message headers and if a match is found with those in message header messages can be directly send to spam folder or to inbox without going through all the steps . And the members in white list and black list could also be removed from the gene library if the hits for each black list or white list is less than or greater than some threshold value. These are the possible future enhancement that could be made such that the processing time and the delay in receiving mail will be reduced.

VI.References

[1] Konstantin Tretyakov “Machine Learining technique in spam filtering” , a MTAT.03.177,DatMining problem oriented seminar ,2004

[2] Christina V,Karpagavalli S,Suganya G,” A study on email spam filters “, International journal of Computer Applications(0975-8887) 2010

[3] “Why Baesian filtering is the most effective antispamtechnology,et”,http://www.gfi.com/whitepapers/why-bayesian-filtering.pdf

[4] Robert Haskins,Rob Kolstad ,”Bayesian Spam Filtering technique”,International journal on Advanced computer engineering and communication technology Vol1-Issue1:ISSN 2278-5140 2003

[5] Graham P : “A plan for spam (2002)” ,http://www.paulgraham.com/spam.html

[6] B.Sirisanyalak and O.Sornil , “An Artificial Immunitybased Spam Detection System”,IEEE Congress on Evolutionary Computation (CEC 2007)

[7] Alaa Abu-Haidar and Luis M Rocha , “Adaptive Spam Detection inspired by Immune System”, Artificial Life XI 2008

[8] Andrej Bratko ,Gordon V Cormack ,Bogdan Fillipic ,Thomson Y Lynam ,Blaz Zuoan “Spam Filtering using statistical Data Compression models ”,Journal of Machine Learning Research 2006