Review on Effective Email Classification for Spam and Non Spam Detection on Various Machine Learning Techniques

Mr. Atul A. Jamnekar, Mr. Falesh M. Shelke
M.E. Computer Science and Engineering
P. R Pote College of Engineering and Management
Amravati, India
atuljamnekar@gmail.com
falesh123@gmail.com

Prof. Praful B. Sambhare
Prof. Department of Computer Science and Engineering
P. R Pote College of Engineering and Management
Amravati, India
Sambharepraful832@gmail.com

Abstract— Some time email receiver or user receives a email which he does not intended to receive or accept, these kind of emails are nothing but spam emails. In other words the unsolicited bulk email is nothing but the spam. Numbers of emails users are increasing day by day, email users communicate around the world using email and internet. Now days a large volumes of spam emails are causing serious problem for Internet service and Internet users. This affects or degrades user search experience, which assists propagation of virus in network or grid, this will increases load on traffic in the network. It also wastes valuable time of user, user's energy for appropriate emails among the spam emails. To avoiding such spam there are so many traditional anti spam techniques includes, rule based system, White list and DNS black holes, IP blacklist, Heuristic based filter, Bayesian based filters. All these techniques are based on links of the mail or content of the email. In this paper, we conferred our study on various existing techniques on spam detection and finding the effective, accurate, and reliable spam detection technique.

Keywords- *Email, Bayesian Filter, Spam Detection, Ham, Spam, Spam Filter*

*****

I. INTRODUCTION

Email spam is abuse of electronic messaging system to send unsolicited messages in bulk. Today Emails are used by number of user to communicate. The growth of internet and email users, there has been breathtaking growth in email spam in recent year and time. Email spam can be originated from any location across world, any place or device where internet access is available. The first spam was sent out to all users on ARPANET. Spam was created by Hornel in 1937 as the world's first canned meat that didn't need to be retreated. This was originally named Hornel Spiced Ham but was eventually changed to the catchier name, SPAM [6]. Normally the spam comes in the form of advertisement or publicity, few times even containing malicious code or explicit content. Email spam has been perceived as problem since the year 1975. According to the statistics from ITU (International Telecommunication Union), 70% to 80% of emails in the internet are spam which has become worldly problem to the information base. As to address growing email spam problem there are many anti-spamming technique.

II. SPAM

Email spam most often considered as electronic junk or junk message posting, some people define the spam as a unsolicited email. Generally email spam is nothing but the email advertisement for any product sent to a group of email users and or mailing list. Email spam is nothing but the any email that was not requested by email user but sent to user and many others, typically (not always) with malicious contents. The identity of a source and sender is anonymous and email spam receiver has no option to stop receiving spam mails in future. Now days a large volumes of spam emails are causing serious problem for the email receiver, and internet services provided by ISP (internet service provider). Like as, this will degrades user search experience through the internet, spam emails also assists in propagation of virus in network, this will increase load of the traffic in network, also wastes the resources such as storage, bandwidth and power of computation, devastation of the user time and energy. to avoid spam's it is advised that spam filter should be used by user, user should not reply the spam email, User should not post email address on web site, and Never buy anything from spam email site which link is received by email.[6][3].

III. ARCHITECTURE OF SPAM FILTERING

The spam filter minimizes the amount of junk email, which receiver does not intended to receive. Spam filtering is nothing but the processing of emails to organize it according to specified benchmark. A common use of mail filters is to organize incoming mail using filters, Elimination of spam emails, Elimination of virus from computer. The spam filter can be implemented at all layers, The Existence of firewall in front of MTA(Mail Transfer Agent) or email server, which will provide an integrated anti-spam and Anti-virus solution offering complete email protection at the network perimeter level and at email server, prior to unwanted or potentially dangerous email reaches the network. ON MDA (Mail Delivery Agent) level also spam filters can be installed as a service to all of their users. As email client user can have a personalized spam filter that can automatically filters all mail according to the selected standard [1][6]. The Fig. 1 shows the typical architecture of spam filter.

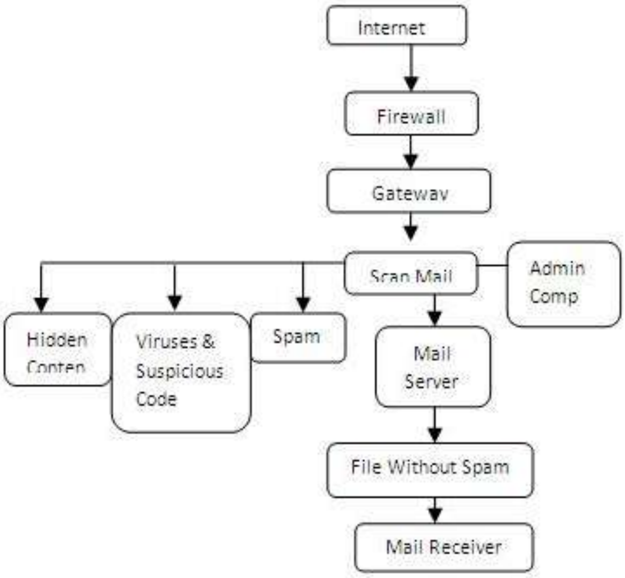

Figure 1: Spam Filtering Architecture

IV. SPAM DETECTION TECHNIQUES

Following are the desired filtering techniques

.

4.1 List Based or Rule Based Filters

The rule based or list based filter attempt to stop spam by categorizing senders as trusted users or spammers and accordingly the allowing their messages or blocking should be done.

4 .1 .1 Blacklist

The black list is the form of rule based filtering that uses one rule to decide which emails are spam and email is not spam email. Black list are the list of IP address of machine or record of email addresses that have been previously used to send spam emails. As incoming message arrives, spam filter checks to see if it's IP or email address is on the black list or not, if found, the message is considered as spam and dropped. Blacklist can be used on both large scale and small scales for spam detection [6].

Advantage of this technique is it can block substantial amount of email spams.

Disadvantage of blacklist provider is that it can block an entire net block range instead of just an individual IP address.

4 .1 .2 White list/Verification Filter

The white listing is used to decide which emails are spam and which are not, White listing is used to decide which emails are ham and assume all other emails are spam[6][7].

4 .1 .3 Black holes [7]

Black holes works with Blacklist, hand in hand. The way in which Black holes work is like someone posts message on websites or send email to group of users, forums etc, by showing their email address. These email address they use is generally a machine account that detects who sent the spam and the IP address of to a DNS Blacklist.

Advantage of white list is the email is received from one of these addresses, sending server can added to a Blacklist for stopping it from sending any more messages, which user does not wish to receive.

Disadvantage of white list is this it can't see any disadvantages to using Black holes in order to detect spam or message, they are important as they enable blacklist to be updated with computers that are sending unwanted emails.

4.1.4 Grey lists

It is relatively new spam filtering technique, which takes the advantage of the fact that many spammers only attempt to send a batch of junk mail once in time. Under the grey list system, initially receiving mail server rejects messages received from unknown users and sends a failure message to the server originating the message. If the mail server attempts to send the message second time- a step most legitimate server will take - the grey list assumes the message is not spam and let it proceed to the inbox of receiver. At this time grey list filter will add the recipient's email or address to a list of allowed senders. Grey list filter require fewer system resources than some other types of spam filters, this also delay in mail delivery, which could be inconvenient when you're expecting time sensitive messages.

4.2 Content Based Filter

Another most commonly used spam filtering technique is Content Based Filter, it is the most commonly used group of methods for spam Filtering. Content base filter act either on the content the information contained in the mail body or message, or on the mail headers like "Subjects" to either classify or specify, accept or reject a message or mail[3].

4.3 Bayesian Filter

The Bayesian filter technique is considered to be a more advanced form of Content Based Filter, which employ the laws of mathematical probability to determine which messages are legitimate and which are spam messages or mails. Bayesian filter learn from both good and spam emails, which result in an adapting and efficient anti spam approach.

4 .3 .1 Mathematical foundation

Bayesian email filter take the advantage of Bayesian theorem which is as follows.

P (spam/word) = [P (word/spam) P (spam)] / p (word)

This theorem in the context of spam filtering, says that the probability that an email is spam, on the basis of that it has contains certain words in it, which is equal to the probability of finding those certain words probability that any email is spam email, divided by the probability of finding those words in any email.

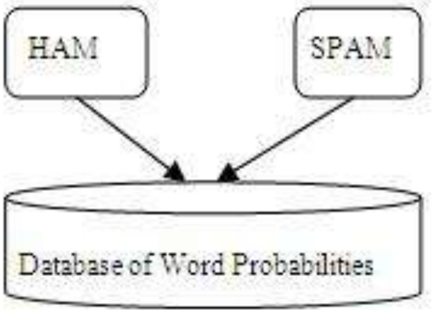

Figure 2: Creating Database for Filter

The Particular words have particular probabilities of occurring in spam email and in legitimate email. In advance this filter does not know these probabilities, so it must first be trained so it can build them. To train the filter user must manually indicate whether a new email is spam or not spam. For all words in each training email, so the filter will adjust the probabilities that each word will appear in spam or legitimate email in its database. For instance, the Bayesian spam filter will typically have learned a very high spam probability for the words "Viagra" and "refinance", but a very low spam probability for words seen only in legitimate email, such as names of family members and friends. After system training the words probabilities are used to compute the probability that an email with a particular set of words in it belonging to list. The each word in the email contributes to the email's spam probability. This contribution is called posterior probability and is computed using Bayes theorem. Then email's spam probability is computed over all words in the email or message, and if the total exceeds a certain threshold (for ex.: 95%), the filter will mark that the email will be a spam. Email marked as spam then be automatically moved to a "Junk" email folder, or even deleted outright [6].

The Bayesian filter can look the words in the body of the message or email, its header (sender and message path) and HTML code, the word pair, Phrases, and Meta information.

The advantages of this filter is it can be trained as per the user basis, the spam that user receive is often related to the online user's activities, which will assign high probability based on the user's specific patterns, the word probability is unique to each user and can evolve over time with corrective training whenever the filter incorrectly classifies an email.

The disadvantage is that they need to be trained properly in order perform the spam filtering to work a most effectively and the training leads to more time.

4.4 Collaborative Spam Filtering

The nature of spam is such that each message is typically sent to a vast number of recipients. Chances are that particular recipient is not the first to receive any particular message it is likely to have not only been received but also recognized as spam by somebody else. Collaborative spam filtering is the process of querying, capturing, and recording these early judgments [3].

V. EFFICIENT SPAM DETECTION TECHNIQUE

The techniques currently used by most anti-spam software are fixed, which mean that it is fairly easy to avoid by twisting the message little. To do so spammer simply examines the latest anti spam techniques and finds the ways how to dodge them. For effectively combat spam, an adaptive new technique is required. This method must be familiar with spammer's tactics as they change over time. It must also able to adapt to the particular organization that it is protecting for the answer lies in Bayesian mathematics.

Why Bayesian filtering is better [4]. The Bayesian method takes the whole message into account- It recognizes key words that find out the spam, but it also identify words that denotes valid mail. A Bayesian filter is constantly self adapting - By learning from new spam and valid departing mails, Bayesian filter expand and adapts to new spam filtering techniques. The Bayesian filter technique is sensitive to the user. A Bayesian anti-spam filter, being adaptive, can be used for any language required. A Bayesian filter is difficult to make fool, as opposed to a keyword filter.

VI. CONCLUSION

This paper has described some spam detection techniques and various problem associated with spam detection. From the study we have conclude that we can't stop the spam but we can reduce it nicely by Bayesian techniques as compare to other techniques.

References:

[1] Christina V, Karpagavalli S, Suganya G, "A Study on Email Spam Filtering Techniques", International Journal of Computer Applications (0975-8887) - Volume12-No.1, December 2010.

[2] Saadat Nazirova,"Survey on Spam Filtering Techniques", Scientific Reaserch-Vol. 3, No. 3, August 2011.

[3] Gordon V.Cormack,David R.Cheriton,"Email Spam Filtering: A Systematic Review ", Foundation and Trends in Information Retrieval-Vol. 1, No.4(2006).

[4] (GFI is Microsoft Gold certified pattern) http://www.gfi.com.

[5] Chi-yao Tseng, Pin-Chieh Sung, and Ming-Syan,"Cosdes:A Collaborative spam Detection System with a Novel E-Mail Abstraction Scheme", IEEE Transactions on Knowledge and Data Engineering, Vol-23, No. 5, May 2011.

[6] G, Ashokkumar, S.Dhineskumar,"Spam Filtering", Department of Computer Science and Engineering, Anna University, Channai- 6000 025, India

[7] David Mertz"Spam filtering Techniques"IBM Developer Works.