IOSR Journal of Computer Science (IOSR-JCE)
e-ISSN: 2278-0661, p-ISSN: 2278-8727
PP 68-72
www.iosrjournals.org

Effective Spam Detection Method for Email

Savita Teli[1], Santoshkumar Biradar[2]

[1]*(Student, Dept of Computer Engg, Dr. D. Y. Patil College of Engg, Ambi, University of Pune, M.S, India)*

[2] *(Asst. Proff, Dept of Computer Engg, Dr. D. Y. Patil College of Engg, Ambi, University of Pune, M.S, India)*

ABSTRACT: *Spam emails are the emails receiver does not wish to receive; it is also called unsolicited bulk email. Emails are used daily by number of user to communicate around the world. Today large volumes of spam emails are causing serious problem for Internet user and Internet service. Such as it degrades user search experience, it assists propagation of virus in network, it increases load on network traffic. It also wastes user time, and energy for legitimate emails among the spam. For avoiding spam there are so many traditional anti spam techniques includes Bayesian based filters, rule based system, IP blacklist, Heuristic based filter, White list and DNS black holes. These methods are based on content of the mail or links of the mail. In this paper, we presented our study on various existing spam detection methods and finding the effective, accurate, and reliable spam detection method.*

Keywords: *Bayesian Filter, Efficient Spam Detection, Ham, Spam, Spam Filter.*

I. INTRODUCTION

Spam is abuse of electronic messaging system to send unsolicited bulk messages. Emails are used by number of user to communicate around the world. Along with growth of internet and email, there has been dramatic growth in spam in recent year. Spam can originate from any location across globe, where internet access is available. Spam was created by Hornel in 1937 as the world's first canned meat that didn't need to be refrigated. It was originally named "Hornel Spiced Ham", but was eventually changed to the catchier name, "SPAM" [6]. Usually they come in the form of advertisement, sometimes even containing explicit content or malicious code. Spam has been recognized as problem since1975. According to the statistics from ITU (International Telecommunication Union), 70% to 80% of emails in the internet are spams which have become worldly problem to the information infrastructure. In order to address growing problem there so many anti-spam methods.

II. SPAM

Spam is abuse of electronic messaging system to send unsolicited bulk messages. Today large volumes of spam emails are causing serious problem for the users, and internet services. Such as, It degrades user search experience, It assists propagation of virus in network, It increase load on the network traffic, It wastes the resources such as bandwidth, storage, and computation power, It also wastes the user time and energy.

General advices to avoid spam's are use the spam filter, Never reply the spam, Don't post your email address on your web site, and Never buy anything from spam.[6][3].

3. SPAM FILTERING ARCHITECTURE

Spam filter it minimize the amount of junk email. Email filtering is the processing of emails to organize it according to specified criteria. Common use of mail filters are Organize incoming mail, Removal of spam emails, Removal of computer virus.

Spam filter implemented at all layers, firewall exist in front of email server or at MTA(Mail Transfer Agent), Email server to provide an integrated anti-spam and Anti-virus solution offering complete email protection at the network perimeter level, before unwanted or potentially dangerous email reaches the network. At MDA (Mail Delivery Agent) level also spam filters can be installed as a

service to all of their customers. At email client user can have a personalized spam filter that then automatically filters mail according to the chosen criteria [1][6]. Fig. 1 shows the typical architecture of spam filter.

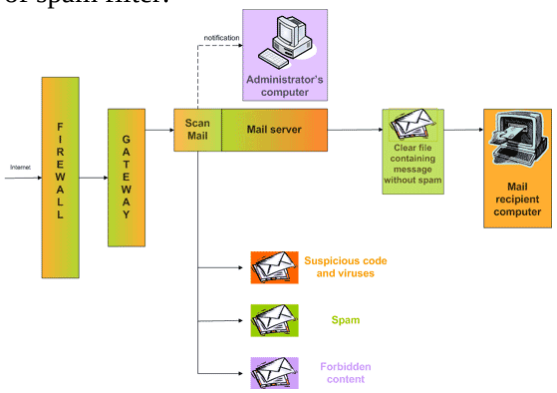

Figure 1: Spam Filtering Architecture

IV. SPAM DETECTION METHODS

Following are the sort list of desired filtering methods

4.1 List Based or Rule Based Filters

List based filter attempt to stop spam by categorizing senders as spammers or trusted users, and blocking or allowing their messages accordingly.

4.1.1 Blacklist

Black list is the form of rule based filtering that uses one rule to decide which emails are spams. Black list are the list of IP address of machine or record of email addresses that have been previously used to send spam. When incoming message arrives, the spam filter checks to see if it's IP or email address is on the black list, if so, the message is considered spam and rejected. Blacklist can be used for on both large scale and small scales [6].

Advantage is it can block substantial amount of email.

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

4.1.2 White list/Verification Filter

While Blacklisting is used to decide which emails are spam, but White listing is used to decide which emails are ham and assume all other emails are spam[6][7].

4.1.3 Blackholes [7]

Spam Blackholes work hand in hand with Blacklist. The way Blackholes work is someone posts message on websites, Usenet, forum, etc, showing their email address. The 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 is the email is received from one of these addresses the sending server can added to a Blacklist stopping it from sending any more messages.

Disadvantage is it can't see any disadvantages to using Blackholes in order to detect spam, they are important as they enable blacklist to be updated with computers that are sending unwanted emails.

4.1.4 Greylists

A relatively new spam filtering technique, it takes the advantage of the fact that many spammers only attempt to send a batch of junk mail once. Under the greylist 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 second time- a step most legitimate server will take – the greylist assumes the message is not spam and let it proceed to the recipient's inbox. At this time greylist filter will add the recipient's email or address to a list of allowed senders. Though greylist filter require fewer system resources than some other types of spam filters, they also delay mail delivery, which could be inconvenient when you're expecting time sensitive messages.

4.2 Content Based Filter

Content Based Filter is the most commonly used group of methods to filter spam. Content filter act either on the content, the information contained in the mail body, or on the mail headers (like "Subjects") to either classify, accept or reject a message [3].

4.3 Bayesian Filter

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

4.3.1 Mathematical foundation

Bayesian email filter take the advantage of Bayesian theorem is

$$\mathbf{P\ (spam/word) = [P\ (word/spam)\ P\ (spam)]\ /\ p\ (word)}$$

Theorem in the context of spam, says that the probability that an email is spam, given that it has certain words in it, is equal to the probability of finding those certain words probability that any email is spam, divided by the probability of finding those words in any email.

4.3.2 Process

Fig.2 creating Word Database for Filter

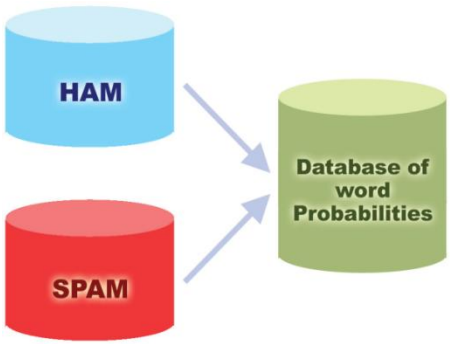

Figure 2: Creating Database for Filter

Particular words have particular probabilities of occurring in spam email and in legitimate email. The filter does not know these probabilities in advance, and must first be trained so it can build them up. To train the filter user must manually indicate whether a new email is spam or not. for all words in each training email, the filter will adjust the probabilities that each word will appear in spam or legitimate email in its database. For instance, 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 friends and family members. After training the words probabilities are used to compute the probability that an email with a particular set of words in it belonging to either category. 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, and if the total exceeds a certain threshold (for ex.: 95%), the filter will mark the email as spam. Email marked as spam then be automatically moved to a "Junk" email folder, or even deleted outright [6].

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

Advantages are it can be trained on per user basis, the spam that user receive is often related to the online user's activities, it 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.

Disadvantage is that they need to be trained properly in order for them work to work at their most effective and
The training leads 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 capturing, recording, and querying these early judgments [3].

V. EFFICIENT SPAM DETECTION METHOD

The methods currently used by most anti-spam spam software are static, mean that it is fairly easy to evade by tweaking the message little. To do this spammer simply examines the latest anti spam techniques and find the ways how to dodge them. To effectively combat spam, an adaptive new technique is needed. 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 identify spam, but it also recognizes words that denotes valid mail. A Bayesian filter is constantly self adapting – By learning from new spam and valid outbound mails, the Bayesian filter evolves and adapts to new spam techniques. The Bayesian method is sensitive to the user. The Bayesian filter is multi-lingual and international- A Bayesian anti-spam filter, being adaptive, can be used for any language required.
A Bayesian filter is difficult to fool, as opposed to a keyword filter

VI. CONCLUSION

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

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.