International Journal of Advance Engineering and Research Development (IJAERD)
Volume 1,Issue 5,May 2014, e-ISSN: 2348 - 4470 , print-ISSN:2348-6406

An Improved Spam Filter for Filtering Repeated Spam E-mails

Sherasiya Firozbhai A.[1]

[1]*Computer Engineering, Noble Engineering College -Junagadh, firozsherasiya@gmail.com*

Abstract—At present, email is one of the primary communication tool used every where due to Fast, Cheap, Secure and Reliable. But in the modern age of competition so many people (Organizations) are misusing this email tools in different ways such as sending email to so many unknown people's for advertisement or for other purpose. So such emails are known as spam emails (Spam mails). Now days so many people are sending such mails repeatedly such as at every hour, at every four hour, daily or weekly etc. Our main aim is to prevent such repeatedly spam mails faster than new incoming spam email.

Keywords-Spam Filter; Bayesian Method; Black List; Spam Mails; Text Classification

I. INTRODUCTION

Electronic mail (email or e-mail) is a method of exchanging digital messages from an author to one or more recipients. Modern email operates across the Internet or other computer networks. Some early email systems required that the author and the recipient both be online at the same time, in common with instant messaging. Today's email systems are based on a store-and-forward model. Email servers accept, forward, deliver, and store messages. Neither the users nor their computers are required to be online simultaneously; they need connect only briefly, typically to a mail server, for as long as it takes to send or receive messages [1].
The main advantages of email are easy to use, Fast, Cheap, Secure, Reliable and Automated.
The main disadvantages of emails are carry Virus, unwanted emails (Spam email) received.

II. SPAM EMAIL

Read an email is nowadays daily habits of many people because of emails are efficient, rapid and cheap mean of communication. This makes it preferred both in professional and personal correspondences. Additionally, Reading occasionally an E-mail from unknown source and content of which is not of the user interest is not really a luck. However, when more than 60% or even 90% of emails are of such kind, and often illegal; this is what one might call a nightmare. This kind of messages is said spam email [2]**.** It is also known as junk email or unsolicited bulk email (UBE).

III. SPAM FILTER

Spam filter is a tool that is used to block unwanted emails (Spam email) to enter into your inbox.

Fig 1: Spam Filter

Various Spam Filters available in market.

1. Mail Washer Pro
2. Spam Fighter

3. Choice Mail One
4. IHate Spam
5. Clean Mail Home
6. Spam Bully
7. Spam Eater Pro
8. Spam Buster

IV. METHODS USED TO IMPLEMENT SPAM FILTER

Various methods are used to implement Spam Filter from which each methods have their own advantages and disadvantages.

Types of SPAM filters	Method	Advantages (Pros)	Disadvantages (Cons)
Blacklist	Blocks mail from banned senders.	Blocks known SPAM messages	Does not block new SPAM
White list	Allows mail only from approved senders.	Blocks mail from unknown senders	Blocks new legitimate mail
Bayesian	Text recognition technology.	• Calculate the probability of the message (SPAM or not). • Self learning technique.	Does not deal with HTML or image mails.
Finger prints	Assign fingerprint for SPAM message.	Construct database for SPAM mails, and prevent them from passing through.	Only effective with identifying repeating emails (after the first one has been fingerprinted).
Password	Passwords are required to be in the email to pass through the filter.	Allows only the emails that have password to pass through.	Blocks new legitimate emails that does not have password yet.
Challenge! Response	Blocks unapproved mail until response arrives,	Allows only legitimate senders to pass through after their response.	• Blocks new legitimate mails. • Annoy legitimate senders by asking for a response with each message.
Community base	Blocks mail based on community agreement.	Block a SPAM that a group decides to block.	Does not block a new SPAM.
Encryption and Trust	Send mail with digital signature.	Digital signature is very hard to fake, and also used to sign and encrypt every message that is sent out.	The encryption technique is too complicated for the users.
Copyright Tokens	Uses the copyright tokens as an anti-SPAM tool.	Emails cannot be received without their own tokens.	Spammers can attach the same tokens to the messages they send.

Fig. 2: Methods used to implement Spam Filter [3]

V. BLACK LIST

A blacklist SPAM filter operates by creating a list of common words or phrases found in the header of the email message and domain name (the main part of the address of a web site, for example murdoch.edu.au - microsoft.com etc.), which can be used to decide if an email should be prevented from passing through the SPAM filter. However, a number of problems may occur if black list filter is the only filter used. For example

if a word "result", is blacklisted, and a user receives an email with a header (your exam "result"), and receives another email with a header (use our product for a quick "result"). Both emails will be blocked by the filter. Spammers may fly to circumvent this type of filter by either changing the contents of the email message or by using random character string for each message (e.g. ofers instead of offers). Another significant disadvantage of blacklists is maintenance. The Internet is unbounded and thousands of new sites are added every day and spammers continually change their identities. The problem with blacklists is the growing blacklist itself. The bigger it gets, the longer the processes required to physically check the black list and block a SPAM email [3].

VI. BAYESIAN FILTERING

Bayesian filtering is an extension of the text classification technology. This filter is a computer program used to recognize the words in a document, and can be implemented in a SPAM filter to search the textual content of an email. Bayesian filtering method uses text categorization algorithms to determine the probability that a certain email is SPAM. The algorithms are capable of categorizing the occurrence of certain words or phrases in terms of how and where they appear in the email message, but not by their existence alone.[3]
The challenge with content filtering is that SPAM emails often contain simply image links (e.g. photographs), which download image-based content to the receiver. Bayesian SPAM filters are capable of analyzing text, but are not capable of analyzing images. To carry out the analysis of images requires pattern matching techniques which is another area of research in itself. This analysis is beyond the scope of this study.[3]
Although the Bayesian filter is quite effective, it needs to be updated regularly. The reason for this is that it divides the incoming email messages into two classes, legitimate or illegitimate. Following this, each email is split into tokens (words, html codes. etc.) so their occurrence in the body of the messages can be computed. Based on this occurrence and using a specific mathematical formula, the probability that an email is SPAM or not can be calculated.[3]

VII. IMPROVEMENT IN SPAM FILTER

In our black list we have list of email address from which we don't want further emails to enter in our inbox. Means black list consist of list of email address from which we want to block further emails. But now a day most of email users receive spam mail from the same mail address repeatedly so we can sort the email address present in our black list to improve the performance (to make it fast). Now the problem is that how to sort these email address or what is the criteria to sort all the email address present in our black list. So at every time when we filter the spam mail place the senders email address at the top of the list. So as soon as the same sender will send the next mail it will compare it with our database and will find in our first comparison so no need to compare for all the other email address. So in such a way we can fast filter the Spam emails that are repeatedly send by the same user from the same email address.

VIII. CONCLUSION

Spam is becoming one of the most annoying and malicious additions to Internet technology. Traditional spam filter tools that uses black list cannot sort the email address present in the black list and can add new email address at the end (bottom) of the list. So at every time you have to check entire list and comparison is done from first email address to last email address. So our aim is to sort our list every time to make it fast.

IX. FUTURE WORK

We can also filter emails by using the image present in the mail. So many emails have image in it. So we can store image and compare it with the image present in incoming mail if the image match then we can filter it as Spam email and increase the count of that image as well as sender email id. Once the count of sender email id's count reached at a specific count (decided by user) add that email id in black list.

REFERENCES

[1] http://en.wikipedia.org/wiki/Email
[2] An Overview of Content-Based Spam Filtering Techniques by Ahmed Khorsi, Informatica 31 (2007) 269-277 269
[3] A study on intelligent adaptive SPAM Filters by Tarek Hassan, Murdoch University, 2006
[4]Evaluating Third-Party Bad Neighborhood Blacklists for Spam Detection by Giovane C. M. Moura, Anna Sperotto, Ramin Sadre, and Aiko Pras University of Twente, 978-3-901882-50-0 c 2013 IFIP

[5]Giovane C. M. Moura, Anna Sperotto, Ramin Sadre, and Aiko Pras, Evaluating Third-Party Bad Neighborhood Blacklists for Spam Detection, 978-3-901882-50-0 @2013 IFIP
[6] http://en.wikipedia.org/wiki/Blacklist_%28computing%29
[7]Trusted Behavior based Spam Filtering by Cong Wang and Jianyi Liu, 2010 International Conference on Web Information Systems and Mining
[8]SOAP: A Social Network Aided Personalized and Effective Spam Filter to Clean Your E-mail Box by Ze Li and Haiying Shen, IEEE @ 2011
[9] Mailbook A social network against spamming by Dimitris Zisiadis, Spyros Kopsidas, Argyris Varalis, Leandros Tassiulas, IEEE @ 2011
[10] A Survey of Text Classification Techniques for E-mail Filtering by Upasana and S. Chakravarty, 2010 Second International Conference on Machine Learning and Computing

