International Journal of Computer Applications Technology and Research
Volume 2– Issue 6, 759 - 762, 2013, ISSN: 2319–8656

A Hybrid approach for enhancing the capability of Spam Filter

Amandeep Kaur
Punjab Institute of Technology
Kapurthala, Punjab,
India

Malti Sarangal
Punjab Institute of Technology
Kapurthala, Punjab,
India

Abstract: In this paper, A hybrid approach for enhancing the capability of spam filter is proposed. This hybrid approach has defined the best features like Whitelisting, Blacklisting, Greylisting, Bayesian filtering, eXpurgate technology, Social closeness that are used to make sure that all fetched mails are checked against spam or not and redirected to user inbox or spam folder respectively**.**

Keywords: UBE; UCE; Bayesian; Blacklist; Whitelist; Greylist; eXpurgate technology.

1. INTRODUCTION

Electronic mail (email) has become a promising component for Internet users. The most common problem observed in maintaining email inbox is the incoming spam mails [7]. UCE or UBCE (Unsolicited Bulk Commercial Email) is the fastest and cheapest method of advertising the commercial websites [3]. UCE stands for Unsolicited Commercial E-mail where the word unsolicited means something unwanted, not requested or invited. These Spam mails are forcefully sent to many users at once. These mails fill the user inbox with thousands of unwanted mails that create the difficulty for user to differentiate between legitimate mails and spam mails. Many copies of the same mail are sent to many users at once. Spammers are the group of people who willingly spread spam in order to advertise their products, commercial websites across the internet [7]. Because sending spam mails through internet is the cheapest and fastest way of advertisement. The most common method of blocking the spam mails and let the only legitimate means useful traffic to pass through is called SPAM Filter. As the spammers enhanced their techniques the spam filters enhanced too to support the needs of user.

Adaptation to these new spam techniques is not observed in the most of spam filters available today. Also these filters do not deal with image spam and currently image spam is become a serious problem that spammers are using now a days. No doubt the commercial available email providers are using unique featured, efficient spam filters but they are not available online. So an individual user cannot use these filters for his machine. Some filters with excellent features come at higher price that a normal user can not afford [7]. So, this hybrid approach defines all unique features that will be available online at free of cost. It is combination of all the excellent features of the already available filters and also will remove the limitations of those filters. It has the ability to learn and adapt from the user's choices and establish a "Blacklist", "Whitelist" and "Greylist" of the messages and SPAM. It will be more effective and accurate in blocking the unwanted messages.

2. PROBLEMS CAUSED BY SPAM

2.1 Cost

Unwanted Spam mails costs a lot to the email providers. Spammers' costs are almost always borne by end users, because spammers often steal hardware and network resources. Spammers use networks of hijacked computers (botnets) as email clients [10]. Spam mails also wastes the network bandwidth by increasing the traffic over internet [2].

2.2 Time Constraint

- It wastes the precious time of the organizations as organizations spend a lot of time in identifying whether the incoming mail is useful for their organization or it is unwanted mail i.e. Spam mail before passing it into spam organization's email inbox. All this time is wasted, costs the company a lot more than the amount spent in initially sending out the spam mail by the spammers [7].
- It wastes the time of user also because he has to spend time in differentiating between the legitimate mails and useful mails. Employee time spent on checking, interacting and removing SPAM emails [2].
- Network administrator's time required to spend dealing with SPAM (scanning, cleaning) and/or associated problems on viruses and malicious applications [2].

2.3 Malware spreading/ Phishing

There are lot of problems that Spam mails create when spread across the internet. Many of the spam carry website links, that on clicking redirect to foreign sites that are harmful to the user's computer. It redirects the users control to phishing sites. Confidential (Personal) information of user is requested through the 'data fields' of such sites using which spammers obtain important personal information such as credit card information of the users [7].

2.4 Blank spam mails

Spam mails can be blank mails with blank body as well as no information even in the subject of the mail. Sender information is also made unavailable to the end recipient. Blank emails are sent by the spammers which enables them to differentiate between valid email address and invalid address under an email provider. Invalid addresses mails bounce back thus providing spammers with only valid email addresses to further send spam mails. Blank mails sometimes also spread malware which can harm the data in the user's computer. Trojans in the form of attachments are sent [7].

2.5 Forwarded mails

These forwarded mails are another problem causes spam. In some spam mails, spammers initialize a spam mail and send it to few users, in order to stop receiving similar spam mails further, the user is forced to forward the mail to some others in the mailing list. Hence, even if half the users forward the mail, the amount of spam created is immense and would require lot of cost to be removed off the internet [7].

2.6 Garbage/ Not legal data

Most of the spam mails prevalent are useless mails consisting of nothing that is meaningless to the user. Spam mails usually contain information about schemes and products that are not of much use to the individuals. Fraudulent schemes, solutions for situations, free advice, links to phishing websites etc. are sent through spam mails that are only contain the garbage material. Illegal content also spread across the Internet via spam mails. In certain countries, laws are implemented against display or spread of certain content. Spammers, against those laws, try to spread out content that is considered illegal through spam mails [7].

3. LITERATURE REVIEW

A solution proposed way back when Internet came into existence was to implement spam filters to avoid spam from filling email inbox to the brim. A SPAM filter is a set of instructions for determining the status of the received email. SPAM filters are used to prevent SPAM email passing through to the recipient. The challenge is how to design an effective SPAM filter that allows desired email mail to pass through while blocking the unwanted SPAM emails. The potential unwanted problem is that often a SPAM filter may identify a legitimate email as a SPAM, and block it (false positive), or identify SPAM email as legitimate email, and allow it to pass through (false negative). Of these two cases, implications on the false positive can be very serious as important legitimate emails may not reach the receiver. A means to quantify the effectiveness of a SPAM filter can be based on the percentage of SPAM emails being blocked, whitelist allowing legitimate emails to pass through to the recipients and blocks the mails that come from unknown sender. Listed below are three commonly used methods in SPAM filtering [2].

3.1 Blacklist Filter

Black list is effectively a list of emails that is not allowed to pass through. This can be based on the assumption that the email could contain a common word or phrase in the header, an IP address, or domain name. The use of a black list SPAM filter in isolation can result in false positive error. Assuming the word "results" is a keyword in the list, the following example will block both emails. If the email header is (your exam results), another email has (use our product for quick results), what is going to happen is the filter will block both emails. (False positive) [4].

3.2 Whitelist Filter

In this case, all the emails are treated as SPAM except the ones in the white list database. The database is built using a confirmation process by the recipient. The problem with this time consuming technique is that it causes unnecessary burden to the users [5].

3.3 Bayesian Filter (Content Focus)

This approach is an extension of text classification technology, which searches the textual content of an email and uses algorithms to identify SPAM email. The algorithms are able to classify the occurrence of certain words and phrases in terms of how and where they appear in the email. The challenge with content filtering is that SPAM emails sometimes contain images, which are difficult to interpret their contents [6].

4. EXISTING WORK

Every Incoming mail is parsed through these three filters step by step after that if mail is identified as Spam then passed to Spam folder or is allowed to enter into user's inbox.

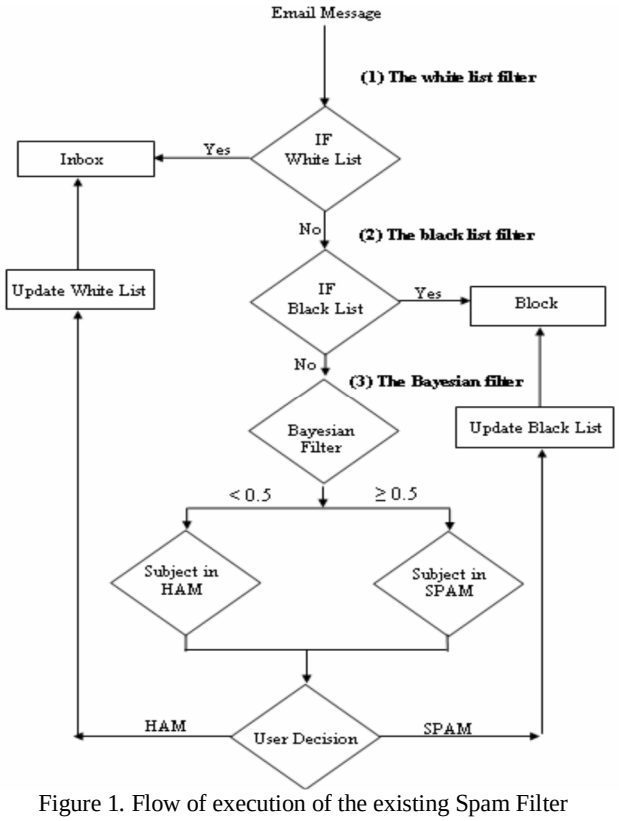

Figure 1. Flow of execution of the existing Spam Filter

Table 1: Comparison between different methods of SPAM filters

Spam Filter Methods	Block known SPAM	Block unknown emails	Self learning	Easy to use
Blacklist	✓			✓
Whitelist		✓		✓
Bayesian			✓	✓
Fingerprint	✓		✓	
Password		✓		✓
Challenge/ Response		✓		
Community Base	✓			✓
Mobile agent	✓			
Encryption and trust		✓		✓
Copyright tokens		✓		

This is the comparison of various filters methods that are used in detecting the Spam mails and and allows only legitimate mails to pass through into the user's inbox. Blacklist filter method is very efficient in blocking the known Spam mails and it is very easy to use. Whitelist filter is very much similar to blacklist filter but it blocks the unknown emails. Bayesian filter is a self learning filter as it automatically learns from new spam mail techniques. Fingerprint filter assign a fingerprint (distinct identifier) for spam message. It constructs the database for SPAM mails and prevent them from passing through. In Password filters passwords are required to be in the email to pass through the filter. But it blocks the new legitimate emails that does not have password yet. Challenge/Response blocks unapproved mail until response arrives and allows only legitimate senders to pass through after their response. But it blocks new legitimate mails and also annoy legitimate senders by asking for response with each message. Community Base filter method blocks mail based on community agreement means blocks a SPAM that a group decides to block but it does not block a new SPAM and also a one major drawback of this filter is that conflict may arise between the users while taking the decision about a particular mail is Spam or not. Encrytion and Trust Send mail with digital signature. Digital signature is very hard to fake and also used to sign and encrypt message that is sent out thus provide the security. But this technique is too complicated for the users and also cost and time wastage for small group of users. Mobile agent is a filter that works on remote system to perform the filtering on email server [2].

Table 2: Comparison between different approaches

Approaches	**Pers-onali-zed**	**Attack-resilient**		**User friendly**
		Imper-sonation	**Poison**	
Content based spam filters				
Static keyword	No	Yes	No	No
Machine-learning	No	Yes	No	No
			No	Yes
Collaborative	No	Yes		
Identity-based spam filters				
Black/white list	No	No	Yes	No
Social-interaction-based	No	No	Yes	Yes
Reputation	No	No	Yes	No
Social network aided content and identity based spam filter				
SOAP	Yes	Yes	Yes	Yes

5. PROPOSED WORK

No perfect SPAM filter has been found so far the following proposed approach [8] is aimed to enhance the capability of spam filter that can block SPAM emails and let legitimate emails to pass through using a combination of techniques including the use of the above approaches. It has defined the best features like Whitelisting , Blacklisting , Greylisting , Bayesian filtering, eXpurgate technology, Social closeness that are used to make sure that all fetched mails are checked against spam or not and redirected to user inbox or spam folder respectively. First of all user login with his details, user's credentials are checked if user is authorized, then before entering into user's inbox some techniques are used to check whether incoming mail is SPAM or not. If the incoming mail is identified as SPAM then it is redirected to SPAM folder otherwise incoming mail is allowed to enter into user's inbox. Firstly, eXpurgate technology is used for SPAM detection. With 2 step checking and adding extra header to incoming mail. The scope of false positive occurrence is decreased with help of eXpurgate technology. Then whitelist filter checks the incoming email against the white list. If the email address is found in the white list, then the filter will allow the message to pass through to the INBOX. If the sender mail id is not present in the white list, and if the mail sent by the sender is identified as spam then the mail id is added to another list called the Greylist. If another spam mail is sent by the same sender for the second time, the sender is then added into the blacklist thus blocking any further incoming mails into the inbox. Blacklist filter checks against the black list and blocks the known spam mails.

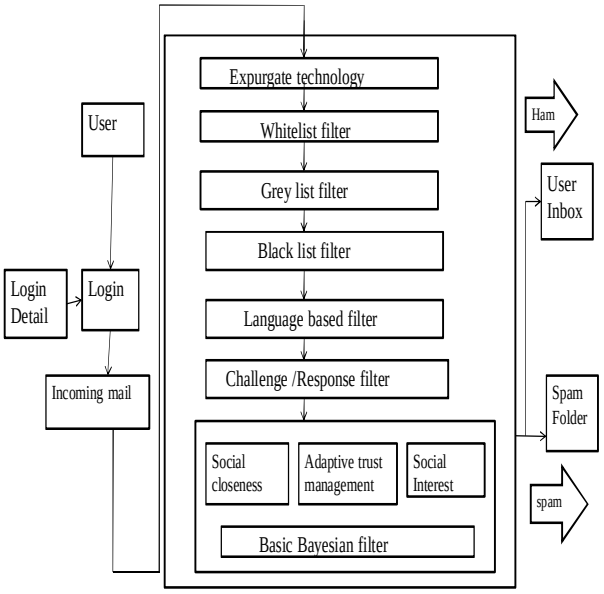

Figure 2. Proposed Approach

Language Based Filter is also used to remove out any incoming mail that is in any other language besides user's mailbox language preference. Challenge/Response filter sends an automated message that asks the sender to provide the return confirmation of his email address. If the filter has not recognized the incoming message as a white list or a black list, the Bayesian filter will be applied on <SUBJECT> field and the content <BODY> of the message. The filter scans through the message, and creates a probability of every word it knows about. Self learning Bayesian filter makes the approach more efficient and accurate. This approach integrates three new components to the Bayesian filter: (1) social closeness-based spam filtering (2) social interest-based spam filtering and (3) adaptive trust management. Based on the three social-based components, after parsing the keywords of an email, It adjusts the weights of the keywords. Then, it resorts to the Bayesian filter for spam evaluation. The weights are adjusted based on the closeness between the receiver and the sender, the receiver's (dis)interests, and the receiver's trust of the sender. If the closeness is high, the likelihood that the emails sent between them are spam is low, and then the weight is decreased otherwise weight is increased. Social closeness-based spam filtering helps filter to be resilient to poison attacks. Adaptive trust management helps filter to be resilient to impersonation attacks. Social interest-based spam filtering component contributes to the personalized feature. After processing the incoming mail through a all these filters, If the incoming mail is identified as SPAM then it is redirected into Spam folder otherwise the incoming mail is allowed to enter into the user inbox if it is identified as HAM means legitimate mail.

6. CONCLUSION

This paper provided the background problem caused by SPAM emails, and it also described the methodology of hybrid approach. This paper comprises a hybrid of the popular White List, Black List, Greylist, eXpurgate technology, Social closeness, Adaptive trust management, Social interest/disinterest and Bayesian Filters approaches that will effectively and accurately block the Spam mails and allow only legitimate mails to pass through to user's inbox based on the user's preferences. It is intelligent in the sense that it learns from the user's feedbacks and it is able to determine whether an incoming email message is a SPAM or not and also it adapts the new spam techniques.

7. REFERENCES

[1] Ze Li, Haiying Shen, "SOAP: A Social Network Aided Personalized and Effective Spam Filter to Clean Your E-mail Box",2011.

[2] Tarek Hassan, Peter Cole, Chun Che Fung, "An Intelligent SPAM filter – GetEmail5", 2006

[3] Khong,W.-K.(2001).The Law and Ecnomics of Junk Emails (SPAM).LAW AND ECNOMICS.Hamburg,University of Hamburg

[4] Moore D., Shannon C., Voelker G. M., and Savage S. "Internet Quarantine: Requirements for Containing Self-Propagating Code" IEEE INFOCOM April 2, 2003.

[5] Eric Allman "Features: Spam, Spam, Spam, Spam, Spam, the FTC, and Spam" Queue- Vol. 1 Issue 6, pages 62 - 69, September 2003

[6] Androutsopoulos I., Koutsias J., Chandrinos K. V., Paliouras G., and Spyropoulos C. D. "An Evaluation of Naïve Bayesian Anti-Spam Filtering" 11th European Conference on Machine Learning- Barcelona, Spain, pp 9-17, 2000

[7] Divya Tara Puvvula(2012) " SPAMKILLER: A TOOL FOR DETECTING AND FILTERING SPAM"

[8] Tarek Hassan, Peter Cole and Chun Che Fung "Towards Eradication of SPAM: A Study on Intelligent Adaptive SPAM Filters" Proceedings of the 5th PEECS Symposium, Perth, Western Australia, pp 203-206, September 2004

[9] C. Binzel and D. Fehr. How social distance affects trust and cooperation: experimental evidence from A slum. In *Proc. of ERF*, 2009

[10] Bindu V,Ciza Thomas "Spam War:Battling Ham against Spam"IEEE 2011

[11] "Anti-Spam Filtering Using Neural Networks and Baysian Classifiers" Proceedings of the 2007 IEEE International Symposium on Computational Intelligence in Robotics and Automation Jacksonville, FL, USA, June 20-23, 2007