INTERNATIONAL JOURNAL FOR ADVANCE RESEARCH IN ENGINEERING AND TECHNOLOGY

WINGS TO YOUR THOUGHTS.....

Email Filtering Based On Text Analysis and File Extension Using Improved Bayesian Algorithm

Neha Thakur[1], Manmohan Sharma[2]

[1]Student of Lovely School of Technology and Sciences
Lovely Professional University
sweetnehathakur85@gmail.com

[2] Faculty of Lovely Professional University
Phagwara, Punjab
manmohan.16073@lpu.co.in

Abstract: *Electronic mail (E-mail) is an electronic message system that transmits messages across computer network. Electronic mail is the easiest and most efficient communication tool for disseminating both wanted and unwanted information. There are many efforts under way to stop the increase of spam that plague almost every user on the internet. Managing and deleting scam or unwanted messages pose negative effects to user's productivity. However the attack of scam on business site also affects the customer. There is an increasing trend of integration of anti-spam techniques into mail transfer agent whereby the mail systems themselves also perform various measures that are generally referred to as filtering, ultimately resulting in spam messages being rejected before delivery or blocked. This paper present a E-mail intelligent system using Bayesian algorithm to reduce overload on mail traffic, shutdown of mailbox and waste of disk storage on mail server.*

Keywords: *E-mail, Bayesian approach, Spam Filtering, Junk mail, Bayes theorem, Scam Filtering.*

1. INTRODUCTION

Electronic mail (email) is the easiest and most efficient way to communicate. Internet users can simply type a letter and at the instantaneously communicate with people in all over the world [1]. Electronic mail (E-mail) is an essential communication tool that has been greatly abused by spammers that cause to become widely known unwanted information that are messages and spread malicious contents to Internet users. E-mail's serves as an archival tool to some people while many users never discard messages because their information contents might be useful at a later date as a reminder of upcoming events [2]. The volume and capacity of E-mail are constantly growing. Electronic messages posted blindly to thousands of recipients and represent one of the most serious and urgent information overload problems. An e-mail message that is unwanted is that in it electronic version of junk mail that is delivered by the postal service. The term spam refers to unsolicited, unwanted, and inappropriate bulk email [3]. Spam is often referred to as Unsolicited Bulk Email (UBE), Excessive Multi-Posting (EMP), Unsolicited Commercial Email (UCE), and Unsolicited Automated Email (UAE), bulk mail or just junk mail [4]. Spammers use many tactics to get email address to send spam and they also used computer programs called robots or spiders to harvest email address from websites. Through the internet, spammers can get the email from news group posting, webpage or mailing list. E-mail allows users to communicate with each other at a low cost as well as provides an efficient mail delivery system. The main problem with spam is that it makes up 30% to 60% of mail traffic and is on the rise. It can make the mail traffic become slow. When spam received and storage in mailbox, the mailbox can cause the problem like shutdown. When dealing with scam, ISP must build a sophisticated program into their system. Other problem at ISP site is server strain [5]. When sending and receiving amount of email in short period of time, server may become strain on ISP resources. They have to upgrade their equipment and pay higher bandwidth bill to deal with the rise of traffic. Sometimes, scammers using multiple

INTERNATIONAL JOURNAL FOR ADVANCE RESEARCH IN ENGINEERING AND TECHNOLOGY
WINGS TO YOUR THOUGHTS.....

combination of common name at popular domain name to send scam.

The risks of not filtering spam are the constant flood of spam networks clogs and corresponding impacts on user inboxes, but also downgrade valuable resources such as bandwidth and storage capacity, productivity loss and interfere with the expedient delivery of legitimate emails [6]. Not only is spam frustrating for most email users, it strains the IT infrastructure of both software and hardware of an organizations and costs businesses to lose billions of dollars in their productivity.

Today, Spammers are exploring the advantages of electronic mail (email) .This is because of its efficiency, effectiveness and it is considered very cheap as they can send the same messages to many email users from addresses gotten by various means [7]. For example, the use of automatic programs called bots such as web crawlers or spiders to scour the Web and Usenet news groups, collecting addresses, or buy email addresses in bulk from other companies at very low prices. Thanks to spoofing, spammers are now able to defraud innocent and greedy victims. In order to address the various growing problem in spam, organization must analyze the tools available to determine how best to counter spam in its environment. Tools, such as the corporate e-mail system, e-mail filtering gateways, contracted anti-spam services, and end-user training, provide an important arsenal for any organization [8].

2. LITERATURE REVIEW

The investigation on the usage of the word "spam" being associated with unsolicited commercial emails is not entirely clear. The fact that SPAM was created by Hormel in 1937 as the world's first canned meat that didn't need to be refrigerated. It was originally named "Hormel Spiced Ham," but was eventually changed to the catchier name, "SPAM." Its connection to email is according to Hormel and many other sources, due to a sketch on the British comedy TV show, Monty Python's Flying Circus. In the skit, a group of Vikings sing "SPAM, SPAM, SPAM" repeatedly, drowning out all other conversation in the restaurant an in-depth research into the history of spam on the internet was carried out by Brad Templeton, founder of Clarinet Communication Corp. According to him, the first email spam was from1978, and was sent out to all users on ARPANET (several hundred users). It was an ad for a presentation by Digital Equipment Corp. Templeton notes that the origin of spam as we know started on Usenet and migrated to email. Fabrício B used content filtering techniques whereby content are blocked or allowed based on analysis of its content rather than it source or other criteria. However there was no a clear security model standard designed to limit the extent of security incidents such as worms which could potentially overload the Internet causing a global denial of service. Developing intelligent and sophisticated content filtering technology with standards and cooperation among ISPs may be the solution. Natarajan 2010[9] provides a third party large-scale blacklist to decide which email is spam. A blacklist is a list of traits that spam emails have, and if the email to be tested contains any of those traits, it is marked as spam. It is possible to organize blacklist based on "From:" fields, originating IP addresses, the subject or body of the message, or any other part of the message that makes sense. A small-scale blacklist works fine if the user gets spam from one particular address. He was unable to provide a solution on a larger scale, where the user does not have any control over the blacklist, there must be a mechanism in place for dealing with accidental blacklisting of other users [10] The report by O' Brien J and Chiarella J (2003) [11] state that it is obvious problem that it is impossible to predict who is going to send email, and anyone previously unknown to the user will be filtered out. One way is to avoid this problem is to read through the filtered email regularly but there is no point in filtering if the user must view all of the email anyways.

Androutsopoulos, [12] in this work I define how Bayesian is different from others because of its learning. To decide that incoming mail is spam or not, the filter needs to know about the mail that user receives. Spam is kept in separate table and that probabilities can be calculated. In this case, the user must manually indicate whether that email is spam or not to train the filter there should be an intelligent mechanism to investigate the required trained word. Grey listing is the technique to temporarily reject messages from unknown sender mail servers as reported in [13]. In a related review Clark et al. [14] presented automated E-mail systems that were able to fill up the incoming E-mail messages into folders and anti-spam using neural network based system. The investigations from the study reveal that the technique is more accurate than several other techniques. The proposed technique mainly deals

INTERNATIONAL JOURNAL FOR ADVANCE RESEARCH IN ENGINEERING AND TECHNOLOGY

WINGS TO YOUR THOUGHTS.....

with clustering or grouping of mails into appropriate folders, rather on e-mail filtering. Wu (2009)[15] used a hybrid method of rule-based processing and back-propagation neural network for spam filtering. A rule-based process is first employed to identify and digitize the spamming behaviors observed from the headers and sys logs of e-mails. Then they utilize the spamming behaviors as features for describing e-mails. This information is then used to train the BPNN. The system produced very low false positive and negative rates. Meizhen et al. (2009)[16] proposed a model for spam behavior recognition based on fuzzy decision tree (FDT). This model can efficiently detect and analyze spammers' behavior patterns, and classify e-mails automatically. They concluded that since absolutely clear attributes does not always exist in the real world, the attribute subordinating degree is more natural and reasonable to describe the characteristics of behavior. Fuzzy decision tree is more adaptive than Crisp decision tree. In the aforementioned related research work, spam filtering methods is devised to work on the receiving end. Merely detecting a user sending out email after email and terminating their access would probably be sufficient to block spammers. The problem does not lie in detecting the spam. The problem is that some ISPs are willing to let spammers use their service to send out thousands of emails. The report in this paper adopts the principle of quantitative and qualitative. The principle of the quantitative technique is asking as much respondents as possible to get adequate results of their search while quantitative is the method of data collection chosen in concordance with the explained methods.

3. ARCHITECTURE OF PROPOSED SPAM FILTERING SYSTEM DESIGN

The proposed system Architecture is based on Bayesian a technique that uses mathematical formulae to analyze the content of a message, learning from the user which is a valid message and which is spam. Bayesian spam filtering is the process of using Bayesian statistical methods to classify documents into categories. Using well known mathematics, it is possible to generate a "spam indicate probability" for each word. Bayesian is different from others because of its learning process. To decide that incoming mail is spam or not, the filter needs to know about the mail that user receives. Spam is kept in separate table and that probabilities can be calculated. Bayesian rule using this probability: for example, most email users encounter the word 'Viagra' in spam email, but rarely want it in other email. The filter doesn't know these probabilities in advance and must be trained first so it can build them up. A Bayesian email filter relies on two things to work effectively: how well the Bayesian analysis formula has been implemented and how good a sample of data it has to work with. According to Wikipedia (2011), Bayesian email filtering is the process of using Bayesian statistical methods to classify documents into categories. Using well known mathematics, it is possible to generate a "spam indicate probability" for each word

Using Bayes' theorem, one can conclude according to equation [j] that: P (spam | words) = P (words | spam) P (spam) / P (words) ……. Eq. j

Where P (spam | words) is the probability of spam where there is word

P (words | spam) is the probability of word where there is spam

P (spam) is the probability of spam

P (word) is the probability of word

3.12 Bayesian Statistical Scam Filter of the Proposed Design.

In probability theory and statistics, Bayes' theorem (alternatively Bayes' law or Bayes' rule) is a method of incorporating new knowledge to update the value of the probability of occurrence of an event. To that end the theorem gives the relationship between the updated probability P(A|B),the conditional probability of A given the new knowledge B, and the probabilities of A and B, P(A) and P(B), and the conditional probability of B given A, P(B|A). In its most common form, Bayes' theorem is:

P (A|B) = P (B|A) P (A)\ P (B)

Based on the theoretical background of Bayesian theory and provided spam (scam or non scam) is obtained, equation [K] is derived. P (scam | non scam) = P (non scam| scam) P (scam) / P (non scam)……..Eq. K

Where

Vol. 1, Issue III, April 2013
ISSN 2320-6802

INTERNATIONAL JOURNAL FOR ADVANCE RESEARCH IN ENGINEERING AND TECHNOLOGY

WINGS TO YOUR THOUGHTS.....

P (scam | non scam) is the probability of scam where there is non scam

P (non scam| scam) is the probability of non scam where there is scam

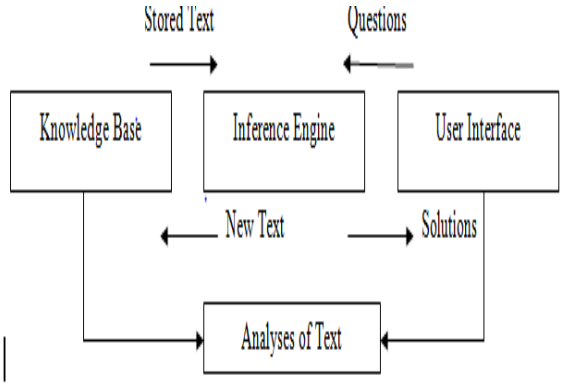
P (scam) is the probability of scam

P (non scam) is the probability of non scam

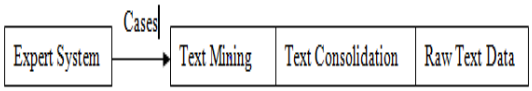
4. METHODOLOGY

This section presents a complete proposed system design, deduced system requirement and implementation. The report in this paper adopts the principle of quantitative and qualitative. The principle of the quantitative technique is asking as much respondents as possible to get adequate results of the research while quantitative is the method of text collection which should obtained by expert system and then analyses of text. Email filtering based on text analysis and file extension using improved Bayesian algorithm.

Step1:- Develop an expert system that known for the expert system concept of machine learning will be used and then known the expert system is that which analyses the text by itself.



Step2:- After developing the expert system, create some cases in the expert system which are useful for the text analysis and also for the better security that the cases can be if the mail is heavy, if the mail is undefined, if the mail is abused, if the mail is spammed, Check the mail by the extension of file and add some more cases while for the implementation.



Step3:- After creating the cases the developing expert system will then create a notification and also the actions of the corresponding actions. In the notification it will consist just sender id, receiver id and type of the file. The actions of the corresponding cases are:

- If the mail is heavy then compress it because, if my mail is heavy den it will take more time to download. This will happen in rare cases.
- If the mail is undefined then define the mail and then resend it.
- If the mail is abused then accurate the mail and resend it to the destination.
- If the mail is spammed, in this another case is developed, if the useful mail is gone in the spam box, and then resends it.
- If the file extension is ".bat", because .bat file extensions most of the time contains malicious viruses. So if the .bat file contains unknown user then it will be discarded. If the .bat file is from the known user then send the condition which includes "write the script in notepad and then send it"

In this way the security will be maintained and that working is a sample working of the project. spammed mails which can be useful for the user like some adds, will not be discarded. We will do further additions in this methodology in the implementation. In this paper we are sending spam mail will send by our system to the source, if again it will resend by source without changing content then our system will block the spam mail. We will modify the exiting algorithm and create of our own working model to carry out the better efficiency in future. We have basically modified the above working model.

5. CONCLUSIONS AND FUTURE WORK

The EMAIL filter software is also designed to remove every form of flooding and illegal spoofing. Over time we have seen detector automatically blocks messages from service provider because such messages wouldn't have been sent on a local INTERNET service provider's web application. The new intelligent system is designed to meet the local INTERNET providers' needs such as an automated view of activity logs of every action carried out by a user, deactivation and activation of clients, auto-train software with new words.

INTERNATIONAL JOURNAL FOR ADVANCE RESEARCH IN ENGINEERING AND TECHNOLOGY

WINGS TO YOUR THOUGHTS.....

Based on the conclusions and findings of this study it is noticed that the fight against filtering messages on EMAIL web application programs is an interesting and growing area of research which could be further investigated to include a variety of functionalities. The scope of the work did not cover for BULK SMS messages. BULK SMS is very cheap and these spammers always try to take advantage of this to defraud innocent citizens. Work is being going on the topic but there are still some areas such as detecting image SCAMs which is still ongoing.

The discussion of the proposed model is centered on our requirement for the design such as the GUI components containing a My Sql server as the database make up the scam filtering system. API was used to develop the system so that any Internet or E-mail Service Provider can easily integrate the system with their existing system. We experiment with Yahoo mail and Google mail using the same ham and scam messages.

In this paper, Bayesian approach Spam filtering application in Architecture of the Proposed Spam filtering system design and bayeain stastical scan filter of the Proposed Design are discussed.

REFERENCES

[1] **Ahmad, B. B. I., 2007. "Spam Filtering Implementation Using Open Source Software"Accessed 8th September, 2011.**

[2] **Chih-Chien, W., 2003. "Sender and Receiver Addresses as Cues for Anti-Spam Filtering,"Journal of Research and Practice in Information Technology, 36(1), 3-7.**

[3] **USIC3 - Internet Crime Complaint Centre Report (2006-2008).www.ic3.gov/media/annualreports.aspx. Accessed8 September.**

[4] **Fabrício B., Tiago R., Virgílio A., Jussara A & Marcos G. (2009); DETECTING SPAMMERSAND CONTENT PROMOTERS IN ONLINE VIDEO SOCIAL NETWORKS; In ACM SIGIRConference, Boston, MA, USA, July 2009.**

[5] **Pantel, P., 1998. "Spamcop—a spam classification and organization program," Proceedings of AAAI-98 Workshop on Learning from Text Categorization. 1998.**

[6] **Sakkis, G., Androutsopoulos, I., Paliouras, G., Karkaletsis, V., Spyropoulos, C. D., & Stama-topoulos, P., 2003. A memory-based approach to anti-Spam filtering for mailing lists. Information Retrieval. 6(1), 48–73.**

[7] **Androutsopoulos, I., Paliouras, G., Karkaletsis, V., Sakkis, G., Spyropoulos, C., and Stama-topoulos, P. 2000c. Learning to filter spam e-mail: A comparison of a naive Bayesian and amemory-based approach. In Proceedings of the Workshop on Machine Learning and Textual Information Access, 4th European Conference on Principles and Practice of Knowl-edge Discovery in Databases (PKDD 2000) (Lyon, France), H. Zaragoza, P. Gallinari, and M.Rajman, Eds. 1—13**

[8] **Symantec Global Internet Security Threat Report Trends for 2010 Volume XVI, Published April 2010. http://www.symantec.com/connect /2011. Accessed 30 June, 2011.**

[9] **Natarajan Arulanand (2010): Payload Inspection Using Parallel Bloom Filter in Dual Core Processor; Computer and Information Science: Vol. 3, No. 4; 2010.**

[10] **Rajkumar, B., Tianchi, M., Rei, S., Chris, S., and Willy, S., 2006, "Domain Specific Blacklists," Proceedings of the Fourth Australian Information Security Workshop (AISW-Net Sec 2006).10 Natarajan Arulanand (2010): Payload Inspection Using Parallel Bloom Filter in Dual Core Processor;Computer and Information Science: Vol. 3, No. 4; 2010.**

[11] **O' Brien J and Chiarella J (2003): AN ANALYSIS OF SPAM FILTERS; Available at; http://web.cs.wpi.edu/~claypool/mqp/spam/mqp.pdf on 9/10/2011.1**

[12] **Androutsopoulos I, Koutsias J, Chandrinos KV, Paliouras G, Spyropoulos C (2000). An evaluation of naïve Bayesian anti-spam filtering. Proc. Of the workshop on machine learning in the new information age: 11th Europe conference on machine learning, pp. 9- 17.**

[13] **Sender and Receiver Addresses as Cues for Anti-Spam Filtering. Journal of Research and Practice in Information Technology, 36(1), 3-7.**

[14] **Clark, J.; Koprinska, I.; and Poon, J. (2003). A neural network based approach to automated e-mail classification.Proceedings of Web Intelligence, Proceedings. IEEE/WIC International Conference, 702- 705.**

[15] **Wu CH (2009).Behavior-based spam detection using a hybrid method of rule-based**

techniques and neural network. Expert Systems with Application. Elsevier, pp. 4321-4330.

[16] **16. W. Meizhen, L. Zhitang, and Z. Sheng , "A Method for Spam Behavior Recognition Based on Fuzzy Decision Tree," IEEE, Ninth International Conference on Computer and Information Technology, pp. 236-241, 2009.**