A Brief Survey on Spam Mail Detection and Filtering Network System

P. Iswarya[1], S. Padmapriya[2]
[1]PG Student, Dept. of Computer Science and Engineering, A.V.C. College of Engineering, Mayiladuthurai, India
[2]Professor & HoD, Dept. of Computer Science and Engg., A.V.C. College of Engineering, Mayiladuthurai, India

***Abstract*: Unsolicited emails, known as spam, are one of the fast growing and costly problems associated with the Internet today. Electronic mail is used daily by millions of people to communicate around the globe and is a mission-critical application for many businesses. Over the last decade, unsolicited bulk email has become a major problem for email users. An overwhelming amount of spam is flowing into user's mailboxes daily. Not only is spam frustrating for most email users, it strains the IT infrastructure of organizations and costs businesses billions of dollars in lost productivity. The necessity of effective spam filters increases. In this paper, we presented an efficient spam filter techniques to spam email based on machine learning and collaborative filtering.**

***Keywords*: Collaborative Filtering, Mission Critical, Machine Learning, Spam Frustrating, Unsolicited Mails.**

1. Introduction

In recent years, internet has become an integral part of our life. With increased use of internet, numbers of email users are increasing day by day. It is estimated that 294 billion emails are sent every day. This increasing use of email has created problems caused by unsolicited bulk email messages commonly referred to as Spam. It is assumed that around 90% of emails sent everyday are spam or viruses. Email has now become one of the best ways for advertisements due to which spam emails are generated. Spam emails are the emails that the receiver does not wish to receive. A large number of identical message are sent to several recipients of email. Increasing volume of such spam emails is causing serious problems for internet users, Internet Service Providers, and the whole Internet backbone network. One of the examples of this may be denial of service where spammers send a huge traffic to an email server thus delaying legitimate message to reach intended recipients. Spam emails not only waste resources such as bandwidth, storage and computation power, but may contain fraudulent schemes, bogus offers and scheme. Apart from this, the time and energy of email receivers is wasted who must search for legitimate emails among the spam and take action to dispose the spam. Dealing with spam and classifying it is a very difficult task. Moreover, a single model cannot tackle the problem since new spams are constantly evolving and these spams are often actively tailored so that they are not detected adding further impediment to accurate detection.

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. A spam filter is a program that is used to detect unsolicited and unwanted email and prevent those messages from getting to a user's inbox. Like other types of filtering programs, a spam filter looks for certain criteria on which it bases judgments. For example, the simplest and earliest versions (such as the one available with Microsoft's Hotmail) can be set to watch for particular words in the subject line of messages and to exclude these from the user's inbox. This method is not especially effective; it may omit legitimate messages (called false positives) and passing actual spam messages. More sophisticated programs such as Bayesian filters or other heuristic filters, attempt to identify spam through suspicious word patterns or word frequency. Filter classification strategies can separated into two categories: those based on machine learning (ML) principles and those not based on ML. ML approaches are capable of extracting knowledge from a set of messages supplied, and using the obtained information in the classification of newly received messages. Non-machine learning techniques, such as heuristics, blacklisting and signatures, have been complemented in recent years with new, ML-based technologies. In the last few years, substantial academic research has taken place to evaluate new ML-based approaches to filtering spam. ML filtering techniques can be further categorized into complete and complementary solutions. Complementary solutions are designed to work as a component of a larger filtering system, offering support to the primary filter (whether it be ML or non-ML based).

2. Literature survey

A. *Can DNS-based Blacklists Keep Up with Bots?*

Anirudh Ramachandran, David Dagon, and Nick Feamster (2006) proposed many Internet Service Providers (ISPs), anti-virus companies, and enterprise email vendors use Domain Name System-based Black hole Lists (DNSBLs) to keep track of IP addresses that originate spam, so that future emails sent from these IP addresses can be rejected out-of-hand. DNSBL operators populate blocking lists based on complaints from recipients of spam, who report the IP address of the relay from which the unwanted email was sent. To be effective in blocking spam, information in the blacklist must have the following properties:

1. Completeness. The blacklist must contain a reasonable fraction of all spamming IP addresses.
2. Responsiveness (i.e., low response time). We term the period of time between when a host first starts sending spam, and when it ultimately becomes listed the response time. The blacklist must have a low response time so that other recipients can subsequently block spam originating from the respective IP addresses.

B. Dynamic Concept Drift Detection for Spam Email Filtering

L. Nosrati, A. Nemaney Pour (2011) proposed nowadays most of Internet user's surfer from spam emails. Filtering technique is one of the effective methods which help us to get rid of the spam emails. One of the problems of filtering is that it cannot detect spam emails accurately when the concepts change or drift happens as time goes by. Therefore, it is required to handle concept drift accurately and quickly. This paper proposes a new algorithm for concept drift detection with three different levels; control, warning, and alarm level. The results show that the proposed algorithm can detect concept drift more accurately compared with the previously proposed ones. In addition, it can detect sudden concept changes more accurately.

C. Spam and the Ongoing Battle for the Inbox

L. Nosrati, A. Nemaney Pour (2011) proposed around the time spam was becoming a major problem in 1997, one of us (Heckerman), along with other colleagues at Microsoft Research, began work on machine learning approaches to spam filtering. In them, computer programs are provided examples of both spam and good (non-spam) email. A learning algorithm is then used to find the characteristics of the spam mail versus those of the good mail. Future messages can be automatically categorized as highly likely to be spam, highly likely to be good, or somewhere in between. The earliest learning approaches were fairly simple, using, say, the Naive Bayes algorithm to count how often each word or other feature occurs in spam messages and in good messages. To be effective, Naive Bayes and other methods need training data—known spam and known good mail—to train the system. When we first shipped spam filters, spam was relatively static. We had 20 users manually collect and hand-label their email. We then used this collection to train a filter that was not updated for many months. Words like "sex," "free," and "money" was all good indicators of spam that worked for an extended period.

D. Spam Mail Detection and Blocking for E-Mail Security by Cascade Hybridization and Collaborative Recommender

Joshuva Goodman (2007) proposed that there are copious ways of communication methods in this digitally advanced world but Electronic mail which is also known as e-mail or email is the utmost competent method to communicate or transfer our data from one to another. There is the likelihood of going astray when transferring or communicating during e-mail. One of those misbehaviors is receiving huge number of undesirable e-mails from a set of unfamiliar senders. A huge number of these mails frequently consist of commercial content. In the current actual system to avoid the undesirable Email receiving Spam method is used. The other terms used for Email spam are unsolicited bulk email (UBE), unsolicited commercial email (UCE), direct mail, third-class mail or junk mail. Sending huge quantity of messages to haphazard set of recipients constitutes spam. This method can differentiate junk messages from other messages in many times but not always. None of the ways we have can be counteracting these undesirable messages or Email receiving in spam method. To clear up this enigma of receiving undesirable messages or Emails we contemplate the abstraction of Spam mail blocking system. In the contemplated method we can permanently counteract the incoming of undesirable messages or Email through Spam mail blocking system.

E. Understanding the Network Level Behavior of Spammers

Adapa Chandrakala, Gangu DharmaRaju, A V S Pavan Kumar (2018) proposed the network-level behavior of spammers, including: IP address ranges that send the most spam, common spamming modes (e.g., BGP route hijacking, bots), how persistent across time each spamming host is, and characteristics of spamming botnets. We try to answer these questions by analyzing a 17-month trace of over 10 million spam messages collected at an Internet "spam sinkhole", and by correlating this data with the results of IP-based blacklist lookups, passive TCP fingerprinting information, routing information, and botnet "command and control" traces. We find that most spam is being sent from a few regions of IP address space, and that spammers appear to be using transient "bots" that send only a few pieces of email over very short periods of time. Finally, a small, yet non-negligible, amount of spam is received from IP addresses that correspond to short-lived BGP routes, typically for hijacked prefixes. These trends suggest that developing algorithms to identify botnet membership, filtering email messages based on network-level properties (which are less variable than email content), and improving the security of the Internet routing infrastructure, may prove to be extremely effective for combating spam.

F. Measuring the Role of Greylisting and Nolisting in Fighting Spam

Anirudh Ramachandran and Nick Feamster (2006) proposed spam has been largely studied in the past years from different perspectives but, unfortunately, it is still an open problem and a lucrative and active business for criminals and bot herders. While several countermeasures have been proposed and deployed in the past decade, their impact and effectiveness is not always clear. In particular, on top of the most common content- and sender-based anti-spam techniques, two minor approaches are popular among system administrators to cope with this annoying problem: greylisting and nolisting. These techniques exploit known features of the Simple Mail Transfer Protocol (SMTP) protocol that are not often respected by spam bots. This assumption makes these two countermeasures really simple to adopt and, at least in theory, quite effective. In this paper we present the first comprehensive study of nolisting and greylisting, in which we analyze these spam countermeasures from different perspectives. First, we measure their world-wide deployment and provide insights from their distribution. Second, we measure their effectiveness against a real dataset of malware samples responsible to generate over 70% of the global spam traffic. Finally, we measure the impact of these two defensive mechanisms on the delivery of normal emails. Our study provides a unique and valuable perspective on two of the most innovative and atypical anti-spam systems. Our findings may guide system administrators and security experts to better assess their anti-spam infrastructure and shed some light on myths about greylisting and nolisting.

G. Experiences with Greylisting

Fabio Pagani, Matteo De Astis and Mariano Graziano (2003) proposed greylisting temporarily rejects mail from un-known sources on the theory that real mailers will retry while spam ware won't. I outline taxonomy of greylisters and report some statistics both on anti-spam effectiveness and its effect on non-spam mail.

H. Towards Better Bayesian Spam Filters

John R. Levine (2005) Spam, that is, unsolicited commercial or bulk email, is a rising problem that has become unavoidable to nearly all email users, and automatic filtering has become a strong area of interest for many users. The most widespread and effective technique for doing so is to use a naïve Bayesian classifier. This technique uses Bayes' theorem to determine the probability that the presence of each word in a given email corresponds to the email being spam or non-spam based on evaluation of previous spam and nonspam emails and analysis of their contents. By combining these probabilities together, the probability that a newly received email is spam or non-spam can be calculated, and filtering can be applied as necessary. The Bayesian filter approach has several advantages: it customizes itself to an individual's behavior, it continuously improves its performance, and it adapts to new spam techniques on its own (Graham, 2002). However, no filter is perfect, and every filter will accidentally allow some spam through (false positives) and occasionally mark a legitimate email as spam (false negatives). False positives are quite undesirable, and thus the filtration algorithm should be optimized to greatly reduce these while still allowing for few false negatives. I will test the effectiveness of multiple Bayesian algorithms on a spam dataset and several variations thereof to find an ideal filtration configuration. I will use a spam dataset created from 4601 emails received in 1997 by a Hewlett-Packard Labs employee (Hopkins, 1999). Each email is described by 58 attributes, of which 54 are numeric values that indicate how often a certain word or symbol appeared in the email, three are analyses of capital letter frequencies, and the last is the predetermined class (spam or non-spam).

I. Improved Bayesian Anti-Spam Filter Implementation and Analysis of Independent Spam Corpuses

P. U. Anitha, Dr. C. V. Guru Rao (2011) proposed spam emails are causing major resource wastage by unnecessarily flooding the network links. Though many anti-spam solutions have been implemented, the Bayesian spam score approach looks quite promising. A proposal for spam detection algorithm is presented and its implementation using Java is discussed, along with its performance test results on two independent spam corpuses – Ling-spam and Enron-spam. We use the Bayesian calculation for single keyword sets and multiple keywords sets, along with its keyword contexts to improve the spam detection and thus to get good accuracy.

J. Time-efficient spam e-mail filtering using n-gram models

Ali C, ıltık, Tunga Gu¨ngo¨ (2008) proposed that spam e-mail filtering methods having high accuracies and low time complexities. The methods are based on the n-gram approach and a heuristics which is referred to as the first n-words heuristics. We develop two models, a class general model and an e-mail specific model, and test the methods under these models. The models are then combined in such a way that the latter one is activated for the cases the first model falls short. Though the approach proposed and the methods developed are general and can be applied to any language, we mainly apply them to Turkish, which is an agglutinative language, and examine some properties of the language. Extensive tests were performed and success rates about 98% for Turkish and 99% for English were obtained. It has been shown that the time complexities can be reduced significantly without sacrificing performance.

3. Proposed system

Collaborative spam filters use the collective memory of, and feedback from, users to reliably identify spam. That is, for every new spam sent out, some user must first identify it as spam for example, via locally generated blacklists or human inspection; any subsequent user who receives a suspect e-mail can then query the user community to determine whether the message is already tagged as spam. In this method spam filtering system uses two key mechanisms to exploit the topological properties of social e-mail networks: the novel percolation search algorithm, which reliably retrieves content in an unstructured network by looking through only a fraction of the network, and the well -known digest-based indexing scheme. Uses machine learning algorithms by first learning from the past data available (seems to be the best at current). Here, follows a brief overview of e-mail spam filtering. Among the approaches developed to stop spam, filtering is an important and popular one. It can be defined as automatic classification of messages into spam and legitimate mail. It is possible to apply the spam filtering algorithms on different phases of email transmission at routers, at destination mail server or in the destination mailbox. The major point of this work is supplying generalized methodology to automatic data segmentation and collecting labeled training data, on the other hand, manipulating windows in real-time employing provisional information.

4. System architecture

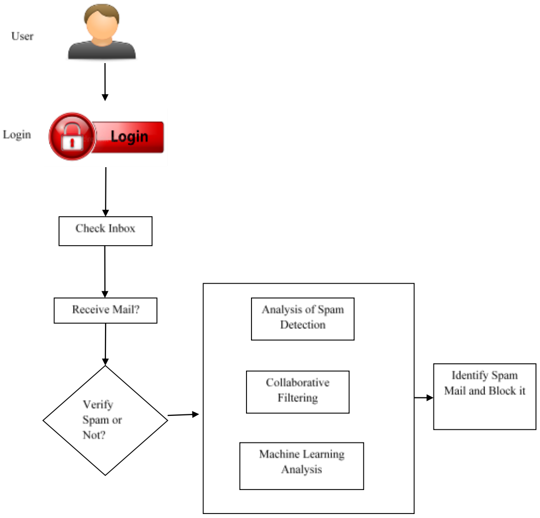
User
Login
Login
Check Inbox
Receive Mail?
Verify Spam or Not?
Analysis of Spam Detection
Collaborative Filtering
Machine Learning Analysis
Identify Spam Mail and Block it

Fig. 1. System Architecture

5. Conclusion

E-mail is an efficient, quick and low-cost communication approach. E-mail Spam is non-requested data sent to the E-mail boxes. Spam could be a huge drawback each for users and for ISPs. According to investigation nowadays user receives a lot of spam emails then non spam emails. To avoid spam/irrelevant mails we'd like effective spam filtering strategies. Spam mails area unit used for spreading virus or malicious code, for fraud in banking, for phishing, and for advertising. Spam messages are nuisance and huge problem to most users since they clutter their mailboxes and waste their time to delete all the junk mails before reading the legitimate ones. They also cost user money with dial up connections; waste network bandwidth and disk space. Bayesian classifier is one of the most important and widely used classifier and also it's the simplest classification method due to its manipulating capabilities of tokens and associated probabilities according to the users‟ classification decision and empirical performance. In this paper, we implemented the system to analyze each and every mail.

6. Future work

Future researches must address the fact that e-mail spam filtering is a co-evolutionary problem, since as the filter attempts to extend its predictive accuracy, the spammers attempt to outdo the classifiers. Hence, an effective approach should find a successful mechanism to identify the drift or evolution in spam features. Among all the traditional approaches discussed so far, the single approach that has achieved tremendous success against spam is content-based spam filtering. Fortunately, machine learning-based systems enable systems to learn and adapt to new threats, reacting to counteractive measures adopted by spammers.

References

[1] C. MacFarlane, (2003), "FTC Measures False Claims Inherent in Random Spam," Federal Trade Commission, http://www.ftc.gov/opa/2003/04/spamrpt.shtm, Accessed Jul. 20, 2011.

[2] L. Nosrati & A. Nemaney Pour, "Dynamic Concept Drift Detection for Spam Email Filtering," Proceedings of ACEEE 2nd International Conference on Advances Information and Communication Technologies (ICT 2011), Amsterdam, Netherlands, pp. 124-126, Dec. 2011.

[3] A. Ramachandran, D. Dagon & N. Feamster, "Can DNS-Based Blacklists Keep Up with Bots?," The Third Conference on Email and Anti-Spam (CEAS 2006), California, USA, pp.1- 2, Jul. 2006.

[4] J. Goodman, "Spam: Technologies and Policies," White Paper, Microsoft research, pp.1- 19, Feb. 2004. International Journal of Network Security & Its Applications (IJNSA), Vol.4, No.2, March 2012 62

[5] A. Ramachandran and N. Feamster, "Understanding the Network-Level Behavior of Spammers," Proceedings of the 2006 conference on Applications, technologies, architectures, and protocols for computer communications (SIGCOMM 2006), Pisa, Italy, pp. 291-302, Sep. 2006.

[6] E. Harris, (2003), "Greylisting: The Next Step in the Spam Control War," http://projects.puremagic.com/greylisting/whitepaper.html

[7] J.R. Levine, "Experience with Greylisting," Proceedings of Second Conference on Email and Anti-Spam (CEAS 2005), CA, USA, pp. 1-2, Jul. 2005.

[8] P. Graham, "Better Bayesian filtering," MIT Spam Conference, Jun. 2003.

[9] H. Yin & Z. Chaoyang, "An Improved Bayesian Algorithm for Filtering Spam E-Mail," IEEE 2nd International Symposium on Intelligence Information Processing and Trusted Computing (IPTC 2011), Huangzhou, China, pp. 87-90, Oct. 2011.

[10] A. Ciltik & T. Gungor, (2008), "Time-efficient spam e-mail filtering using n-gram models," Pattern Recognition Letters, Vol. 29, No. 1, pp. 19-33.