International Journal of Scientific Engineering and Applied Science (IJSEAS) – Volume-2, Issue-3, March 2016
ISSN: 2395-3470
www.ijseas.com

Review on Email Spam Filtering Techniques

Shiva Sharma[1], U. Dutta [2]
Computer Science and Engineering Department, Maharana Pratap College of technology
Putli Ghar Road, Near Collectorate, Gwalior-474006, Madhya Pradesh, India

ABSTRACT- In this paper we present email spam filtering and email authorship identification. Electronic mail is used by millions of people to communicate around the world daily and is a mission-critical application for many businesses. Over the last 10 years, unsolicited bulk email has become a major problem for email users called senders and receivers. And the recent years spam became as a big problem of internet and electronic communication known as users. There is developed lot of techniques to fight them. We presents the overview of existing e-mail spam filtering methods is given. The classification, valuation and juxtaposition of traditional and learning based methods are provided. I think using this paper user can understand easily spam filtering techniques and user can protect from spammed mails to financial damage, mentally damage to companies and annoying individual users.

Keywords: Spam, E-mail spam, Unsolicited Bulk Messages, Filtering, Traditional Methods, String Matching.

I. INTRODUCTION

The internet has become an Indiscernible part of everyday life and email has become a powerful tool for information interchange. Along with the growth of the Internet and e-mail, there has been a choreographic growth in spam in recent years. Spam is the use of electronic messaging systems (including most broadcast media, digital delivery systems) to send unfaithful bulk messages recklessly.

Written communication that employs the use of computers using email is referred to as computer mediated communication (CMC), and email is one of the most necessary and widely used forms of CMC. Other types of CMC include online messages, blogs, forums, whatsapp and instant messaging services. Among all CMC, email has remained a key source of written communication, especially in the last few years. Due to its salient features, it is the preferred source of written communication for almost every population (a part from illiterate) connected to the Internet. It is a very quick, asynchronous written communication channel that is used for various purposes ranging from formal to informal communication. Email messages can be sent to a single receiver or broadcasted to groups known as users. An email message can reach to a number of receivers simultaneously and instantly at singe time. These days, the majority of individuals even cannot imagine the life exclusive of email. For these and countless other motives, email has also become a widely used medium for communication of the people having ill intentions [1]. But we know that everywhere have some pitfalls by destructive persons who want to loss to other persons using spam emailing.

Spam is flooding the Internet with many copies of the same message in an attempt to force the message on people who would not otherwise choose to receive it. Spam is a combination of: unsolicited commercial e-mail (UCE), unsolicited non commercial e-mail (UNCE) and unsolicited bulk e-mail (UBE). [2, 3, 4]
The three main characteristics of spam email are; [2]

1. not requested by the recipients;
2. has commercial value;
3. always sent in bulk.

Spam is a problem because it wastes the email storage, user's time, email recipients time to open, read and delete the email. It targets

individual users with direct mail messages and bulk messages for multiple users. Creation of email spam lists is done by scanning Usenet postings, stealing Internet mailing lists, or searching the Web for addresses.

II. SPAM-UNSOLICITED BULK EMAIL

E-mail spam, known as unsolicited Pile Email (UPE), junk mail, advertisements or unsolicited commercial email (UCE), is the practice of sending unwanted e-mail messages for promotions as well as to get information with unauthenticated way, frequently with commercial content, in large quantities to an unscrupulous set of recipients. The technical definition of spam is an electronic message is called "spam" if

(A) The recipient's personal equality and out of context because the message is equally applicable to many other forceful Consignee; and

(B) The Consignee has not verifiably granted intentionally, explicit, and still-revocable permission for it to be sent.

The risks in filtering spam are sometimes legitimate mails may be rejected or disallowed and legitimate mails may be marked as spam. The risks of not filtering spam are the constant flood of spam clogs networks and contrarily impacts user inboxes, but also drain valuable resources such as bandwidth and storage competence, productivity lower and intervene with the expedient delivery of legal emails. General consultation to avoid spam is, Avoid giving your "real" email address to all but close associates, Setup web mail accounts (Google, hotmail, yahoo, rediffmail etc.) for registering with web sites or for communicating with people you do not know, edify your contacts to exercise caution with email address, Do not open junk email, just delete, remove it, Never click to unsubscribe to a mailing however you are sure it is a venerable entity.

III. INTRODUCTION CLASSIFICATION OF SPAM-FILTERING METHODS

Depending on used techniques spam filtering methods are generally divided into two categories.

1. Methods to Avoid Spam Distribution

Legislative measures limiting spam distribution, development of e-mail protocols using sender authentication, blocking mail servers which distribute spam are the methods which avoid spam distribution in origin.

Using these methods alone doesn't give considerable results. For example, there are many hard legislative re-striations for spam distribution in USA; nevertheless, the greatest amount of spam is distributed from this region. One of the reasons is an existence of high level broad-band Internet access in USA. There is a number of the approaches, offering to make spam sending economically unprofitable. One of these statements is to make sending of each e-mail paid. The payment for one e-mail should be the extremely insignificant. In this case for the usual user it will be imperceptible. For spammers who send thousand and millions messages the cost of such mailing becomes considerable that makes it economically un-profitable. This type of methods avoiding spam in their origins is a subject of author's another papers [5, 6]. They should be implemented together with the methods described in the next section, which filter spam at the destination point.

2. Methods to Avoid Spam Receiving

Methods which filter spam in destination point can be divided into the following categories:

- ✓ Depending on used theoretical approaches: traditional, learning-based and hybrid methods;
- ✓ Depending on filtration area: server side, client side and filtration in public mail-servers.

2.1. Classification of Spam Filtering Methods Depending on Theoretical Approaches

As we noted above depending on used theoretical approaches spam filtering methods are divided into traditional, learning-based and hybrid methods. In traditional methods the classification model or the data (rights, pat-terns, keywords, lists of IP addresses of servers), based on which messages are classified,

is defined by expert. The data storage collected by experts is called as the knowledge base. There are also used trusted and mistrusted senders lists, which help to select legal mail. Actually it makes sense only creation of the "white" list, because spammers use fictitious e-mail addresses. This technique can't represent itself as a high-grade anti-spam filter, but can reduce considerably amount of false operations, being a part of e-mail filtration system based on other classification methods.

In learning-based methods the classification model is developed using Data Mining techniques. There are some problems from the point of view of data mining as changing of spam content with time, the proportion of spam to legitimate mail, insufficient amount of training data are characteristic for learning-based methods.

Traditional methods Traditional methods are divided into the following categories:

1) ***Methods based on analysis of messages:*** The received e-mail is analyzed for specific signs of spam on the base of: formal signs; content using signature in updated database; content applying statistic methods based on Bayes theorem; content by means of use SURBL (Spam URL Real-time Block Lists)[7], when run search for located references in e-mail and their verification under base of SURBL. This method is effective if instead of advertisement, the reference of website with advertisement is located in e-mail.

2) ***Detectors of mass distribution****:* Their task is to detect distributions of similar e-mails to the bulk of users. The following methods are used for the detection:

- ✓ user's voting(Razor / Pyzor)[8,9];
- ✓ analysis of e-mails coming through mail system (DCC) [10];

3) ***Methods based on acceptance of sender as a spammer*****:** These methods rely on different black hole lists of IP and e-mail addresses. It is possible to apply own black hole and white lists or to use RBL services (Real-time Black hole List) and DNSBL (DNS-based Black hole List) for address verification. Advantage of these methods is detection of spam in early step of mail receiving process. Disadvantage is that the policy of addition and deletion of addresses is not always transparent. Often the whole subnets belonging to providers get to the Black lists. For such systems it is actually impossible to estimate the level of false positives (the legitimate e-mail wrongly classified as spam) on real mail streams.

4) ***Methods based on verification of sender's e-mail address and domain name*****:** This is the simplest method of filtration if DNS request's name is the same with the domain name of sender. But spammers can use real ad-dresses, so that current method is ineffective. In this case it may be verified with possibility of sending the message from current IP address. Firstly, the Sender ID technology can be used where sender's e-mail ad-dress is protected from falsification by means of publishing the policy of domain name use in DNS. Secondly, there can be used SPF (Sender Policy Framework) technology, where DNS protocol is used for verification of sender's e-mail address. The principle is that if do-main's owner wants support SPF verification, then he adds special entry to DNS entry of his domain, where indicates the release of SPF and ranges of IP addresses from where may become an email from users of current domain.

5) ***Method based on SMTP server response emulation:*** If the real mail delivery systems, which follow the SMTP protocol correctly, observe such error, they get some interval (1-2 hours) and repeat attempt again. But the majority of spam-bots has very short time out periods. So filters based on this method slow down the SMTP transaction to the point that some SPAM senders will fail but where real mail delivery systems will still continue and deliver mail successfully.

IV. HISTORY OF EMAIL SPAM FILTERING

So well aware is the fact that internet started to gain notability in the 1990's, and soon it started to be used for advertisement. It was then that Spam gained popularity and started to be used widely to send thousands of emails. The usage

of the word 'spam' is attributed to the British comedy troupe Monty Python [9] the historic significance lies in it being adopted to refer to unsolicited commercial electronic mail sent to a large number of addresses, in what was seen as drowning out normal communication on the Internet.

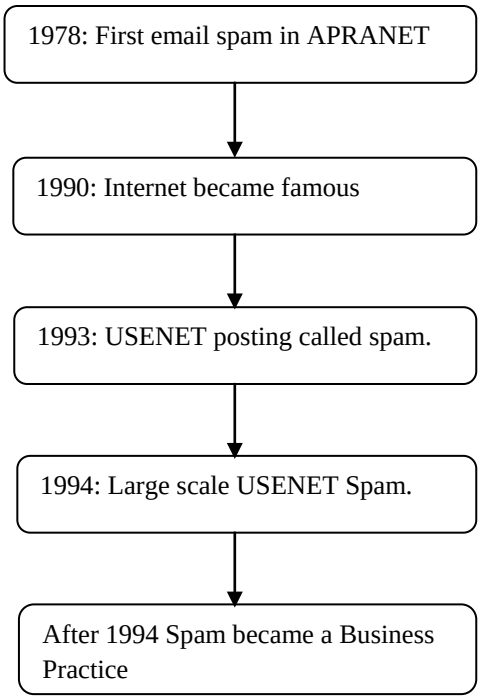

Figure 1: Spam Filtering History.

In figure 1 we explained the history of spam filtering [13].

V. SOME FAMOUS SPAMMERS

The Register of Conversant Spam Operations (ROCSO) database modulate information and evidence on known professional spam operations that have been finished by a minimum of 3 Internet Service Providers for spam offenses. To be placed on the ROCSO list a spammer must first be terminated by a minimum of 3 ISPs for AUP violations. Once listed in ROCSO, IP addresses under the control of ROCSO listed spammers [14] are directly and preemptively listed in the Spam Block List (SBL).

- Robert McGee
- Daniel Alvarez
- Yair Shalev
- Dante Jimenez
- Michael Lindsa

With the increase in spam and spammers it's important to come up with methods and techniques to prevent or stop Spam from piercing our mailbox and waste our time, resources and money.

VI. SPAM FILTERING TECHNIQUES

While new computer security threats may come and go, spam remains a constant. At a minimum, spam can interrupt your work, forcing you to spend time opening and deleting emails with useless medical information or fake investment opportunities. On a serious note, spam could inject a nasty virus in your organization's network, affecting your servers and desktop machines.

According to specialist the rate of spam is anywhere from 50 to 90 percent of all emails on the Internet. Anti-spam techniques typically use one or more filtering methods to identify spam and stop it from reaching a user's inbox. Spam filtering can be broadly done in two ways.

One is to stop spam before delivery: This is done by techniques that use filtering criterion, so that the unwanted content does not reach the mailbox itself. In this there are predefined set of rules that are set, another way to do is on the basis of IP Address, one can see from the span Master List of IP Addresses created and block or ignore or report as spam all messages coming from that IP address. Although the drawback with this technique was that smart spammers frequently switched their IP addresses and hence made it tough to identify the Spam.

- ✓ URL Based Filtering[15, 16]
- ✓ List Based Filtering[15]

Another way is destination Spam Filtering, in which the mail is classified as Spam or Ham based on the filtering techniques applied here.

They are mainly traditional methods of checking the spam signature and keywords. Black Listing is one such major technique. Machine learning techniques use data mining and AI to identify Spam.

- ✓ Key Word Based Filtering
- ✓ Content Based Filtering [16]

VII. CONCLUSION

In this paper, we presents the impact of spam email has been explored in many studies, and it's clear that email spam is grave both from individual nature as well as organizational. Spam arrives in the mailbox in form of Viruses, Advertisements and many more; all commercial and bulk, Phishing emails, antivirus or Political mails. Since 30 years, spam is diffused and causing trouble in daily internet based affairs leading to indecent use of personal and clandestine information as well as resources. In this paper various techniques that can help fight Spam like Web URL based filters, listing based and reserved keyword based filtering and content based filtering pledge Bayesian filtering combined with string matching to make it more athletic are discussed. With the use of these techniques and further enhancements one can fight Spam Hence; we can create many types of algorithms with the help of this paper. But also many scope for research in spam filtering field. Like classifying text, audio, video, multimedia messages and so on.

References

[1] Nizamani S, Memon N, Wiil UK, Karampelas P. CCM: a text classification model by clustering. In: 2011 International conference on advances in social networks analysis and mining (ASONAM), IEEE; 2011. p. 461–7.

[2] Izabella Miszalska, Wojciech Zabierowski, Andrzej Napieralski,"Selected Methods of Spam Filtering in Email, ", CADSM'2007,February 20-24, 2007, Polyana, UKRAINE ,p 02

[3] Liu Ming, Li Yunchun , Li Wei "Spam Filtering by Stages" presented at the International Conference on Convergence Information Technology, IEEE 2007, pp 01-04

[4] "Spam email percentage in mailbox" McAfee Managed Mail Protection Always On, Automatic Mail Protection. Available: Website:http:// www.mcafee.com/us/resources/misc/web-protectioninfographic. pdf [Accessed: Feb 5, 2014]

[5] S. A. Nazirova, "Anti-Spam Module for Filtering the Outgoing Correspondence," in Russian, *Transactions of ANAS, Informatics and Control Problems*, Vol. XXVIII, No. 3, 2008, pp. 158-162.

[6] S. A. Nazirova, "New Anti Spam Methods," *Proceedings on the Second International Conference on Problems of Cybernetics and Informatics*, Baku, 10-12 September 2008, pp. 89-92.

[7] Spam URL Realtime Block Lists. http://www.surbl.org/.

[8] Razor's homepage. http://razor.sourceforge.net/.

[9] Pyzor's homepage. http://sourceforge.net/apps/trac/pyzor/.

[10] DCC Spam Control Delayed Your E-Mail. http://mail.cc.umanitoba.ca/grey/.

[11] Symantec Brightmail Anti-Spam. http://www.symantec.com/business/premium-antispam.

[12] "History of Email Spam, Origin of Spam, Email Spam" Website: http://en.wikipedia.org/wiki/History_of_email_spam

[13] P. G. Juneja, R. K. Pateriya,"Survey on Email Spam Types and Spam Filtering Techniques", in International Journal of Engineering Research & Technology, March-2014, Vol. 3, Issue 3.

[14] "Famous Spammers in the World", Website:
http://www.spamhaus.org/statistics/

[15] Yang Li1,2, Bin-Xing Fang1, Li Guo1 "TTSF: A Novel Two Tier Spam Filter" IEEE Proceedings of the Seventh International Conference on Parallel and Distributed Computing, Applications and Technologies (PDCAT'06), 2006, pp 01-02.

[16] Jiansheng Wu 1, Tao Deng 2 "Research in Anti-Spam Method Based on Bayesian Filtering" IEEE 2008 IEEE Pacific-Asia Workshop on Computational Intelligence and Industrial Application, 2008 pp 01-02