A Survey on Email Spam Types and Spam Filtering Techniques

Prachi Goyal Juneja
M.Tech Scholar
Maulana Azad National Institute of Technology
Bhopal(M.P) India-462001

R. K. Pateriya
Maulana Azad National Institute of Technology
Bhopal(M.P) India-462001

***Abstract*- With the increase in usage of internet there is an exponential increase of spam in internet world. Spam is extraneous content sent to users without their consent, mostly associated with email. Spam wastes the user's time by fumbling up the inbox, consumes resources and spreads viruses leading to reduced efficiency. To make internet access more reliable and user friendly one needs to eliminate spam from our mailboxes so that the usage becomes fast and convenient. Various techniques in the past have been devised to tackle the problem of Email Spam: List based filtering; URL filtering, key word based filtering, content based: Bayesian filtering. In this paper it's discussed about spam, its evolution and impact. In this paper you will see different types of spam and spam filtering techniques.**

***Keywords*- Spam, Spam filtering, Email Spam, URL, Bayesian, String Matching.**

I. INTRODUCTION

Communication is necessary since always, be it in the Stone Age to alert each other of predators and hunt for food or in the Iron Age to talk, share ideas and come up with different tools or in the Machine age to build gadgets according to ones needs. There have been many ways of communications prevalent since the old ages: pigeons, human carriers, telegrams, letters, book posts, telephones and most recently Emails.

The evolution of email has transformed the age old method of communication for good, in terms of cost, usability and speed. But with every good thing comes a bad counterpart, so is the case with emails. The gift of emails is been noised by some people with unwanted mails. And hence junk emails largely known as Spam arrived. It is unfortunate that most email users are not aware of the real impacts of the spam emails either at the individual level or the organizational level, that eventually affect the economy of the country.

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) and unsolicited bulk e-mail (UBE). [1, 2, 3]

The three main characteristics of spam email are; [1]

(i) not requested by the recipients;
(ii) has commercial value;
(iii) always sent in bulk.

Spam is a problem because it wastes the email storage, email recipients time to open, read and delete the email. It targets individual users with direct mail messages. Creation of email spam lists is done by scanning Usenet postings, stealing Internet mailing lists, or searching the Web for addresses.

II. SPAM FILTERING IMPORTANCE

Email spam's typically cost users money. Many people - anyone with measured phone service - read or receive their mail while the phone line/ broadband connection is running, so to speak spam costs them additional money. On top of that, it costs money for ISPs and online services to transmit spam, and these costs are transmitted directly to subscribers.

Email is mainly used as communication tools other than telephone and any other tools. In an organizational setting, employees are more alert on incoming emails rather than telephone calls [1]. A majority of employees view the email within 6 seconds of its arrival which is faster than letting the telephone ring three times. So with the increase in tight deadlines and heavily trafficked schedules email spam wastes a lot of precious time.

Spam causes many problems in the daily computer savvy world, some of them being, slower access to mails due to network traffic or missing out on important mails or in today's world very specifically usage of tablets for mail access may cause hang outs and delays on the device due to large spam size. Creating effective communication with the customers is the most important aspect in service marketing [1], therefore, it is up to the marketers creativity to attract the interest of customers on their products or services.

Nowadays, email is more portable and email users can receive the email on their mobile phone personally which is always with them. Thus, the response on email notification is quicker than before. In business, email marketing is among the best way of advertising the business products of services online [1]. The ability of email to reach the global customers as their target market with the cheapest method is the main reasons email is widely used in marketing. Comparing to other methods, email is also popular because of the usability where different files can be attached such as pictures, documents and videos.

As per the announcement of McAfee Managed Mail Protection [4], approximately 25% to 50 % of all mails received by organizations are spam. Spam consists of more than 70% of the mails in the personal mailbox. To go through all these spam mails and deleting them is a waste of both time and energy. Spam not only wastes precious time but also costs money to users with dial up connections, wastes network resources, spread computer virus and expose under aged recipients to unsuited content. Hence these spam emails have to be segregated and its essential to find methods to prevent spam.[3]

III. DIFFERENT TYPES OF SPAM

Spam can be broadly classified into the various areas.
Unsolicited Advertisements: These are hundreds of billions of email advertisements sent daily selling weight loss cures, knock-off merchandise, online degree programs etc. They mainly include topics of Health and Medicine, Education and IT.
Phishing Spam: One of the hardest types of email spam to spot is phishing spam [5] emails. These emails are designed to look like official emails from financial institutions, e commerce websites and online greeting card services [6] but actually direct victims to equally official looking scam sites. This tricks people into giving away their usernames and passwords, which are then used by the site owners, the scammers, to make illegitimate transactions.
Nigerian 419 Spam: Many times you receive emails with offering a large sum of money with amazing offer by some stranger from a faraway land, a lottery service, an employer or even some business deal. They ask for some nominal amount of money as shipping or some hidden charges and trap you.
Email Spoofing: More of a technique used to make other email spam tactics seem more believable, many spammers will send messages which appear to originate from a different email address than they actually do. This spoofing technique [5] makes it appear as though a fraudulent email actually came from a trusted source, company or organization. This builds the trust of the victim, making them more likely to take part in whichever scam is included in the message.
Trojan Horse Email: Considered Prevalent long time back in the email spam world, email worms are bugs which not infect the victim computer and also send itself to everyone in the victim's contact list. The most famous email worm was the ILOVEYOU bug which debuted in 2000. [5, 7]
Commercial Advertisements: This includes when legit websites and companies that you use send out advertisements, newsletters and other junk messages. Most websites these days ask you if you'd like to be included in their communications however some will automatically add you to their mailing list simply for signing up for their site.
Anti Virus Spam: No one wants a virus so when victims receive emails saying that their computer is infected, some will believe the claim out of fear. Victims think they're downloading security software but they are actually infecting their computers with nasty viruses [7].
Political or Terrorist Spam: Part scare tactic and part attempt to steal personal information, this type of email spam appears to be from a politician or well-known government office, such as the FBI, claiming that you're in danger. To clear up the threat, the email asks the victims to fork over personal information and sometimes cash. The trick does get people to volunteer their personal information to untrusted sources.[8] Discussed above are various types of spam emails that are present in the inter world. While some of these only waste time, some even lead to leakage of personal information to unreliable sources and some result in monetary loss. Hence it’s important for the internet users to understand spam and its consequences.

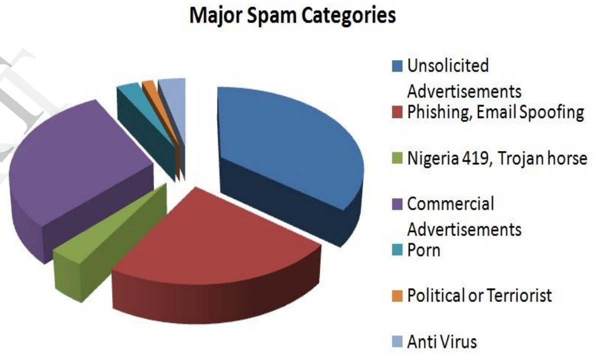


Figure 1: Spam Categories distribution

IV. HISTORY OF EMAIL SPAM FILTERING

So well aware is the fact that internet started to gain popularity in the 1990’s, and soon it started to be used for advertising. 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] (who allegedly staged a skit in which a group of Vikings sing a chorus of "SPAM, SPAM, SPAM..." at increasing volumes and attempt to drown out other conversation), 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.

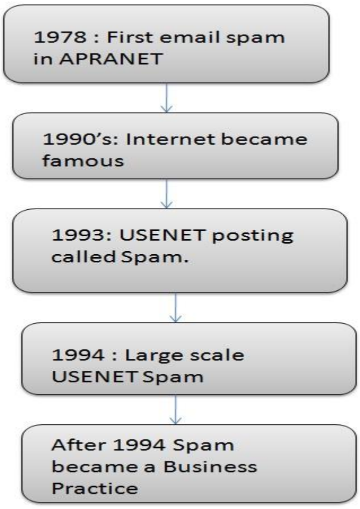

1978 : First email spam in APRANET
1990's: Internet became famous
1993: USENET posting called Spam.
1994 : Large scale USENET Spam
After 1994 Spam became a Business Practice

Figure 2 : Spam Filtering History

As per Brad Templeton, founder of ClariNet Communication Corporation, the first email spam was from 1978, and was sent out to all users on ARPANET. It was an ad for a presentation by Digital Equipment Corporation. [9]

It was after 1993 that a USENET posting was called Spam. On January 18, 1994, the first large-scale USENET spam occurred. A message with the subject "Global Alert for All: Jesus is Coming Soon" was posted to every available and possible newsgroup. Its controversial message sparked debates all across USENET, according to Templeton, the student who posted this message got in trouble later on. [9]

In April 1994, spamming became a business practice for the first time. Two lawyers from Phoenix, Canter and Siegel, hired a programmer to post their "Green Card Lottery- Final One?" message to as many newsgroups as possible.

V. SOME FAMOUS SPAMMERS

The Register of Known Spam Operations (ROKSO) database collates information and evidence on known professional spam operations that have been terminated by a minimum of 3 Internet Service Providers for spam offenses. ROKSO is a "3 Strikes" register. To be placed on the ROKSO list a spammer must first be terminated by a minimum of 3 ISPs for AUP violations. Once listed in ROKSO, IP addresses under the control of ROKSO-listed spammers [10] are automatically and preemptively listed in the Spamhaus 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 entering 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 experts 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 brodly 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[11, 12]
- ➢ List Based Filtering[11]

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 [12]

VII. VARIOUS SPAM FILTERING TECHNIQUES

Many techniques have evolved over the years to provide users a spam free mailbox, though to achieve a 100 % result is still not possible. But with the work going on in this field, a large number of spam messages are detected and moved to a spam folder. Many forms of spam filtering techniques are present, some of the major ones going to be discuss here and also each technique has its own merits and demerits. The usage of a spam filtering technique has to be gauged keeping in mind its merits and demerits and research is still going on to come up with a 100 % full proof technique. It's really tough to provide complete spam removal, because the spammers nowdays are so smart. In 1997, about 10% of emails received by corporate networks are spam. However, in 2003, Dunn reports that 75% of emails in an account are spam emails [13]. This raises the need for effective measures to reduce the spam emails.

(i) Preventive techniques

As widely heard Prevention is better than cure", in these types of techniques, the incoming mail is beforehand checked for authenticity and then allowed to enter the users mailbox.

URL Based: There appears a type of spam filtering method based on URL [14]. In this the incoming URL is first tested to be legitimate or not and if found clean is then allowed to drop the mail in the mailbox of the recipients. There is a repository thus created of these URL's that can be checked and updated from time to time. Because of the rapid rises of different URLs, the maintenance cost will be very large and thus results in the heavy cost of the end users. Spam analysis is done with URL normalization followed by the analysis to decide whether the URL is legitimate or not.[15]

However, the drawback of this approach is that when reporting spam messages to the network, also ham domains probably contained in multi-URL messages are discredited. Meanwhile, it lacks the online process ability. Moreover, it will misjudge a lot of ham with the URL whose domain has been blacklisted before.

List Based : This is an origin-based filtering method has been used before a message is received by the receiver[16]. It is based on the network information in order to identify whether it is spam or not.

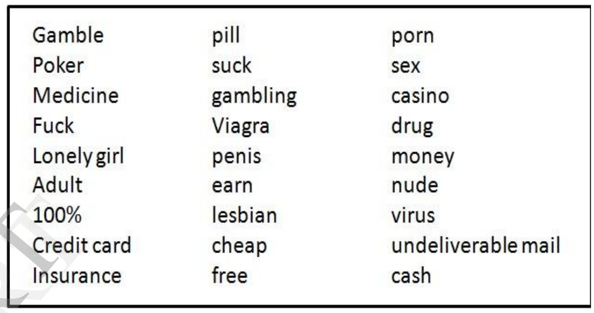
- ✓ *Blacklists:* The Blacklists are used to decline IP or TCP connections from spam originators, but also to decline mail if the domain name specified at the MAIL FROM command [17]. Blacklist Filter eliminates loads of bulk E-mail automatically by using public blacklists. Third-party blacklist service providers check the IP addresses and URLs in the configured blacklists. The IP addresses from mail headers are confirmed to find out whether the sender has tried to use a fake IP address to send the email [18].
- ✓ *Whitelists:* White-lists are used to classify users email addresses as legitimate. Emails addresses are saved within one's address book are automatically considered to be 'white-listed'. White lists can support in blocking unnecessary messages and allowing only legitimate mails, but it could not be always a correct mail. White-lists are used to decrease the occurrence of false positives.

In case of the list based techniques one can see that the size of these lists to be maintained keeps on growing as the spammers frequently change the URL links and the IP addresses and keep on hopping from one DNS to the other. Also a previously blacklisted URL may be now a ham and science of misclassifying a ham into spam could occur.

(ii) Curing Techniques

There are a lot of spammers keep on changing the URL from spam to ham and vice versa. At the same time they frequently switching the IP's so that tracing of SPAM becomes tough. In such a case where the Preventive Techniques can be time consuming and require large database to be kept of URL's etc, there are Curing Techniques. In Curing techniques the email is gauged after it arrives in the Mailbox. Once the email is there, it is filtered based on some criteria to be classified ad Ham or Spam.

Keyword Based : Keyword-based filtering is another approach for spam email filtering with effective results and is commonly employed in commercial software [19]. Here a list of spam words is maintained and the incoming emails are matched for those words. Emails containing the Spam words are classified as Spam.



Gamble	pill	porn
Poker	suck	sex
Medicine	gambling	casino
Fuck	Viagra	drug
Lonely girl	penis	money
Adult	earn	nude
100%	lesbian	virus
Credit card	cheap	undeliverable mail
Insurance	free	cash

Figure 3: A list of spam words for spam emails filtering

An advantage of Keyword based approach is that it's lightweight, fast and effective. It gives users a leverage to modify the list of spam words as per their experience or knowledge. But on the other hand one see that this type of technique is ad-hoc in nature. Here decisions are to be made so as to classify a word as spam or not first and then that word is to be relevantly searched for in the email, hence it has to make tough decisions. Another practical thing that remains unhandled here is uncertainty. Nothing in the world is certain and one need to device ways that can handle uncertainty as well.

Content Based: Content-based filtering approaches are based on the assumption and it reads the text in order to discover distinctive features which are used for sake of classifying a message [20]. It is used to analyses the headers, subject, and body of an email message using feature (token) matching or statistic methods to determine whether it is spam or legitimate email. Many content-based filtering methods have been proposed to filter spam from e-mail.

An important part in content based filtering is feature selection. There are several popular feature selection methods such as: Document Frequency Thresholding, Entropy, Term Frequency Variance, Information Gain and Odd Ratio [13]. With the increase in content filtering techniques, methods are designed and applied to automatically filter spam.

Bayesian Classification: Bayesian classification theory was derived from the Bayes's theorem in probability theory. Bayes's Theorem is a mathematical formula used for calculating conditional probabilities.

Probability-based Bayesian filtering technique is an advanced keyword filtering algorithm. It does not require pre-set rules and analysis of message content. If the calculated probability value is higher than the preset threshold, the message is classified as a spam, and treated accordingly.

Bayes' theorem : It shows how to determine inverse probabilities: knowing the conditional probability of A given B, what is the conditional probability of B given A? This can be done, but also involves the so-called prior or unconditional probabilities of A and B. Bayes' theorem relates the conditional and marginal probabilities of events *A* and *B*, provided that the probability of *B* does not equal zero:

$$\boldsymbol{P(A|B) = \frac{P(B|A).P(A)}{P(B)}}$$

- P(*A*) is the *prior* probability (or "*unconditional*" or "*marginal*" probability) of *A*. It is "prior" in the sense that it does not take into account any information about *B*; however, the event *B* need not occur after event *A*.
- P(*A*|*B*) is the conditional probability of *A*, given *B*. It is also called the posterior probability because it is derived from or depends upon the specified value of *B*.
- P(*B*|*A*) is the conditional probability of *B*, given *A*. It is also called the likelihood.
- P(*B*) is the prior or marginal probability of *B*, and acts as a normalizing constant.

In Bayesian Spam Filtering[21] , following steps are there:

1. Create database of Spam and Non Spam messages.
2. Calculate appearance rate for each independent word (in each Database).
3. Create hash tables for each. These store token to appearance rate.
4. For each incoming mail, calculate the probability of word in both hash tables.
5. Create a new hash table to store this token.
6. Use Bayes formula to calculate the probability of message being spam.
7. If the value is higher than threshold, its SPAM.

The issue faced in this is selection of the token words sometimes. Another modified approach is to use string matching along with Bayesian Classification [20,22]
The system proposed here does the follows:

- Uses private and public list to filter incoming emails based on header information.
- Text processing is done- cleaning, tokenization, stemming and drop stop words.
- Once text processing is done the message comes as a set of tokens at the matching stage.
- Here the Bayesian classifier is used to classify message into spam or ham.
- During string matching which is done in the previous step, bit parallel approximate string algorithm is used.

The string matching algorithm used here is an inexact string matching algorithm and various experimental results are thus calculated in terms of precision and recall. To improve the performance of the system further the string matching algorithm can be replaced by further more improved and advanced classifiers.

VIII . CONCLUSION

The impact of spam email has been explored in many studies, and it's clear that email spam is serious both from individual perspective as well as organizational.
Spam arrives in the mailbox in form of Viruses, Advertisements both commercial and bulk, Phishing emails, antivirus or Political mails. Since 30 years, spam is prevalent and causing trouble in daily internet based affairs leading to improper use of personal and confidential information as well as resources. In this paper various techniques that can help fight Spam like URL based filters, list based and keyword based filtering and content based filtering encompassing Bayesian filtering combined with string matching to make it more robust are discussed. With the use of these techniques and further enhancements one can fight Spam. Awareness is to be brought so that people become more alert and cautious towards email spam.

REFERENCES

1. Yanti Rosmunie Bujang , Husnayati Hussin "Should We Be Concerned with Spam Emails? A Look at Its Impacts and Implications", presented at the *5th International Conference on Information and Communication Technology, IEEE 2013*, p 01.
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-protection-infographic.pdf [Accessed: Feb 5, 2014]
5. Aayad Al Hajj, "Cyber Crimes: Threats and Protection", presented at the International Conference on Networking and Information Technology, IEEE 2010, pp 01-02
6. Cynthia Dhinakaran and Jae Kwang lee, Dhinaharan Nagamalai, ""Reminder: please update your details": Phishing Trends", presented at the First International Conference on Networks & Communications, IEEE 2009, p 01
7. Asoke K. Talukder, Vedula Bhaskar Rao, Vaibhav Kapoor and Devashish Sharma," Artificial Hygiene: A Critical Step towards Safety from Email Viruses", presented at the IEEE India Annual Conference 2004. Indicon, IEEE 2004, p 01
8. Robin Gandhi, Anup Sharma, William Mahoney, William Sousan, Qiuming Zhu, And Phillip Laplante, " Dimensions of Cyber Attack", IEEE, IEEE Technology And Society Magazine 1932-4529/11/$26.00©2011IEEE Magazine, 2011

9. ”History of Email Spam, Origin of Spam, Email Spam” Website: http://en.wikipedia.org/wiki/History_of_email_spam [Accessed: Feb 8, 2014]
10. “Famous Spammers in the World”, Website: http://www.spamhaus.org/statistics/spammers [Accessed: Feb 8, 2014]
11. 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.
12. 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
13. Hu Yin, Zhang Chaoyang, “An improved Bayesian Algorithm for Filtering Spam E-mail”, IEEE, 2011 International Symposium on Intelligence Information Processing and Trusted Computing p 02
14. Yang Li1,2, Bin-Xing Fang1, Li Guo1, Zhi-Hong Tian3, Yong-Zheng Zhang1 and Zhi-Gang Wu1, “UBSF: A Novel Online URL-Based Spam Filter” , IEEE 2008, p 01
15. Wang Zhongtao, Peng Xin, Wang Yuling, Luo Yaohua, Cai Biao, Huang Li, “Analysis on the Characteristics of URL Spam”, presented at the International Conference on Computer Science and Electronics Engineering, IEEE 2012, pp 01-04
16. Mithilesh Kumar Paswan , P. Shanthi Bala, G. Aghill, “Spam Filtering: Comparative Analysis of Filtering Techniques”, IEEE, International Conference On Advances In Engineering, Science And Management (ICAESM -2012) March 30, 31, 2012, pp 01
17. Yanhui Guo, Yaolong Zhang, Jianyi Liu, Cong Wang "Research on the Comprehensive Anti-Spam Filter", presented at the 2006 IEEE International Conference on Industrial Informatics, IEEE, 2006, p 01
18. “Spam Filter Software”, Website: http://www.spamihilator.com [Accessed: Feb 10, 2014]
19. Tak-Lam Wong, Kai-On Chow, Franz Wong, “Incorporating Keyword-Based Filtering To Document Classification For Email Spamming” , presented at the Proceedings of the Sixth International Conference on Machine Learning and Cybernetics,IEEE, Hong Kong, 19-22 August 2007, pp 01-03
20. Monther Aldwairi nad Yahya Flaifel, “Baeza Yates and Navarro Approximate String Matching for Spam Filtering “, IEEE, 2012, p 18
21. Ray Hunt and James Carpinter,” Current And New Developments In Spam Filtering”, Department of Computer Science and Software Engineering University of Canterbury, New Zealand, p 01
22. Rushdi Shams and Robert E. Mercer “Classifying Spam Emails using Text and Readability Features”, presented at the IEEE 13th International Conference on Data Mining, IEEE 2013, pp 03-04