Journal of Information Engineering and Applications www.iiste.org
ISSN 2224-5782 (print) ISSN 2225-0506 (online)
Vol.3, No.6, 2013- Selected from International Conference on Recent Trends in Applied Sciences with Engineering Applications

A Review of Spam Filtering and Measures of Antispam

Ms.Rachana Mishra
Asst.Prof.(Department IT), Oriental College of Technology, Bhopal(M.P.)
Dr..Ramjeevan Singh Thakur
Associate Prof.(Department of computer application), M.A.N.I.T., Bhopal(M.P.)
rachanamishra812@gmail.com

Abstract: Spam is commonly defined as unsolicited email messages, and the goal of spam categorization is to distinguish between spam and legitimate email messages. Our main aim is classification of spam mail and solving various problem is related to web space. So we discuss the measuring parameter which are helpful for reduce the spam or junk mail. In this paper, we describe procedure that can help eliminate unsolicited commercial e-mail, viruses, Trojans, and worms, as well as frauds perpetrated electronically and other undesired and troublesome e-mail.so this research is helpful for best classification and categorization method of email spam detection.
Keywords: spam, anti spam factor.

INTRODUCTION:

The Internet is gradually becoming an integral part of everyday life. Internet usage is expected to continue growing and e-mail has become a powerful tool intended for idea and information exchange, as well as for commercial and social lives. Along with the growth of the Internet and e-mail, there has been a dramatic growth in spam in recent years [1,2,6]. The majority of spam solutions deal with the flood of spam.However, it is amazing that despite the increasing development of anti-spam services and technologies, the number of spam messages continues to increase rapidly.

The increasing volume of spam has become a serious threat not only to the Internet, but also to society. For the business and educational environment, spam has become a security issue. Spam has gone from being annoying to being expensive and risky. The enigma is that spam is difficult to define. What is spam to one person is not necessarily spam to another. Fortunately or unfortunately, spam is here to stay and destined to increase its impact around the world. It has become an issue that can no longer be ignored; an issue that needs to be addressed in a multi-layered approach: at the source, on the network, and with the end-user. Consequently, spam filtering is able to control the problem in a variety of ways. Identification and spam removal from the e-mail delivery system allows end-users to regain a useful means of communication. Many researches on spam filtering have been centered on the more sophisticated classifier-related issues. Currently, machine learning for spam classification is an important research issue . The success of machine learning techniques in text categorization has led researchers to[1] explore learning algorithms in spam filtering. In particular, Bayesian techniques, support vector machines (SVM) effectively used for text categorization which influences researchers to classify the email is based on a special case of TC (text categorization), with the categories being spam and non-spam.

SPAM PROBLEMS

Spam is difficult to define. In fact, there is no widely agreed or clear workable definition, We all recognize spam when we see it, but the truth is that what is spam to one person may not be spam to another. So, the notion of spam is subjective [1,3]. The most well-known definition seems to be 'unsolicited commercial email' (UCE) and 'Unsolicited Bulk Email' (UBE). The content of spam ranges enormously from advertisements for goods and services to pornographic material, financial advertisements, information on illegal copies of software, fraudulent advertisements and/or fraudulent attempts to solicit money. More recently, spam has been spreading at an increasingly rapid rate, and while groups of spammers were relatively small in the past, the wide availability of 'spam kits' over the Internet has spread the practice from the United States to China, Russia and South America .[3] The scale of the problem is perhaps best highlighted when the growth of spam since 2001 is considered, as well as the percentage of spam, which was 7 per cent of all received e-mail . By 2002 this had grown to 29 per cent, and by the end of 2003 the total stood at 54 per cent . In March 2004 the percentage had increased to 63 per cent and this is set to continue rising. According to Message Labs, a US consultancy firm, spam now accounts for around 65% of all e-mail traffic. The following subsections outline a number of reasons which can explain why spam has become a serious problem.

"SPAM" A Serious Problem

"Cost-Shifting"

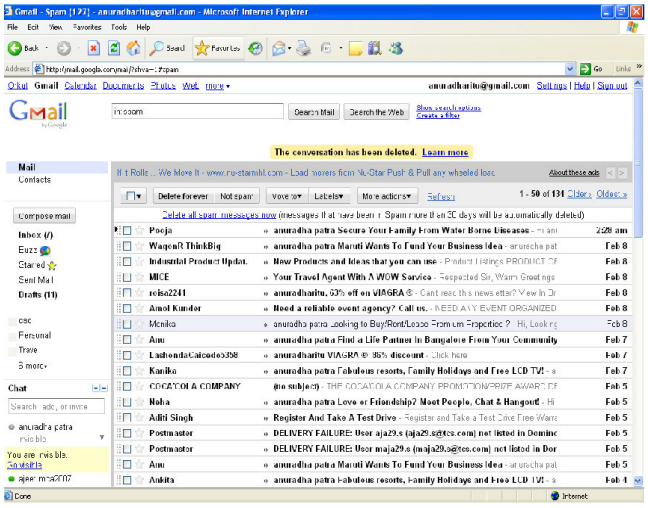
Sending spam is extremely cheap (for the sender).The costs of spamming are paid by others: network mainainers,recipients,etc."Fraud" Often,spams pretend to be replies or follow-ups to previous inqiuiries to get

people into opening messages."Disguise origin" Spammers can easily disguise the originof their spam messages.Otherwise it would be just too easy to filter spam and spamming would be rendered useless.

Problem formulation

The problem with spam is that it tends to swamp desirable e-mail. In my own experience, a few years ago I occasionally received an inappropriate message, perhaps one or two each day. Every day of this month, in contrast, I received many times more spams than I did legitimate correspondences. On average, I probably get 10 spams for every appropriate e-mail. In some ways I am unusual -- as a public writer, I maintain a widely published e-mail address; moreover, I both welcome and receive frequent correspondence from strangers related to my published writing and to my software libraries. Unfortunately, a letter from a stranger -- with who-knows-which e-mail application, OS, native natural language, and so on, is not immediately obvious in its purpose; and spammers try to slip their messages underneath such ambiguities. My seconds are valuable to me, especially when they are claimed many times during every hour of a day. All mail server are not control spam mail, so that I can try to do some more work on spam data analysis, and give some accurate point to control them.

Sample data of spam:



ANTI –SPAM MEASURES

Why finding out secondary impacts of spam filtering is so difficult

From a spam filtering point of view we can use two main information resources for studying secondary impacts of spam filtering. First, we can analyze block lists and the IP addresses they list. Technically this is rather straightforward and we will cover results in the next section. Second, we can analyze spam messages and where they (allegedly) come from. In addition, it is necessary to take into account information from other, non-technical sources to understand how spamming is affecting the networked world.

Analyzing email in the context of spam filtering: For a number of reasons including avoiding legal liability and protecting their valuable infrastructure, spammers try to disguise the true origin of their messages. Naturally this behavior means that collecting data regarding secondary, unwanted effects of anti-spam measures beyond false-positive rates (which are computed locally) is rather difficult.

Identifying fake Received headers in email

For example imagine that we receive an email with header information including the following (for a more detailed discussion see 2006):

Received:	from	Echo	by	Destination
Received:	from	Delta	by	Echo
Received:	from	Charlie	by	Delta
Received:	from	Bravo	by	Charlie

Received: from Alpha by Bravo

Type of spam

Text spam:the capability to sort incoming emails based on simple string found in specific header fields.its capability is very simple and doesn't even include regular expression matching.in text spam we filter the particular keyword which are unwanted.

Image spam: Image spam is a kind of E-mail spam where the message text of the spam is presented as a picture

in an image file. Since most modern graphical E-mail client software will render the image file by default, presenting the message image directly to the user, it is highly effective at circumventing normal E-mail filtering software. Currently, the surest known countermeasure for image spam is to discard all messages containing images which do not appear to come from an already white listed E-mail address. However, this has the disadvantage that valid messages containing images from new correspondents must either be silently discarded.Content based spam: involves a number of methods, such as repeating unrelated phrases, to manipulate the relevance or prominence of resources indexed by a search engine, in a manner inconsistent with the purpose of the indexing system.

How to filter the spam:

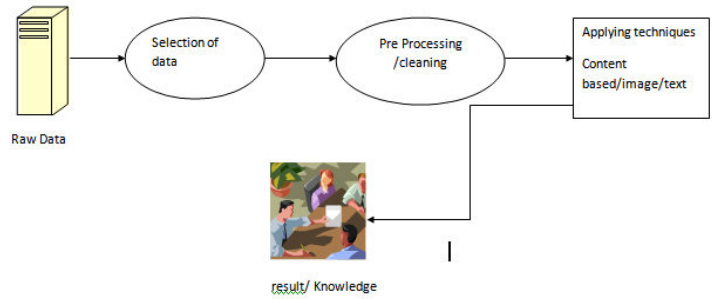
Step 1 : collect the spam and anti spam data.

Step2: preprocessing the data in which we have to reduce the noise,means we reduces the unwanted field.we have to clean the data .

Step 3: identify the specific bad sender, read the header body.

Step 4: applying particular classification tools or techniques for specified content, text and images. We have to apply these technique as per our requirement .

Step 5: store the result in the form of content based or text based or image based.



Techniques for anti spam filtering:

White /list filter

Rule based filter

Content based filter

Pattern detection

In our study we conclude that spam can be classified into two wide categories .The first categories is spam with attachment & the second categories is spam without attachment .Spam with attachment into four type such as spam mail containing image file(.gif),text message,URL,spam containing image, text and spam containing worms,virus,Trojans as an attachment. Spam without attachment are small in size but spam with attachment bigger in size.our aim is to introduce the survey spam techniques.

Conclusion

Spam measures are helpful for categorization of spam mail. Our main aim is finding best filtering techniques for spam detection, so given measures are helpful for features selection . We can conclude that number of SPAM is a very serious problem which is drastically widespread. In future ,we plan to propose and implement a new spam classification and spam detection approach .

References:

[1] Fulu Li, Mo-han Hsieh, “An empirical study ofclustering behavior of spammers and Group based Anti-spam strategies”, CEAS 2006, pp 21-28, 2006.

[2] Dhinaharan Nagamalai, Cynthia.D, Jae Kwang Lee,” ANovel Mechanism to defend DDoS attacks caused by spam”,
International Journal of Smart Home, SERSC, Seoul, July 2007, pp 83-96.

[3] Calton pu, Steve webb: “Observed trends in spam construction techniques: A case study of spam evolution”, CEAS 2006, pp 104-112, July 27-28, 2006.

[4] Anirudh Ramachandran, David Dagon, Nick Feamste,“Can DNS-based Blacklists keep up with Bots”, CEAS 2006,CA, USA, July 27-28, 2006.

[5] SpamCop http://spamcop.net.

[6] Internet User Forecasts by Country http:// www. etforecasts.Com.

[7] Nigerian fraud mail Gallery http://www.potifos.com/fraud/.

[8] Fairfax Digital http://www.smh.com.au/articles/2004/10/18.