Available at: www.researchpublications.org

Spam Filtering Methods for Email Filtering

Akshay P. Gulhane
Final year B.E. (CSE)
P.R. M. I. T. & R.,
Amravati, Maharashtra, India
E-mail: akshaygulhane91@gmail.com

Shraddha A. Jalan
Final year B.E. (CSE)
P.R. M. I. T. & R.,
Amravati, Maharashtra, India
E-mail: shraddha.jalan10@gmail.com

Sakshi Gudadhe
Third year B.E. (CSE)
P.R. M. I. T. & R.,
Amravati, Maharashtra, India
E-mail: gudadhe.sakshi25@gmail.com

Ajinkya R. Bijwe
Final year B.E. (CSE)
P.R. M. I. T. & R.,
Amravati, Maharashtra, India
E-mail: ajinkya.bijwe@gmail.com

Abstract

Millions of people uses Electronic mail to communicate around the globe is a very important Application for many businesses. Over the last decade, unsolicited bulk email has become a major problem for email users. A myriad amount of spam is sent to users' mailbox daily. Spam not only frustrates the most of the Email users but also strains the IT infrastructure of Organization and costs billions of Dollars in lost productivity. The necessity of the effective Spam Filters increases day by day.

The Spam filtering approaches constantly face new evasion techniques attempted by the spammers. For example, text-based approaches, including those using Bayesian classifiers on email messages, may be evaded by sending text in images or minimizing the text of an ad and shifting most details to a web site by adding a link. The links (URLs) spammers typically use as a feedback mechanism present perhaps the only piece of "immutable" information in a spam message because each URL must be precisely, spelled out to link to a Web page1.

As spammers continue to develop new techniques to evade text-based filtering, combinations of spam filtering solutions are likely to be employed. Our approach can make such combined filtering solutions more accurate because it relies on an additional information source making it harder for spammers to evade classification. In this Paper, we presented our study on various methods used to filter the Spam.

Keywords: Spam, Email Spam's, Spam Filters.

1. Introduction

Electronic mail is a method of exchanging digital messages from an author to one or more recipients. Modern email operates across the Internet or other computer networks. Email spam is a subset of electronic spam involving nearly identical messages sent to numerous recipients by email. Clicking on links in spam email may send users to phishing web sites or sites that are hosting malware. Spam email may also include malware as scripts or other executable file attachments. Definitions of spam usually include the aspects that email is unsolicited and sent in bulk. Most spam filtering approaches in use today rely on the text of messages for classification. In many cases, these approaches can be easily evaded by spammers. For example, a spammer can shift most of the details of an ad in a message to a Web site, to evade filters. The resulting spam message can be very small, consisting of neutral words (e.g. the ones that commonly occur in legitimate e-mail for most users) and a link to a web site.

Email spam is a subset of electronic spam involving nearly identical messages sent to numerous recipients by email. Clicking on links in spam email may send users to phishing web sites or sites that are hosting malware. Spam email may also include malware as scripts or other executable file attachments. Definitions of spam usually include the aspects that email is unsolicited and sent in bulk.

Spammers can also put ads in images. A good spam filtering approach must not only have significantly low false positive and negative rates, but also take into account that it is far worse to misclassify a legitimate message than spam. [1, 2]

2. Spam

Email spam, also known as junk email or unsolicited bulk email (UBE), is a subset of electronic spam involving nearly identical messages sent to numerous recipients by email. Clicking on links in spam email may send users to phishing web sites or sites that are hosting malware. Spam email may also include malware as scripts or other executable file attachments. Definitions of spam usually include the aspects that email is unsolicited and sent in bulk. One subset of UBE is UCE (unsolicited commercial email). The opposite of "spam", email which one wants, is called "ham", usually when referring to a message's automated analysis (such as Bayesian filtering). Like other forms of unwanted bulk messaging, it is named for Spam luncheon meat by way of a Monty Python sketch in which Spam is depicted as ubiquitous and unavoidable.

Email spam has steadily grown since the early 1990s. Botnets, networks of virus-infected computers, are used to send about 80% of spam. Since the expense of the spam is borne mostly by the recipient, it is effectively postage due advertising.

ISPs have attempted to recover the cost of spam through lawsuits against spammers, although they have been mostly unsuccessful in collecting damages despite winning in court.

Spammers collect email addresses from chat-rooms, websites, customer lists, newsgroups, and viruses which harvest users' address books, and are sold to other spammers. They also use a practice known as "email appending" or "e-pending" in which they use known information about their target (such as a postal address) to search for the target's email address. Much of spam is sent to invalid email addresses. Spam averages 78% of all email sent. According to the Message Anti-Abuse Working Group, the amount of spam email was between 88–92% of email messages sent in the first half of 2010. [6]

Fig. 2.1 Spam Folder in Mail Inbox

3. Email Filtering

Email filtering is the processing of email to organize it according to specified criteria. Most often this refers to the automatic processing of incoming messages, but the term also applies to the intervention of human intelligence in addition to anti-spam techniques, and to outgoing emails as well as those being received.

For its output, it might pass the message through unchanged for delivery to the user's mailbox, redirect the message for delivery elsewhere, or even throw the message away. Some mail filters are able to edit messages during processing.

Common uses for mail filters include organizing incoming email and removal of spam and computer viruses. A less common use is to inspect outgoing email at some companies to ensure that employees comply with appropriate laws. Users might also employ a mail filter to prioritize messages, and to sort them into folders based on subject matter or other criteria.

Mail filters can be installed by the user, either as separate programs, or as part of their email program (email client). In email programs, users can make personal, "manual" filters that then automatically filter mail according to the chosen criteria. Most email programs now also have an automatic spam filtering function. Internet service providers can also install mail filters in their mail transfer agents as a service to all of their customers. Due to the growing threat of fraudulent websites Internet service providers filter URLs in email messages to remove the threat before users click. Corporations often use filters to protect their employees and their information technology assets.

Mail filters can operate on inbound and outbound email traffic. Inbound email filtering involves scanning messages from the Internet addressed to users protected by the filtering system or for lawful interception. Outbound email filtering involves the reverse - scanning email messages from local users before any potentially harmful messages can be delivered to others on the Internet. One method of outbound email filtering that is commonly used by Internet service providers is transparent SMTP proxying, in which email traffic is intercepted and filtered via a transparent proxy within the network. Outbound filtering can also take place in an email server. Many corporations employ data leak prevention technology in their outbound mail servers to prevent the leakage of sensitive information via email. [1, 2]

4. Spam Filtering

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, too often omitting perfectly legitimate messages (these are called *false positives*) and letting actual spam through. More sophisticated programs, such as Bayesian filters or other heuristic filters, attempt to identify spam through suspicious word patterns or word frequency.

4.1 Techniques used for Spam Filtering

There are various spam filtering techniques used now a days. Few are explained below.

4.1.1 List Based Filtering

List-based filters attempt to stop spam by categorizing senders as spammers or trusted users, and blocking or allowing their messages accordingly.

➢ **Blacklist**

This popular spam-filtering method attempts to stop unwanted email by blocking messages from a preset list of senders that you or your organization’s system administrator creates. Blacklists are records of email addresses or Internet Protocol (IP) addresses that have been previously used to send spam. When an incoming message arrives, the spam filter checks to see if it’s IP or email address is on the blacklist; if so, the message is considered spam and rejected.

Though blacklists ensure that known spammers cannot reach users' inboxes, they can also misidentify legitimate senders as spammers. These so-called false positives can result if a spammer happens to be sending junk mail from an IP address that is also used by legitimate email users. Also, since many clever spammers routinely switch IP addresses and email addresses to cover their tracks, a blacklist may not immediately catch the newest outbreaks.

➢ **Real-Time Blackhole List**

This spam-filtering method works almost identically to a traditional blacklist but requires less hands-on maintenance. That’s because most real-time Blackhole lists are maintained by third parties, who take the time to build comprehensive blacklists on the behalf of their subscribers. Your filter simply has to connect to the third-party system each time an email comes in, to compare the sender’s IP address against the list.

Since Blackhole lists are large and frequently maintained, your organization's IT staff won't have to spend time manually adding new IP addresses to the list, increasing the chances that the filter will catch the newest junk-mail outbreaks. But like blacklists, real-time Blackhole lists can also generate false positives if spammers happen to use a legitimate IP address as a conduit for junk mail. Also, since the list is likely to be maintained by a third party, you have less control over what addresses are on — or not on — the list.

➢ **Whitelist**

A Whitelist blocks spam using a system almost exactly opposite to that of a blacklist. Rather than letting you specify which senders to block mail from, a Whitelist lets you specify which senders to allow mail from; these addresses are placed on a trusted-users list. Most spam filters let you use a Whitelist in addition to another spam-fighting feature as a way to cut down on the number of legitimate messages that accidentally get flagged as spam. However, using a very strict filter that only uses a Whitelist would mean that anyone who was not approved would automatically be blocked.

Some anti-spam applications use a variation of this system known as an automatic Whitelist. In this system, an unknown sender's email address is checked against a database; if they have no history of spamming, their message is sent to the recipient's inbox and they are added to the Whitelist.

➢ **Greylist**

A relatively new spam-filtering technique, Greylist take advantage of the fact that many spammers only attempt to send a batch of junk mail once. Under the Greylist system, the receiving mail server initially rejects messages from unknown users and sends a failure message to the originating server. If the mail server attempts to send the message a second time — a step most legitimate servers will take — the Greylist assumes the message is not spam and lets it proceed to the recipient's inbox. At this point, the Greylist filter will add the recipient's email or IP address to a list of allowed senders.

Though Greylist filters require fewer system resources than some other types of spam filters, they also may delay mail delivery, which could be inconvenient when you are expecting time-sensitive messages. [7, 8]

4.1.2 Content Based Filters

Rather than enforcing across-the-board policies for all messages from a particular email or IP address, content-based filters evaluate words or phrases found in each individual message to determine whether an email is spam or legitimate.

➢ Word-Based Filters

A word-based spam filter is the simplest type of content-based filter. Generally speaking, word-based filters simply block any email that contains certain terms.

Since many spam messages contain terms not often found in personal or business communications, word filters can be a simple yet capable technique for fighting junk email. However, if configured to block messages containing more common words, these types of filters may generate false positives. For instance, if the filter has been set to stop all messages containing the word "discount," emails from legitimate senders offering your nonprofit hardware or software at a reduced price may not reach their destination. Also note that since spammers often purposefully misspell keywords in order to evade word-based filters, your IT staff will need to make time to routinely update the filter's list of blocked words.

➢ Heuristic Filters

Heuristic (or rule-based) filters take things a step beyond simple word-based filters. Rather than blocking messages that contain a suspicious word, heuristic filters take multiple terms found in an email into consideration.

Heuristic filters scan the contents of incoming emails and assigning points to words or phrases. Suspicious words that are commonly found in spam messages, such as "Rolex" or "Viagra," receive higher points, while terms frequently found in normal emails receive lower scores. The filter then adds up all the points and calculates a total score. If the message receives a certain score or higher (determined by the anti-spam application's administrator), the filter identifies it as spam and blocks it. Messages that score lower than the target number are delivered to the user.

Heuristic filters work fast — minimizing email delay — and are quite effective as soon as they have been installed and configured. However, heuristic filters configured to be aggressive may generate false positives if a legitimate contact happens to send an email containing a certain combination of words. Similarly, some savvy spammers might learn which words to avoid including, thereby fooling the heuristic filter into believing they are benign senders. [5]

➢ Bayesian Filters

Bayesian filters, considered the most advanced form of content-based filtering, employ the laws of mathematical probability to determine which messages are legitimate and which are spam. In order for a Bayesian filter to effectively block spam, the end user must initially "train" it by manually flagging each message as either junk or legitimate. Over time, the filter takes words and phrases found in legitimate emails and adds them to a list; it does the same with terms found in spam.

To determine which incoming messages are classified as spam, the Bayesian filter scans the contents of the email and then compares the text against its two-word lists to calculate the probability that the message is spam. For instance, if the word "valium" has appeared 62 times in spam messages list but only three times in legitimate emails, there is a 95 percent chance that an incoming email containing the word "valium" is junk.

Because a Bayesian filter is constantly building its word list based on the messages that an individual user receives, it theoretically becomes more effective the longer it's used. However, since this method does require a training period before it starts working well, you will need to exercise patience and will probably have to manually delete a few junk messages, at least at first. [6, 7]

5.1.3 Other Filtering Methods

In addition to list- and content-based filtering techniques, some anti-spam applications employ one or more additional methods.

➢ Challenge/Response System

Filters that use a challenge/response system block undesirable emails by forcing the sender to perform a task before their message can be delivered. For instance, if you send an email to someone who's using a challenge/response filter, you'll likely receive an email right back that asks you to visit a Web page and enter the code displayed their in a form. If you successfully complete this task, your email (and all future emails) will be delivered to the recipient. If you don't complete the challenge after a certain time period, the message is rejected.

This system works to fight spam because the "challenge" is typically only one that a human can solve. Spammers usually rely on automated mailing programs to send out millions of emails at once, and they rarely check to see what emails come back in response. And even if they did see a challenge message, they aren't likely to respond and risk revealing themselves as a spammer.

However, challenge/response filters might also block email newsletters you subscribe to, as these messages are typically sent by automated programs. Another downside is that some of your organization's constituents may not take the time to complete the challenge or may not understand the challenge email, meaning that their messages will not reach the recipient. And there's always the slight chance that if both the sender and recipient are using challenge/response systems, their anti-spam applications will continue to challenge each other, locking the email in an undeliverable loop. [4, 5, 7]

- **Collaborative Filters**

Collaborative content filtering takes a community-based approach to fighting spam by collecting input from the millions of email users around the globe. Users of these systems can flag incoming emails as legitimate or spam and these notations are reported to a central database. After a certain number of users mark a particular email as junk, the filter automatically blocks it from reaching the rest of the community's inboxes.

When a collaborative content filtering system involves a large, active user base, it can quickly quell a spam outbreak, sometimes within a matter of minutes. One potential downside to the collaborative-content method is that if a group of spammers mobilize in large numbers and pretend to be legitimate users of the system, they could skew results by falsely labeling spam emails as legitimate messages. [5]

- **DNS Lookup Systems**

While not a particularly reliable method on its own, several anti-spam methods use the domain name system (DNS) — which all mail servers on the Internet use to identify themselves — to identify and foil spammers.

DNS Mail Exchange (MX) attempts to verify that the domain name in the email address of the sender — the part after the “at” symbol (@) — exists. It does this by searching the domain name system to see whether the domain name has a valid MX record, which indicates the presence of a real mail server; if there's no match, the anti-spam program assumes that the message is junk. A filter will also perform a reverse DNS lookup using the IP address off the mail server that sent the questionable message. This lookup will reveal the domain name associated with the server.

While DNS lookups can be useful in weeding out emails from spammers attempting to disguise themselves, they are not as effective or reliable on their own (when compared to other spam-fighting methods) in stopping general junk mail. In particular, reverse DNS lookups have been known to produce false positives — legitimate messages marked as spam — since it's technically possible that legitimate senders can send email from a domain different from their own. [2, 4, 5]

5. Bayesian Filter

Bayesian spam filtering is a statistical technique of e-mail filtering. It makes use of a naive Bayes classifier to identify spam e-mail. Bayesian classifiers work by correlating the use of tokens (typically words, or sometimes other things), with spam and non-spam e-mails and then using Bayesian inference to calculate a probability that an email is or is not spam.

Bayesian spam filtering is a very powerful technique for dealing with spam, that can tailor itself to the email needs of individual users, and gives low false positive spam detection rates that are generally acceptable to users.

Particular words have particular probabilities of occurring in spam email and in legitimate email. For instance, most email users will frequently encounter the word "Viagra" in spam email, but will seldom see it in other email. The filter doesn't know these probabilities in advance, and must first be trained so it can build them up. To train the filter, the user must manually indicate whether a new email is spam or not. For all words in each training email, the filter will adjust the probabilities that each word will appear in spam or legitimate email in its database. For instance, Bayesian spam filters will typically have learned a very high spam probability for the words "Viagra" and "refinance", but a very low spam probability for words seen only in legitimate email, such as the names of friends and family members.

After training, the word probabilities (also known as likelihood functions) are used to compute the probability that an email with a particular set of words in it belongs to either category. Each word in the email contributes to the email's spam probability, or only the most interesting words. This contribution is called the posterior probability and is computed using Bayes' theorem. Then, the email's spam probability is computed over all words in the email, and if the total exceeds a certain threshold (say 95%), the filter will mark the email as a spam.

As in any other spam filtering technique, email marked as spam can then be automatically moved to a "Junk" email folder, or even deleted outright. Some software implements quarantine mechanisms that define a time frame during which the user is allowed to review the software's decision.

The initial training can usually be refined when wrong judgments from the software are identified (false positives or false negatives). That allows the software to dynamically adapt to the ever evolving nature of spam.

Some spam filters combine the results of both Bayesian spam filtering and other heuristics (pre-defined rules about the contents, looking at the message's envelope, etc.), resulting in even higher filtering accuracy, sometimes at the cost of adaptiveness.

Bayesian email filters take advantage of Bayes' theorem. Bayes' theorem is used several times in the context of spam:

- A first time, to compute the probability that the message is spam, knowing that a given word appears in this message;
- A second time, to compute the probability that the message is spam, taking into consideration all of its words (or a relevant subset of them);
- Sometimes a third time, to deal with rare words. [4,5,7]

6. Conclusion

We discussed the basic definitions regarding spam filter and various techniques to reduce the spam in Email. One of the most widely used spam filtering techniques, i.e., The Bayesian technique was also discussed in the seminar.

Spam or unsolicited e-mail has become a major problem for companies and private users. This paper explored the various problems associated with spam and different methods and techniques attempting to deal with it. From the study we identified that, many of the filtering techniques are based on text categorization methods and there is no technique can claim to provide an ideal solution with 0% false positive and 0% false negative. There is lot of scope for research in classifying text messages as well as multimedia messages.

7. References

[1] Christina V, Karpagavalli S, *A Study on Email Spam Filtering Techniques,* International Journal of Computer Applications (0975 – 8887) Volume 12– No.1, December 2010

[2] Microsoft Corporation. Exchange intelligent message filter. *http://www.microsoft.com/exchange/techinfo/security/imfoverview.asp*, 2003.

[3] Mason J. et al Sergeant M. Spamassassin presentations. *http://eu.spamassassin.org/presentations.html*, 2003.

[4] V.V.Prakash. Vipuls razor documentation: Collaborative filtering.*http://razor.sourceforge.net/docs*, 2004.

[5] M. Sahami, S. Dumais, D. Heckerman, and E. Horvitz. A bayesian approach to filtering junk e-mail. In *AAA1 Workshop pp. 55-62*, 1998.

[6] Vikas P. Deshpande, Robert F. Erbacher, and Chris Harris, *An Evaluation of Naïve Bayesian Anti-Spam Filtering Techniques*, Proceedings of the 2007 IEEE Workshop on Information Assurance United States Military Academy, West Point, NY 20-22 June 2007

[7] A white paper on *"Why Bayesian filtering is the most effective anti-spam technology"* By GFI Software, Inc.

[8] David Mertz, *"Comparing a Half-Dozen Approaches to Eliminating Unwanted Email"*, August 2002