# A New Probability based Analysis for Recognition of Unwanted Emails

Shashi Kant Rathore
Department of Computer Science & Engineering,
Lovely Professional University, Jalandhar, Punjab

Palvi Jassi
Department of Computer Science & Engineering,
Lovely Professional University, Jalandhar, Punjab

Basant Agarwal
Department of Computer Science & Engineering,
Central University of Rajasthan, Kishangarh, Ajmer

## ABSTRACT
Electronic mail is used as a mean for personal and business communication. The volume of unwanted messages or mails that are received is growing as well. Cost of sending this type of Email is very low for sender, so several people and companies use it to quickly distribute unsolicited bulk messages, also called spam, to a large number of recipients. The reasons for sending spam vary and may include marketing of products and services. Moreover, many people uses spam as a medium for attacks and distributing harmful content such as viruses, trojan horses, worms and other malware. Spam has become a major threat for business users, network administrators and even ordinary users. In addition to regulations, several technical solutions including commercial and open source products have been proposed and deployed to block this problem. In this work proposed and implement mechanism for block spam mails by implementing anti spam filters at the network gateway.

## Keywords
Spam filtering; text categorization; machine learning; legitimate emails; unsolicited commercial e-mail; spam.

## 1. INTRODUCTION
At the moment of writing emails are widely used in our social or professional life, most email systems are based on SMTP (Simple Mail Transfer Protocol) [13]. It is used for standard mechanism for transporting emails among different hosts over the internet. But the major parts of all emails that are received are unsolicited (spam). It will decrease the usefulness of email. Spam is become the primary threat to the survival of e-mail as a useful communication medium. The percentage of spam in mail traffic in 2009 came to an average of 85.2%, or 3.1% higher than in 2008. The highest percentage of spam recorded was 93% on 22 February, while the low for the year was 72.8% on 26 April. At this time several organizations are being formed to fight the war against spam. Organizations make it possible to create solutions in a structured manner. A large amount of software is being developed to stop spam, from which majorities are to make filters. Despite the availability of solutions to spam, users are unable to use them. This is primarily due to the lack of transparency and relatively difficult use of the solutions.

Recently, there are strong demands for preventing spam emails. There was developed many more techniques for blocking spam, almost all ways cannot be proper effective for the administration of email system. The major problem is that every email must be received by email server. Therefore, it is difficult to shut all spam senders perfectly. However, it accidentally can delete or reject even legitimate email. From those reason, we approach spam from a different aspect. Even our proposed system which is working as a kind of proxy server detects spam, it changes its state and here it blocks the spam. Otherwise it relay the normal email quickly as the normal mode. The proposed system can guarantee that legitimate emails are transferred.

### 1.1 Working of e-mail system
Most email systems are based on SMTP (Simple Mail Transfer Protocol). It is used the standard mechanism for transporting emails among different hosts over the internet [3].

The SMTP operation of transferring message is constructing in 4 steps:

1. Starting session
2. Confirming the domains and a sender address in the envelope
3. Sending the message
4. Quitting the session

Step 2 and 3 are independent of step 1, SMTP client can transfer any number of message in one session [17].

### 1.2 How spam comes
Most of emails are transferred by the following two ways that is directly or indirectly [8].

1. Spam sender directly connects to target SMTP server. Then it transfers spam. It is defined as direct connection.
2. Spam sender use the open rely SMTP server for multi hop systems. The spam sender can use any number of SMTP clients to transmit lots of spam simultaneously. It is defined as indirect connections.

## 2. VARIOUS FILTERING METHODOLOGIES
To prevent e-mail spam, both end users and administrators of email systems use various anti spam techniques. No one technique is a complete solution to the spam problem, and each has disadvantages between incorrectly rejecting legitimate email vs. not rejecting all spam, and the associated costs in time and effort. Following methods are commonly used for blocking of spam mails.

### 2.1 Memory based filtering
This method include text categorization or memory based (or instance based) methods. In this technique they store all incoming messages in a memory structure, and use them directly for classification. Classification is usually performed through a variant of the basic k- Nearest-Neighbor (k-NN) algorithm [18]. There are also some disadvantages, like it focuses on keyword

similarity. This approach is incapable of capturing more complex relationships at a deeper semantic level [18].

## 2.2 Machine learning based classification

Apart from collecting separately spam and legitimate training messages, the learning process is fully automatic. When a mail is reported as a spam, the contents of mail are automatically added to the spam database. There are also some disadvantages, like anti spam filtering differs from other electronic mail and news categorization tasks, in that spam messages cover a very wide spectrum of topics, and hence are much less homogeneous [17].

## 2.3 Bayesian spam filtering

Naive Bayes classifier [6] is a simple probabilistic classifier based on applying Baye's theorem with strong (naive) independence assumptions. Baye's classifier assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature. Bayesian spam filtering (a form of e-mail filtering) is the process of using a Naive Bayesian classifier to identify spam email. Bayesian spam filtering is susceptible to Bayesian poisoning, a technique used by spammers in an attempt to degrade the effectiveness of spam filters that rely on Bayesian filtering [3][6][9].

## 2.4 Checksum based filtering

Checksum based filter exploits the fact that the messages are sent in bulk, that is they will be identical with small variations. Checksum based filters strip out everything that might vary between messages, reduce what remains to a checksum, and look that checksum up in a database which collects the checksums of messages that email recipients consider to be spam. The disadvantage is that spammers can insert unique invisible gibberish known as hash buster [2] into the middle of each of their messages, thus making each message unique and having a different checksum [14].

# 3. METHODOLOGY ADOPTED

Here we will discuss the methodology which is adopted by proposed system to solve the problem of blocking spam. The first section and remaining part of this section will discuss the method we have used. And remaining part of this section will describe introduction of system.

The proposed system should be large such like an ISP. The email system should transfer huge amount of emails to its users. And all emails should pass through the proposed system. It means the system plays a role as a proxy server for email [1].

Proxy server receives a request message from SMTP client (sender side). It established a connection between SMTP server process and Proxy server, and passes the connection to a sub process. Now this sub process request a new connection to the SMTP server on the controlled domain and then established a connection between Proxy server and SMTP server (Controlled domain). On the sub process system judges that received Email is spam or not. If it is not a spam, the module quickly transfer it to SMTP Server, otherwise the process reject it [1].

The filters relied on a probabilistic method known as Bayesian filtering [9]. The system works with database of words, which are mostly comes in spam emails. These words are taken from the various spam details sites. Then we checked each word of the incoming email with this database, and then find out the local probability of each email.

Firstly filter tokenizes the whole email in small words. The individual probabilities of each word appearing in a spam are independent of one another. The overall probability that the new e-mail is a spam is then computed as following method. For two tokens with probabilities a and b the combined probability is computed as

The combined probabilities for three tokens with probabilities a, b, c would be computed:

$$p = \frac{abc}{abc - (1-a)(1-b)(1-c)}$$

And so on [3].

This approach is an extension of text classification technology, which searches the textual content of an email and uses algorithms to identify spam email. The algorithms are able to classify the occurrence of certain words and phrases in terms of how and where they appear in the email, not by their existence alone. And generate the probability of each word.

If this probability is more than a certain level (predefined). Then that email is considered as a spam. Otherwise it is not spam.

When proxy receives a connection from SMTP client, it creates a new sub process to deal with the SMTP session. At this time spam detection module judges whether the IP address is considered as spam server or not. The sub process store the information for, then check it to detect. In each case, when spam detection module detects spam, the sub process rejects connection to that SMTP client.

# 4. EVALUATION

Now this section describes the construction of proposed system. Figure shows the simple layout of proposed system. In which a proxy server is there for controlling transfer of emails passing between two mail servers. There are one proxy server, single SMTP server (Receiver side) and four SMTP Server (Sender side). These details of all systems are given below.

1. Proxy server: - This machine is running with FEDORA 7.1. and has an IP address 192.168.164.128.The program of this proxy server is implemented in C language.
2. SMTP Server (receiver side):- This machine is running with FEDORA 7.1. And has an IP address 192.168.164.129.
3. SMTP Server 1(sender side):- This machine is running with FEDORA 7.1. and has an IP address 192.168.164.130
4. SMTP Server 2(sender side):- This machine is running with FEDORA 7.1. And has an IP address 192.168.164.131.
5. SMTP Server 3(sender side):- This machine is running FEDORA 7.1. and has an IP address 192.168.164.132

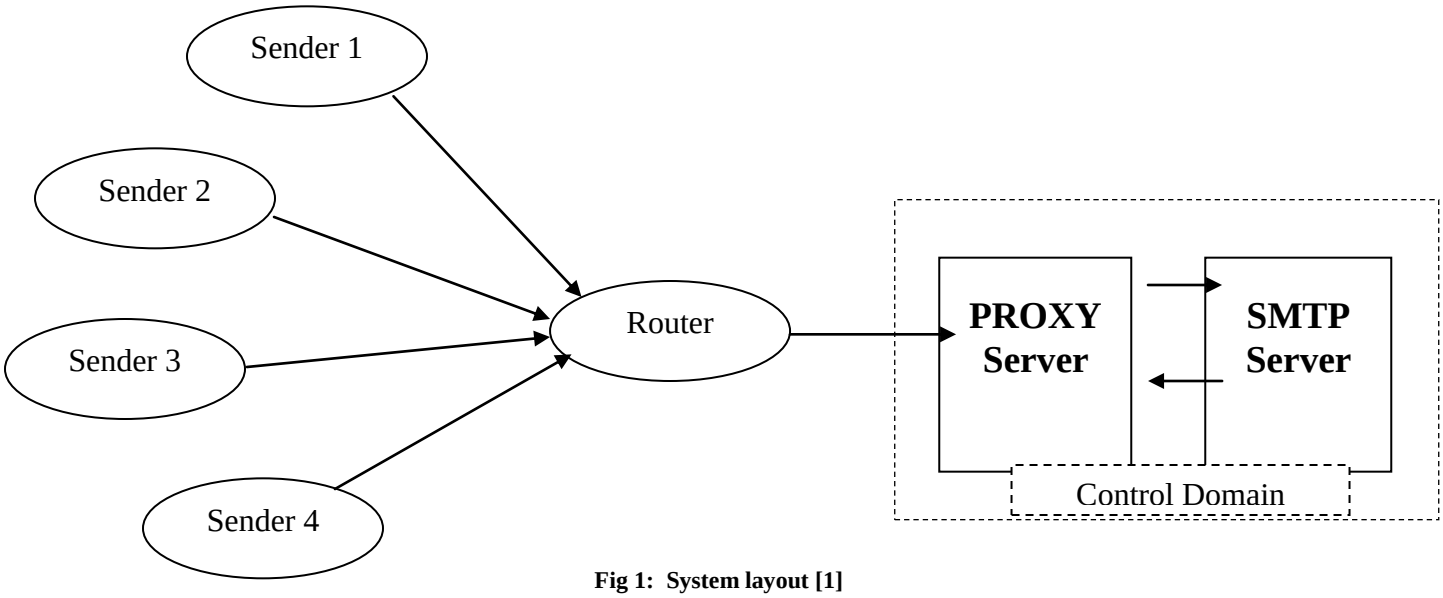


**Fig 1: System layout [1]**

## 5. GENERATION OF E-MAILS

The two types of emails are generated: one is the legitimate email, and other is spam mails. The legitimate email is generated with obeying the following rules[7]:-

1. The "subject" consists of random 64 strings.
2. The "body" consists of random 512 strings.

On the other hand, spam from a certain sender is assumed to be same format. They are an advertisement for a certain services or goods. Since sender transfers a lot of spam to an ISP at a time, SMTP server in the controlled domain receives them. Therefore in this experiment, spam emails are generated with obeying the following rules.

1. The "subject" consists of constant strings.
2. The "body" consists of constant strings.

On proxy server each mail is checked for spam. It finds out the probability of whole email by comparing each word to its database. If the probability of whole mail is greater then a certain level then it is declare as a spam. In experiment we send 5-5 emails in each group. And note down the results of both legitimate emails and spam emails. Here we assume that probability greater then 0.7 necessary to be spam.

## 6. RESULTS

This section shows the various results when legitimate or spam mails are sent by senders to receivers.

### 6.1 Results when legitimate emails are sent

Table 1 shows the results, when the legitimate mails are sent. In experiment legitimate emails are sent in groups and result is showing details of each group of 5 emails. Here is the some of results from them.

**Table 1: Results when legitimate E-mails are sent**

| Email groups | Over all probability | Result | True Negative | False Negative |
|---|---|---|---|---|
| 1st Group of 5 emails | .67 | 2 mails are SPAM | 3 Mails | 2 Mails |
| 2nd group of 5 emails | .49 | 0 mails are SPAM | 5 Mails | 0 Mails |
| 3rd group of 5 emails | .53 | 0 mails are SPAM | 5 Mails | 0 Mails |
| 4th group of 5 emails | .62 | 1 mails are SPAM | 4 Mails | 1 Mails |

1. Total number of Legitimate Emails are sent = 20
2. Number of Emails found spam positive is = 3
1. Number of Emails found spam positive is = 17
2. Over all Throughput for Legitimate Mails = 85%

### 6.2 Results when spam emails are sent

Table 2 shows the results, when the spam mails are sent. Experiment has been done with a lot of emails. But here is the some of results from them.

**Table 2: Results when SPAM Emails are sent**

| Email | Probability of whole mail | Results | True/False |
|---|---|---|---|
| 1st mail | .82 | SPAM | True |
| 2nd mail | .73 | SPAM | True |
| 3rd mail | .68 | Not SPAM | False |
| 4th mail | .97 | SPAM | True |
| 5th | .87 | SPAM | True |

| | | | |
|---|---|---|---|
| mail | | | |
| 6th mail | .70 | SPAM | True |
| 7th mail | .92 | SPAM | True |
| 8th mail | .42 | Not SPAM | False |
| 9th mail | .74 | SPAM | True |
| 10th mail | .88 | SPAM | True |
| 11th mail | .79 | SPAM | True |

1. Total number of legitimate emails are sent = 20
2. Number of emails found spam positive is = 2
3. Number of emails found spam positive is = 18
4. Over all Throughput for Legitimate Mails = 91%
5. True Positive(P) =9
6. False Positive(R)= 2
7. True Negative(S)= 0
8. False Negative(Q)= 0

## 7. CONCLUSION

It is now well known that not a single technique can be claimed alone to be the ideal solution with 0% false positive and 0% false negative. Currently being used anti spam systems couples several machine learning techniques for content classification. Spam assassin uses the genetic programming to generate its bayesian classifier for each release. Text classification techniques, such as bayesian classifiers and neural networks offer a good theoretical and practical background to fight the problem of spam.

In this paper, we provide the design, implementation of a new anti spam system. For the administration of email system, and our system can guarantee that legitimate emails are transferred to the users. This property is important for both users and administrators. This technique can filter present day spam acceptably well using nothing more than a Bayesian combination of the spam probabilities of individual words. Using a slightly tweaked Bayesian filter, now it misses only less than 5 per 1000 spam emails, with 0 false positives. Even if it is misjudges then no problem because it is utilized in proxy server, not in user's system. Even if it misjudges, the legitimate email must be delivered.

## 8. REFERENCES

[1] D. Heckerman, "Anti SPAM System: Another Way of Preventing SPAM", Proceeding of the 16th international on database and Expert Systems Applications 1529-4188/05 on 5-9 Sept. Page(s):1 – 5, 2005.

[2] Tarek Hassan, Peter Cole, "An Intelligent Spam Filter", School of information Technology, 2008. 28-30 Page(s):466 – 473, Aug. 2008.

[3] J. Provost, "Naïve-Bayes vs. rule-learning in classification of email," The University of Texas at Austin, Department of Computer Sciences, Technical Report AI-TR-99-284, : 4-13, 2003.

[4] E.-S. M. El-Alfy and R. E. Abdel-Aal, "Spam filtering with abductive networks," in Proceedings of the IEEE International Joint Conference on Neural Networks (IJCNN'08), Hong Kong, :1-3, June 2008.

[5] E.-S. M. El-Alfy, "Learning Methods for Spam Filtering," International Journal of Computer Research, vol. 16, no. 4, 2008.

[6] Y. Yang, S. Elfayoumy, "Anti-spam filtering using neural networks and Bayesian classifiers," in Proceedings of the 2007 IEEE International Symposium on Computational Intelligence in Robotics and Automation, Jacksonville, FL, USA, : 13-22, June 2007.

[7] M. Sahami, S. Dumais, D. Heckerman, and E. Horvitz, "A Bayesian approach to filtering junk e-mail," in Proceedings of AAAI'98 Workshop on Learning for Text Categorization, Madison, WI, :8-12, July 1999.

[8] Cournane, A., & Hunt, R. "An Analysis of the tools used for the generation and prevention of spam." Computer & Security, 23 (2), 154- 166, 2004.

[9] Bayesian Technique. Retrieved 10th September 2004. (Accessed from http://classifier4j.sourceforge.net ).

[10] T. A. Meyer and B. Whateley. Spambayes: "Effective open-source, bayesian based, email classification system". In Proceedings of the First Conference on Email and Anti-Spam (CEAS), 2004.

[11] Spertus, E. Smokey: "Automatic Recognition of Hostile Messages". Proceedings of the 14th National Conference on AI and the 9th Conference on Innovative Applications of AI, pp. 1058–1065, Providence, Rhode Island, 2000.

[12] Payne, T.R. and Edwards, P. Interface Agents that Learn: "An Investigation of Learning Issues in a Mail Agent Interface". Applied Artificial Intelligence, 11(1):1–32, 1999.

[13] Hall, R.J. "How to Avoid Unwanted Email". Communications of ACM, 41(3):88–95, 2004.

[14] Patrick Pantel and Dekang Lin. Spamcop: "A spam classification and organization program". In Learning for Text Categorization: Papers from the 2006 Workshop, Madison, Wisconsin, AAAI Technical Report, 2006.

[15] John Aycock & Nathan Friess, "Spam Zombies from Outer Space" Department of Computer Science University of Calgary, 15th Annual EICAR Conference, pages: 23-31, 2001.

[16] Cohen,W. "Learning rules that classify e-mail". In Proceedings of the AAAI Spring Symposium on Machine Learning in Information Access. Palo Alto, California, 18–25, 2003.

[17] Quinlan, J.R. "C4.5: Programs for Machine Learning". Morgan Kaufmann, pages: 44-59, 2002.

[18] G. Sakkis, I. Androutsopoulos, and G. Paliouras, "A memory-based approach to anti-spam filtering," Information Retrieval, vol. 6, pp. 49- 73,8-3-2003.3.