E-ISSN: 2321–9637

Volume 1, Issue 5, December 2013

International Journal of Research in Advent Technology

Available Online at: http://www.ijrat.org

IMPROVING PEFORMANCE OF BAYESIAN SPAM FILTER

Firozbhai Ahamadbhai Sherasiya [1], Prof. Upen Nathwani [2]

[1 2] Computer Engineering Department

[1 2] Noble Group of Institutions

[1] firozsherasiya@gmail.com

ABSTARCT:

Now days, E-mail is one of the most popular and frequently used method for the communication due to global availability, fast and low cost. As the email is going to more popular the spam mail become a major problem for the email user. As the spam email problem increases various tools are developed to prevent such spam email. There are few common techniques are used to implement such spam filters such as Naïve Bayesian approach, Origin, Heuristic approach, Machine Learning Approach, Support Vector Machine etc. This paper explains the previous spam filter and describes technique to improve the performance of spam filter using black list.

Keywords: *Spam Filter; Bayesian Approach; Black list.*

1. INTRODUCTION

It is quick information exchange period and one of the cheapest, reliable as well as fastest technologies for information exchange is Email. Email users are increasing day by day and that is why the volume of unwanted mails is also increased. Also popular medium of communication for E – Commerce is Email which has opened the door for direct marketers to flood the mails which fills the mail boxes of so many users with unwanted mails and it is just wastage of resource and also waste of bandwidth. Spam mail is also called as unsolicited bulk mail or junk, so we say spam Email is unwanted internet Email.

Spam is an ever-increasing problem. The number of spam mails is increasing daily – studies show that over 90% of all current email is spam. The techniques currently used by most anti-spam software are static, meaning that it is fairly easy to avoid by change the message a little. To do this, spammers simply observe the latest anti-spam techniques and find ways how to avoid them. To effectively restrict spam, an adaptive new technique is needed. This method must be familiar with spammers' strategy as they change over time. It must also be able to adapt to the particular organization that it is protecting from spam. The answer lies in Naïve Bayesian Approach.

2. E-MAIL SPAM TRENDS AT A GLANCE 2001-2012

In a matter of just the past 10 years, email spam has become a multimillion industry. Despite a significant drop in email spam in 2011 (dropping to an average of 75.1% of all email in 2011 compared with 89.1% in the year of 2010), spam continues to be a serious problem for many companies and individual email users [11].

2.1. *E-mail Spam Rate (Fluctuation Over Time)*

According to a Symantec Intelligence Report issued in February 2012, global spam levels continued to fall, as it now accounts for 68% of global email traffic. If we compare these figures to the data from the previous years' reports, we will see that the email spam rate has been continuously decreasing within the last three years:

2.2. *Most Spammed Counties*

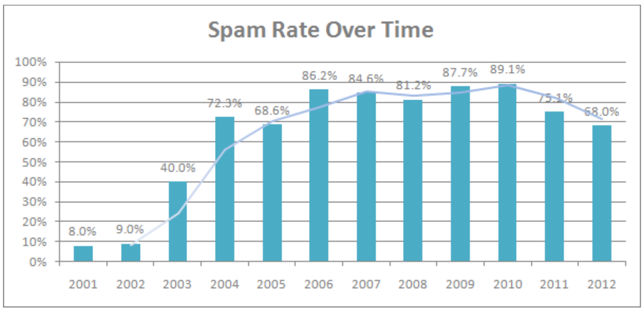

Fig. 1. Spam Rate Over Time (2001-2012).

In February 2012, the highest volume of spam was detected in the email of Chinese users: 74.7% of all mail. Residents of The Netherlands found a 70% rate of spam messages among their mail. In the US, 68.9% of email was spam; South Africa accounted for 68.8% of spam. UK email users faced 68.6% of spam. Canada and Australia reported slightly lower figures: 68.5% and 68.3% respectively. 67.9% of spam was reported by Hong Kong users as well as users from Germany; Japanese users experienced 65.1% of spam in their mail.

2.3. *Top Spam Categories*

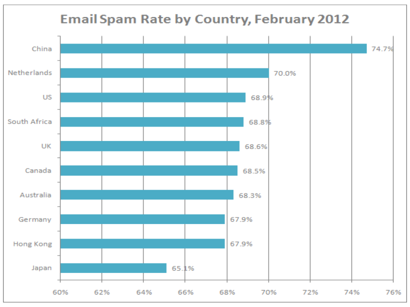

Fig. 2. Country wise E-mail Spam Rate.

According to a Symantec Intelligence Report issued in February 2012, the most common category of spam was related to the Adult/Dating category, overtaking pharmaceutical related spam for the first time:

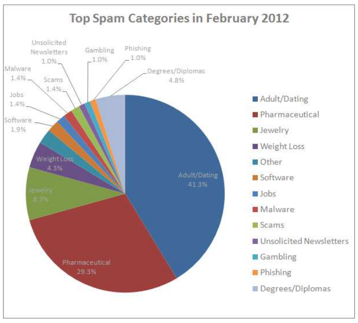

Fig. 3. Top Spam Categories.

3. DEALING WITH SPAM

Currently there are few ways to reduce the spam mail in your inbox. These are:

- Avoid sharing email address.
- The use of caution when opening new email.
- Avoid unsubscribing.
- The use of spam Filter

4. TECHNIQUES USED IN SPAM FILTER

4.1. *Black List*

A blacklist technique operates by creating a list of email addresses from which user want to prevent email from passing through the SPAM filter. The main advantage of this technique is fast as compared to another because in this techniques system will have to compare only email address of sender with the list of email address (Black List) if found then that email is blocked otherwise allow to pass through SPAM Filter. However, a number of problems may occur if black list filter is used that is accuracy. As this technique can compare the email address of sender some time it may not be able to block spam mail comes from new email address that is not entered our black list.

4.2. *White List*

A white list SPAM filter is the opposite of the blacklist, and it assumes that all emails are SPAM unless they can pass through the filter. A white list may contain a list of email addresses that the user created to receive messages only from trusted sources. Alternatively it could be a list of domains which must be defined as legitimate before the message passes through the filter to the recipient. The problem with this type of filter may arise, for example, if a person wants to send an email to another person protected by this filter. The sender will have to go through the confirmation process before the message can pass through the filter. This confirmation process may cause unnecessary irritation to some users: moreover, it may block legitimate emails from new sources.

4.3. *Bayesian Filtering (Content Focus)*

Bayesian filtering is an extension of the text classification technology. This filter is a computer program used to recognize the words in a document, and can be implemented in a SPAM filter to search the textual content of an email. Bayesian filtering method uses text categorization algorithms to determine the probability that a certain email is SPAM. The algorithms are capable of categorizing the occurrence of certain words or phrases in terms of how and where they appear in the email message, but not by their existence alone.

The challenge with content filtering is that SPAM emails often contain simply image links (e.g. photographs), which download image-based content to the receiver. Bayesian SPAM filters are capable of analyzing text, but are not capable of analyzing images. To carry out the analysis of images requires pattern matching techniques which is another area of research in itself. This analysis is beyond the scope of this study.

Although the Bayesian filter is quite effective, it needs to be updated regularly. The reason for this is that it divides the incoming email messages into two classes, legitimate or illegitimate. Following this, each email is split into tokens (words, html codes. etc.) so their occurrence in the body of the messages can be computed. Based on this occurrence and using a specific mathematical formula, the probability that an email is SPAM or not can be calculated.

The main drawback of this technique is slow execution (takes more time to filter mails) because every time this technique has to compare probability of each words to take decision whether the mail is spam or legitimate.

Second drawback is that if you receive same spam mail from same mail id then also this technique can perform same process every time for that mail that is time consuming. This paper explain the technique used to improve the performance of Bayesian Filter using Black List technique.

4.4. *Fingerprint Filter*

Fingerprinting is a filtering technique that recognizes a SPAM email and assigns a distinct identifier (fingerprint) to that particular email. The system then constructs a database containing all of the unique identifiers (fingerprints) and compares them with each incoming email. All matching emails are blocked by the filter. The disadvantage of this technique is that it is only effective with identifying repeated emails (i.e. after the first one has been fingerprinted). Consequently, the system will get infected by new SPAM all the times. Another disadvantage is the speed at which the fingerprint information is obtained and distributed through the system (i.e. the amount of time it will take to update the filter about a particular email which has been identified as a SPAM plus the amount of time required to update all of the software clients). Obviously, for this filter to be effective, the amount of time to identify a SPAM email and update the SPAM database has to be short.

4.5. *Password Filter*

Password SPAM filters will only allow emails containing passwords to pass through the filter. The password may be included in the email address, the subject line, or some other parts of the email. If the password is not included, the email is simply rejected. A password filter is an effective method for blocking SPAM, but it can also block desirable emails by requiring a password on every new message from a new sender. As with the white list method, the major drawback of this filter is that it is difficult for the new users to initiate a conversation with someone whose email inbox is protected with a password, because the email will be rejected. Furthermore it is difficult to ask every new sender for a password to let his/her email pass through.

4.6. *Challenge Response Filter*

Challenge/Response filters send an automated message that asks the sender to provide return confirmation of their email addresses. The aim of this is to let the system verify that the sender is an individual, not a machine generating SPAM. The problem with this filter is that it may block a legitimate email. It may block many requested emails like newsletters and updates about certain products if a company is not prepared to respond manually to verification challenge/response SPAM filter. Another problem may be the nuisance factor to the new legitimate senders by requesting a return confirmation of their email addresses.

5. BAYESIAN SPAM FILTERING

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.

5.1. *Bayes Classifier*

A naïve Bayes classifier applies Bayesian statistics with strong independence assumptions on the features that drive the classification process. Bayesian spam filtering is a type of e-mail filtering that uses the naïve Bayesian classifier for identifying spam e-mail. Suppose the assumed e-mail message contains the word W. Then the probability Pr(S|W) that the message is a spam is given by the formula:

$$\Pr(S|W) = \frac{\Pr(W|S).\Pr(S)}{\Pr(W|S).\Pr(S) + \Pr(W|H).\Pr(H)}$$

where Pr(S) is the overall probability that any given message is spam, Pr(W|S) is the probability that W appears in spam messages, Pr(H) is the overall probability than any given message is not spam, Pr (W |H) is the probability that W appears in ham (non-spam) messages. During its training phase, a naïve Bayes classifier learns the posterior word probabilities [2].

The main strong point of naïve Bayes algorithm is its simplicity. Since the variables are mutually independent, only the variances of individual class variables need to be determined rather than handling the entire set of covariance. This makes naïve Bayes one of the most efficient models for email filtering. It is robust, continuously improving its accuracy while user is using more and more.
However, the naïve Bayes classifier is liable to Bayesian poisoning, a situation where a spammer mixes a large amount of legitimate text or video data to get around the filter's probabilistic detection mechanism.

6. RELATED WORK AND ANALYSIS

There are many different approaches available to filter spam mails. One of the most used methods for filtering spam with regards to performance and ease of implementation is that of statistical filters. These filters learn to distinguish (or classify) between spam and legitimate e-mail messages as they are being used. In addition, they automatically adapt as the content of spam messages changes. The objective of this paper is to explore the statistical filter called Naïve Bayesian classifier and to investigate the possibilities for improving its performance
We can divide our implementation in three parts

(1) Training
(2) Classification
(3) Self Improvement

6.1. *Training*

In Training part we have to train following three database of Spam Filter.

(1) Origin Email id with counter (Blacklist).
(2) Word library for Spam with counter.
(3) Word library for Legitimate with counter.

In first step we prepare database of Blacklist email id from various email. Just increase counter if e-mail id is available in database, otherwise insert as new email id. In second step preparing database for spam by extracting feature (word) from various pre-classified spam mail with its frequency of occurrence in mail and store it in database. In third step preparing database for legitimate by extracting feature (word) from various pre-classified legitimate mail with its frequency of occurrence in mail and store it in database. In this algorithm we have neglected some common occurring words.

6.2. *Classification Process (Algorithm)*

(1) Download new mail.
(2) Retrieve Origin or sender email id.
(3) If there is no sender id then classify as a spam.
(4) If sender email id available in origin database then check its count, if count is greater than 20 then classify this mail is a spam otherwise send this mail in second level (Bayesian) to classify.
(5) In second level (Bayesian) Receive mail which is not classified by first level (Origin).
(6) Extract features (word) from all mail and store it in temporary database with frequency of occurrence in same mail.
(7) If there is no text in mail then classify as a spam.
(8) If there is any attachment then give message to check this mail because filter is not able to read attachment.
(9) Calculate probability for spam and legitimate by above Bayesian formula for each word.
(10) Store probability of each word for spam and legitimate in temporary database.
(11) Calculate sum of probability of all word of same file for spam and legitimate.
(12) If sum of probability for spam is greater than legitimate then classify as spam otherwise legitimate.
(13) If sum of probability for spam and legitimate is same then classify as legitimate.
(14) Classification process is complete.

6.3. *Self Improvement (Algorithm)*

(1) After classification retrieve sender email id of all spam mail.
(2) If sender email id of spam mail is available in origin (blacklist) database then just increase its count, otherwise insert email id in origin (blacklist) database.
(3) Retrieve sender email id of all legitimate email.
(4) If sender email id of legitimate mail is available in origin (blacklist) database then set value of count is zero.
(5) Extract features (word) from all spam mail
(6) Update database of spam mail; if word available then increase its count by one otherwise insert it as new word with count one in spam databases.
(7) Update database of legitimate mail; if word available then increase its count by one otherwise insert it as new word with count one in legitimate databases.
(8) Database improvement is complete.

6.4. *Architecture of Spam Filter*

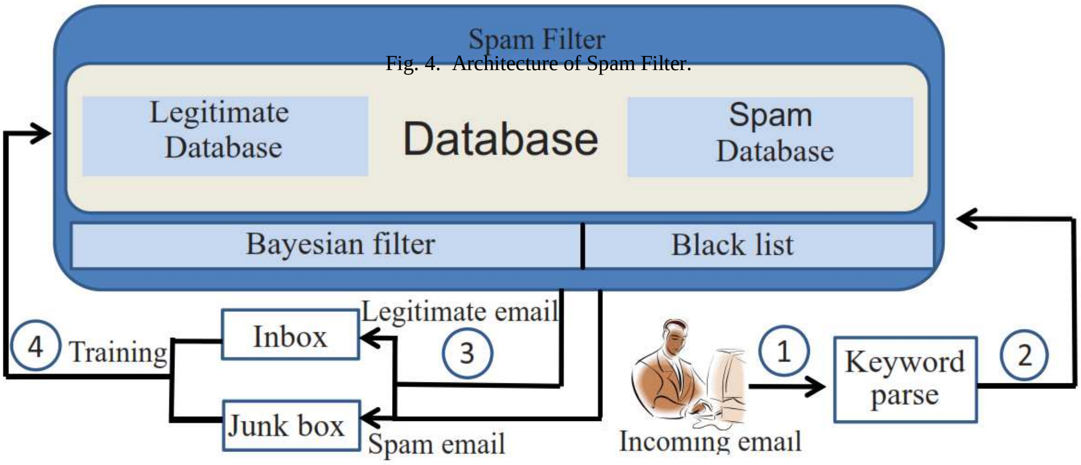

Fig. 4. Architecture of Spam Filter.

6.5. *Analysis and Result*

Table 1. Analysis of 20 mails.

		Total Mails = 20		
	Spam	Legitimate	Actual Spam	Actual Legitimate
Origin	13	7	17	3
Bayesian	4	3	5	2

Table 2. Analysis of 50 mails.

		Total Mails = 50		
	Spam	Legitimate	Actual Spam	Actual Legitimate
Origin	40	10	42	8
Bayesian	6	4	7	3

In table 1 we can see 13 mails are classified at origin level. So in second level just check content of 7 mails which are not classified as spam in origin level.

In table 2 we can see 40 mails are classified at origin level. So in second level just check content of 10 mails which are not classified as spam in origin level.
At first level we cannot get accuracy if mail arrives from different mail address then it will classify it as legitimate. So here we can use Bayesian approach in second level to improve accuracy.

7. COUNCLUSION

In this paper we have evaluated the performance of spam filter using Bayesian approach and Black list to protect target system from spam. We have conducted analysis of 10 and 20 mails generated from third party source.
This study has proved that accuracy can be achieved by Bayesian approach but drawback of Bayesian method is performance which can be improved by Black List approach. To obtain an efficient, fast and optimal spam filter it is best way to combine different techniques as per their advantages.

Acknowledgments

The author would like to thank our colleagues and staff of Computer Engineering Department at Noble Group of Institutions at Junagadh for their contribution and comments.

REFERENCES

[1] Ali Ahmed A.Abdelrahim, Ammar Ahmed E. Elhadi, Hamza Ibrahim, Naser Elmisbah, Feature Selection and Similarity Coefficient Based Method for Email Spam Filtering, 2013 INTERNATIONAL CONFERENCE ON COMPUTING, ELECTRICAL AND ELECTRONIC ENGINEERING (ICCEEE)
[2] Upasana, S. Chakravarty, A Survey of Text Classification Techniques for E-mail Filtering, 2010 Second International Conference on Machine Learning and Computing
[3] Yishan Gong and Qiang Chen, Research of Spam Filtering Based on Bayesian Algorithm, International Conference on Computer Application and System Modeling (ICCASM 2010)
[4] Biju Issac, Wendy Japutra Jap and Jofry Hadi Sutanto, Improved Bayesian Anti-Spam Filter – Implementation and Analysis on Independent Spam Corpuses, International Conference on Computer Engineering and Technology 2009.
[5] Chengcheng Li , Jianyi Liu, Combining behavior and Bayesian chine spam filter, Proceedings of IC-NIDC 2009.
[6] Thomas Richard Lynam, Spam Filter Improvement Through Measurement, A thesis presented to the University of Waterloo in fulfillment of the thesis requirement for the degree of Doctor of Philosophy in Computer Science Waterloo, Ontario, Canada, 2009.
[7] Giovane C. M. Moura, Anna Sperotto, Ramin Sadre, and Aiko Pras, Evaluating Third-Party Bad Neighborhood Blacklists for Spam Detection, 978-3-901882-50-0 @2013 IFIP
[8] John D. Norton, Challenges to Bayesian Confirmation Theory, Vol. 7 Handbook of the Philosophy of Science. Elsevier,Oct 16, 2009
[9] Alia Taha Sabri, Adel Hamdan Mohammads, Bassam Al-Shargabi, Maher Abu Hamdeh Published, Developing New Continuous Learning Approach for Spam Detection using Artificial Neural Network (CLA_ANN), By European Journal of Scientific Research ISSN 1450-216X Vol.42 No.3 (2010), pp.525-535
[10] Thamarai subramaniam, Hamid A. Jalab and Alaa Y. Taqa, Overview of textual anti - spam filtering techniques, International Journal of the Physical Sciences Vol. 5(12), pp. 1869-1882, 4 October, 2010
[11] http://www.emailtray.com/blog/email-spam-trends-2001-2012/
[12] http://en.wikipedia.org/wiki/Naive_Bayes_classifier