

Probability Modeling for Improving Spam Filtering Parameters

S. C. Chiemeké¹
University of Benin
Nigeria

O. B. Longe²
University of Ibadan
Nigeria

Abstract

As efforts at combating electronic mail Spam rages on, users and technocrats are becoming fuzzy in their judgment of what actually constitute Spam. Most Spam filters depend on routing information, sender addresses, message title, evidence of bulk mailing, non-solicitation and mail contents to identify Spam messages. Some researchers have posited that it is the act of blindly mass-mailing a message that makes it Spam and not the actual content of the message. This to our understanding relegates the use of message content as an important factor in designing filters for identifying Spam mails. To identify the effectiveness of filtering with or without considering mail contents, we calculated the probability of the occurrence of events; Spam and real mails (called Ham) in terms of mutually inclusive and exclusive events. We then used probability theory to model parameters for identifying their instances. Our results showed that the efficiency of filtering Spam mails using other parameters and without consideration for the message content in an inclusive mail corpus reduces the volume of false positives by 8.3%. For an exclusive Spam corpus, that is, mails that are identified as Spam, filtering without consideration for message content reduced the efficiency of the filter by 7.7%.

Keywords: Nigeria, antispam, 419 mail, filters, ham, scam, spammers, yahoo-boys.

Introduction

Techniques currently employed to prevent Spam revolves around using filters to examine various parts of mail message to determine whether or not it is Spam. These systems are either origin or address-based filters which typically use network information for classifying mails and content filters that examine the contents of email messages. Rule-based or heuristic filters (by extension a content filter) search the email message for patterns that indicate Spam. These patterns could include specific words or phrases, malformed message headers, large amounts of exclamation marks and capital letters. Detection of a specific pattern attribute that exceeds a set threshold classifies a message as Spam. Rule-based filters were the most common type of Spam filter until 2002, when Bayesian filters became popular (Graham 2002). Bayesian (also known as statistical) filters work by analyzing the words inside an email message to calculate the probability that it is Spam. This probability is based on not only those words that provide evidence that a message is Spam, but also on those words that provide evidence that a message is

not Spam. Words that are not generally found in Spam messages contribute to the probability value in very much the same way as words that are frequently found in Spam messages.

To calculate an e-mail's Spam probability with a good degree of accuracy, Bayesian filters need to be trained by being given examples of what constitutes a Spam email and what does not. The advantage of this technique is that, given appropriate time and training data, Bayesian filters can achieve a combination of extremely high accuracy rates with a low percentage of false positives. The low amount of false positives generated by a Bayesian filter is useful, as users generally regard the classification of legitimate emails as Spam as an order of magnitude worse than receiving Spam incorrectly classified as legitimate (false negatives).

Related Literature

Data mining, machine learning, text classification and social network techniques, such as Reputation Network Analysis techniques and global collaboration filters have been applied to categorizing text for filtering purposes. Some of these include, density-based filters, feature selection and feature categorization (Lewis, 1992), memory-based approach (Sakkis, 2003), artificial neural networks (Drewes, 2002), machine learning for filtering (Sebatiani, 1999), subject line labeling (FTC 2005), pattern discovery based system (Isidore and Tien, 2005), sender and receiver addresses as cues for filtering (Chi-chien, 2003). Rajkumar et al (2006) proposed a technique which revolves around using intrusion detection system (IDS) techniques, specifically audit log analysis, to stop Spam messages before they enter the network. Longe & Chiemeke (2006) proposed a hybrid collaborative content-specific filters (SPAMAng) for trapping Nigerian scam mails at the origin.

Mail Constituents

In order to design effective Spam filters, elements that have been used to identify Spam include routing information, sender addresses, message title, evidence of bulk mailing, non-solicitation and mail contents. We disagree with the school of thought (Ion et al, 2000), that posited that it is the act of blindly mass-mailing a message that makes it Spam and not the actual content of the message

Given the universal corpus of what constitute Spam mails as:

Spam Mail = S_m and identifying elements as
 Routing information = R_t
 Sender addresses = S_a
 Message title = M_t
 Bulk Mails = B_m
 Mail contents = M_c
 Not-solicited = N_s

Then we can express S_m mathematically as:

$$S_m = f(R_t, S_a, M_t, B_m, M_c, N_s) \quad (1)$$

On the other hand 'Ham' are also electronic mail messages but not exact opposite of Spam as

some authorities might want to posit. Ham and Spam share few characteristics and the fact that they shared it at all is one of the major challenges in designing effective antispam systems.

Ham can be defined by the following properties.

- Ham mails = H_m .
- Routing information = R_t^* (the determinant for white-listing or blacklisting in origin based filters)
- Sender addresses = S_a^* (A determinant for challenge-response, white-listing or blacklisting in origin based filters)
- Message title = M_t^* (Used in content-based filters)
- Not Bulk Mail = N_{Bm} (Very subjective)
- Mail contents = M_c^* (Used in content-based filters)
- Solicited Mail = S_s (Used in origin-based filters)

Ham (H_m) can therefore be expressed mathematically as:

$$H_m = f(R_t^*, S_a^*, M_t^*, N_{Bm}, M_c^*, S_s) \quad (2)$$

Using set theory, the union of (1) and (2), U_m , constitute the universal set of e-mails (U_m) (that is the entire pool of mails containing both ham and Spam)

Expressed mathematically,

$$U_m = S_m \cup H_m = f(R_t, R_t^*, M_t, M_t^*, S_a, S_a^*, B_m, M_c, M_c^*, N_{Bm}, N_s, S_s) \quad (3)$$

The Union of any of these two corpuses of mail parameters or the existence of any of them singly can cause an email to be recognized as either Ham or Spam.

The specific identification of a mail exclusively as Spam (E_s) is therefore expressed as the intersection of (1) and (2). This expression can be two-fold:

- (a) Taking the mail content into consideration (Wc)
- (b) Without taking the mail content into consideration (Wtc) and

$$(Wc) = S_m \cap H_m = f(R_t, S_a, M_t, M_c) \quad (4)$$

$$(Wtc) = S_m \cap H_m = f(R_t, S_a, M_t) \quad (5)$$

These expressions are based on the understanding that $R_t \equiv R_t^*$, $M_t \equiv M_t^*$, $S_a \equiv S_a^*$ and $M_c \equiv M_c^*$.

From (4), we can deduce that Routing information (R_t), Sender addresses (S_a), Message title (M_t) and Mail contents (M_c) are common denominators to Ham and Spam and therefore remains very powerful parameters for separating and classifying electronic mails as Spam or Ham. Efforts at designing credible filters must therefore be directed at these parameters. Among these parameters, the subtlest in the hand of the spammer is the message content. Whether directed as mass mailing or singly at a victim, the spammer, through the choice of the right words (manipulated and otherwise) can gain tremendous advantage at the expense of the recipient.

We introduce probability concepts to identify the extent of efficiency for filtering Spam with or without consideration for the *message content*.

When content is considered the following holds:

$$\begin{aligned} \text{No of Universal parameter for identifying mails} &= 12 \\ \text{No of parameters for exclusively identifying Spam} &= 6 \\ \text{No of parameters for exclusively identifying Ham} &= 6 \end{aligned}$$

$$\text{Probability of Mail Containing Spam} = \frac{n(S_m)}{U_m} = \frac{6}{12} = 0.5 \quad (6)$$

$$\text{Probability of Mail Containing Ham} = \frac{n(H_m)}{U_m} = \frac{6}{12} = 0.5 \quad (7)$$

When content is not considered the following holds:

$$\begin{aligned} \text{No of Universal parameter for identifying mails} &= 10 \\ \text{No of parameters for exclusively identifying Spam} &= 5 \\ \text{No of parameters for exclusively identifying Ham} &= 5 \end{aligned}$$

$$\text{Probability of Mail Containing Spam} = \frac{n(S_m)}{U_m} = \frac{5}{10} = 0.5 \quad (8)$$

$$\text{Probability of Mail Containing Ham} = \frac{n(H_m)}{U_m} = \frac{5}{10} = 0.5 \quad (9)$$

Equations 6/7 and 8/9 imply that in the universal corpus, the chances of Ham and Spam occurring are the same (50 – 50).

Corpus of Filtering

Inclusive Filtering

To identify the effectiveness of filtering with or without contents, we calculate the probability of the events S_m and H_m in terms of mutual inclusiveness and exclusiveness.

The chances or probability that a sent mail could be either Spam or Ham given consideration to mail content or not is derived below:

$$P_{wc}(S_m \cup H_m) = P(S_m) + P(H_m) - P(S_m \cap H_m)$$

$$P(S_m) = \frac{1}{2},$$

$$P(H_m) = \frac{1}{2},$$

$$P_{wc}(S_m \cap H_m) = \frac{4}{12} = \frac{1}{3}$$

$$P_{wtc}(S_m \cap H_m) = \frac{3}{12} = \frac{1}{4}$$

Efficiency of Filtering with contents (E_{fc}) is given:

$$\begin{aligned} P_{wc}(S_m \cup H_m) &= \frac{1}{2} + \frac{1}{2} - \frac{1}{3} \\ &= 1 - \frac{1}{3} \\ &= \frac{2}{3} \\ &= 0.6667 \\ &= 66.67\% \end{aligned}$$

Efficiency of Filtering without contents (E_{fwc}) is given:

$$\begin{aligned} P_{wtc}(S_m \cup H_m) &= P(S_m) + P(H_m) - P(S_m \cap H_m) \\ &= \frac{5}{12} + \frac{5}{12} - \frac{1}{4} \\ &= \frac{10}{12} - \frac{1}{4} \\ &= \frac{10 - 3}{12} \\ &= \frac{7}{12} \\ &= 0.5833 \\ &= 58.33\% \end{aligned}$$

Exclusive Filtering

In the event that the occurrence of Spam and Ham are mutually exclusive (that is the mails going into the filter are either totally Spam or Ham), the efficiency of the filtering system when mail contents are considered is given as:

$$\begin{aligned}
& P_{wc}(Sm \cap Hm) \\
&= P_{wc}(Sm) \times P(Hm) \\
&= \frac{5}{12} \times \frac{5}{12} \\
&= \frac{25}{144} \\
&= 0.1736 \\
&= 17.3\%
\end{aligned}$$

In the exclusive event in which the mail content is not considered as part of the filtering parameters, the efficiency of the filtering system is given as:

$$\begin{aligned}
& P_{wc}(Sm \cap Hm) \\
&= P_{wc}(Sm) \times P(Hm) \\
&= \frac{1}{2} \times \frac{1}{2} \\
&= \frac{1}{4} \\
&= 0.25 \\
&= 25\%
\end{aligned}$$

Experimental

We pursue exclusivity in our experimental in order to validate our theoretical proposition. The subjective factors of bulk mailing and non-solicitation were removed as elements that do not absolutely identify a mail as Spam. Using common parameters for Spam, mails messages already classified as Spam by our scanning system were taken through the filter designed as part of the preliminary framework for the SPAMAng project. Selected parameters were activated in sequence in the filtering process. Results obtained are depicted in the table below.

Table 1. Filtering by Parameter Selection

Parameters	Number Of Messages	Number Identified As Spam By Filter	% Precision	No Of False Positives	%Misses
R_t	153	44	28.75	109	71.24
R_t, S_a	153	73	47.71	80	52.28
R_t, S_a, M_t	153	136	88.88	17	11.11
R_t, S_a, M_t, M_c	153	147	96.07	06	3.92

Key: R_t = Routing information; S_a = Sender addresses; M_t = Message title; M_c = Mail contents

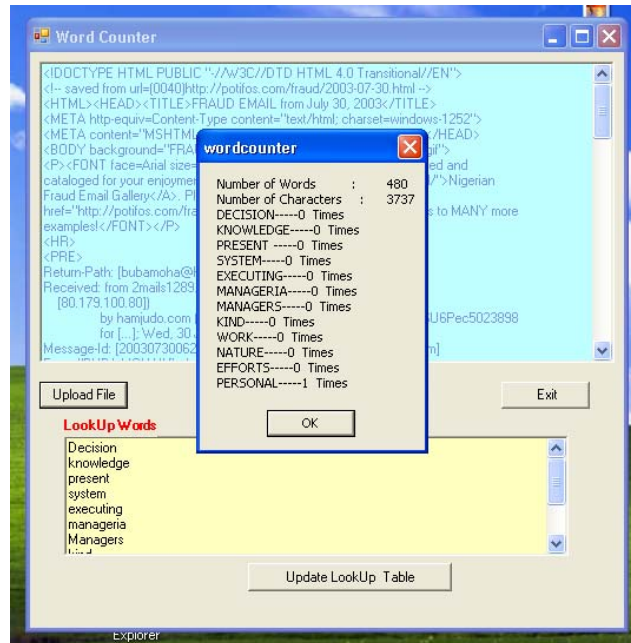


Figure 1. Text Scanning System for Identifying Text Categories

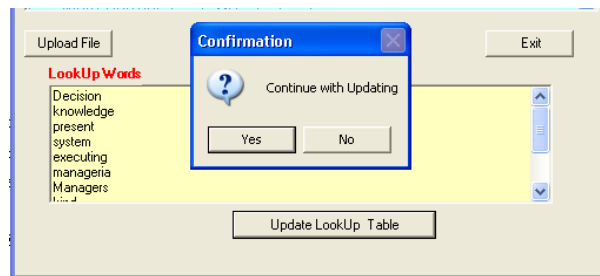


Figure 2. Text Category Interface for Including New Text/Message Content

Results and Discussions

The interpretation of the results for the inclusive probability is that there are 58.33% chances that Spam mails can be correctly filtered when mail content is not considered as part of the filtering parameters. Whereas with the inclusion of mail content, the efficiency increases to 66.67%.

$$E_{fc} - E_{fwc} = (66.67 - 58.33)\% = 8.34\%$$

The expression above shows that chances are that 8.34% of already filtered mails are Spam and they escaped the filtering system. That means the rate at which the filtering system produce false positives is 8.34%. Obviously, this percentage affects the throughput of the system.

For the exclusive filtering we have the probability that there are 25% chances that Spam mails can be correctly filtered when mail content is considered as part of the filtering parameters. Whereas without considering contents, the efficiency decreases to 17.3%.

$$E_{fc} - E_{fwc} = (25.0 - 17.3) = 7.7\%$$

Here, the result is interpreted as 7.7% of mails that are Spam are categorized as being non-Spam.

From Table 1, the efficiency of filtering improves as we descend the parameter column that is as more features/factors are introduced. The low efficiency experienced with the use of Routing Information (R_i) and Sender Addresses (S_a) is not unconnected with the number of IP and Sender addresses either blacklisted or whitelisted in our database. This part of the system is collaborative and will contribute more towards increased efficiency as the filtering system evolves. The marked improvement when message titles are introduced is not unconnected with the fact that text categories within message titles form part of the corpus for training the text classification engine. As our research has progressed, we already identified peculiar words/text categories in the Nigerian Yahoo-boys mail corpus (419 mails). An additional improvement was noticed as message content is introduced thereby justifying the additional efficiency expressed in the theoretical probability foundation. In this case, the margin is 7.19 (96.07 – 88.88) as compared to 7.7 in the probability computations.

Conclusion

Most Spam filters depends on routing information, sender addresses, message title, evidence of bulk mailing, non-solicitation and mail contents to identify Spam messages. The most subtle tool in the hand of the spammer among the various parameters mentioned remain the message content which can be easily manipulated to fool Spam filters. To calculate an e-mail's Spam probability with a good degree of accuracy, Spam filters need to be trained not just by being given examples of what constitutes a Spam email and what does not but by providing absolute parameters by which Spam can be identified. The advantage of this technique is that, given appropriate time and training data, Spam filters can achieve a combination of extremely high accuracy rates with a low percentage of false positives. The low amount of false positives generated by a Bayesian filter is useful, as users generally regard the classification of legitimate emails as spam as an order of magnitude worse than receiving Spam incorrectly classified as legitimate. For the instances considered above, results showed that the exclusion or relegation of message contents as a filtering parameter, considerably affects the throughput of antispam systems.

Future Direction

Different forms of Spam have peculiarities in the use of language. Therefore an understanding of the nature of text categories and possible manipulations by Spammers to fool Spam filters within specific Spam corpus that we are dealing with is imperative in the design of effective filters. Our future work will concentrate on expanding the understanding of the various forms of manipulations that Nigerian yahoo-boys carry out in order to fool Spam filters. We also intend to critically study mail messages for which the efficiency of our current filtering is low in order to identify new words within the mail content. This will enable us increase the vocabulary for

training our text classification engine.

References

- Chih-Chien, W. (2003). Sender and Receiver Addresses as Cues for Anti-Spam Filtering. *Journal of Research and Practice in Information Technology*, 36(1), 3-7.
- Drewes, R. (n.d.). *An artificial neural network Spam classifier*. Retrieved August 2002 from <http://www.interstice.com/drewes/cs676/spam-nn/spam-nn.html>
- Graham, P. (2002). *A Plan for Spam*. Retrieved from www.paulgraham.com/spam.html
- FTC (2005). *Subject Line Labelling As a Weapon Against Spam*. Federal Trade Commission Publications.
- Ion, A., John, K., Konstantinos, V., George, P., & Constantine, D. (2000). An Evaluation of Naive Bayesian Anti-Spam Filtering. Proceedings of the workshop on *Machine Learning in the New Information Age*, G. Potamias, V. Moustakis and M. van Someren (Eds.), 11th European Conference on Machine Learning (pp. 9-17). Barcelona, Spain.
- Isidore, R., & Tien, H. (2005). Pattern-discovery-based System for the Automatic Identification of Unsolicited E-mail Messages Discovery Based System for Filtering Spam. Bioinformatics and Pattern Discovery Group IBM Thomas J Watson Research Center Yorktown Heights, NY 10598, USA
- Lewis, D. (1992). Feature Selection and Feature Extraction for Text Categorization. Proceedings of the *DARPA Workshop on Speech and Natural Language* (pp. 212-217). Harriman, New York.
- Longe, O. B., & Chiemekwe, S. C. (2006, June). The Design and Implementation of an E-Mail Encryptor for Combating Internet Spam. Proceedings of the 1st International Conference of the International Institute of Mathematics and Computer Sciences (pp. 1-7). Covenant University, Ota, Nigeria.
- Rajkumar, B., Tianchi, M., Rei, S., Chris, S., & Willy, S. (2006). Domain Specific Blacklists. Proceedings of the Fourth Australian Information Security Workshop (AISW-NetSec 2006).
- Sakkis, G., Androutsopoulos, I., Paliouras, G., Karkaletsis, V., Spyropoulos, C. D., & Stamatopoulos, P. (2003). A memory-based approach to anti-Spam filtering for mailing lists. *Information Retrieval*. 6(1), 48-73.
- Sebastiani, F. (1999). *Machine Learning in Automated Text Categorization*. Technical Report B4-31, Istituto di Elaborazione dell'Informazione, Consiglio Nazionale delle Ricerche, Pisa. Retrieved from <http://faure.iei.pi.cnr.it/~fabrizio>
-

- ¹ Dr. Stella C. Chiemeke lectures and currently heads the Department of Computer Science, University of Benin, Benin City, Nigeria. She can be reached at: schiemeke@yahoo.com; Phone: +2348023158911
- ² Mr. Babatope O. Longe teaches computing and Information Technology at the Department of Computer Science, University of Ibadan, Ibadan, Nigeria. He can be reached at: longeolumide@yahoo.com; Phone: +2348024071175