Email Spam Filter using Bayesian Neural Networks

Nibedita Chakraborty[1], Anjani Patel[2]
Department of Computer Science & Engineering[1, 2]
K. G. Polytechnic, Raigarh, India [1, 2]

Abstract

***Nowadays, e-mail is widely becoming one of the fastest and most economical forms of communication but they are prone to be misused. One such misuse is the posting of unsolicited, unwanted e-mails known as spam or junk e-mails. This paper presents and discusses an implementation of a spam filtering system. The idea is to use a neural network which will be trained to recognize different forms of often used words in spam mails. The Bayesian ANN is trained with finite sample sizes to approximate the ideal observer. This strategy can provide improved filtering of Spam than existing Static Spam filters**.*

Keywords

Spam, Bayesian Filtering, False Positive, False Negative

1. Introduction

Today Emails are efficient, rapid and cheap mean of communication. This makes it favorite both in professional and personal correspondences. But receiving E-mail from unknown source and content of which is not of the user interest is not really a misfortune. These kind of unwanted emails are called spasm but not all of them are spam. Another term would be unsolicited commercial email, but unfortunately spam is not only advertising material. Email spam also known as junk mail is “the practice of sending unwanted email messages frequently with commercial content, in large quantities to random set of recipients”.

Hence there is a need of classifiers that can distinguish useful email from spam. The simplest and most common approaches are to use filters that screen messages based upon the presence of common words or phrases common to junk email. Other simplistic approaches include blacklisting and white listing.

The Bayesian classifiers identify attributes (usually keywords or phrases common to spam) that are assigned probabilities by the classifier. The product of the probabilities of each attribute within a message is compared to a predefined threshold, and the messages with products exceeding the threshold are classified as spam.

2. Types of Spam

Although there are at least three main different kinds of spam [Le Zhang, et al, 2004] and especially advertising spam could be divided into many subcategories sub categories , all the spam has also a content-free characteristic. All of these characteristics could be the base for further filter solutions.

1. Advertisement Spam

Most spam is commercial advertisement, often a direct product offer. Spam costs the sender very little to send, compared to other advertisement methods. The most common subcategories of the advertisement spam are:

- Online Pharmacy spam
- Penny Stock spam
- Porn or dating spam
- Pirate Software spam
- Fake Degrees spam

2. Financial Spam

While advertisement spam have at least a little probability, that the responder could get something for the sent money, the financial spam only tries to fool people and get their money somehow, without the chance to buy anything. The most common financial spam are Lottery spasm which tells readers that ‘You have already won X Million’ in order to try to extract transfer fees etc.

3. Phishing

Phishing spam [Xin Jin, et al, 2006] is fake alert from banks (mostly Citibank), Papal, embay etc, and it asks for confirmation, validation or monitoring of details in order to defraud people of their personal information. Phishing spam are usually linked to fake login sites, which can be used to capture user details (e.g. passwords) in order to use this information to steal money or goods. Phishing emails use mostly the listed methods:

- Using the company’s Image

- Links to the real company's site
- Email appears to be from the spoofed company

3. Need of a spam filter

E-mail reading is nowadays a daily habit of many people. Indeed, emails are efficient, rapid and cheap mean of communication. Reading occasionally an E-mail from unknown source and content of which is not of the user interest is not really a misfortune. However, when more than 60% or even 90% of E-mails are of such kind, and often illicit; this is what one might call a nightmare. The cost induced by productivity and resources loss, filtering software, and support caused by only one unsolicited E-mail to from 1$ up to 2$; multiplied by the number of spam sent and received every day, the one dollar becomes then millions. Studies show that over 70% of all current email is spam.

Most email readers must spend a non-trivial amount of time regularly deleting spam messages, even as an expanding volume of junk email occupies server storage space and consumes network bandwidth. For example, if a company has no filter and a worker receives 6 spam messages daily and it takes an average of 5 seconds to read and delete each spam message this means, that the worker will spend almost 3 hours per year to read and delete spam. An ongoing challenge, therefore, rests within the development and refinement of automatic classifiers that can distinguish legitimate email from spam.

The primary flaw in the normal Filter and Blacklisting is that the spammers can change their identities or to alter the style and vocabulary of their sales pitches. White listing risks the possibility that the recipient will miss legitimate email from a known or expected correspondent with a heretofore unknown address, such as correspondence from a long-lost friend. Therefore the need of special classifier or filter arises. Here is the basic spam filter which shows the process of detecting spam.

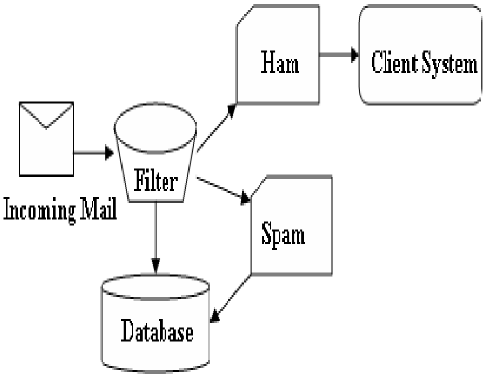

Fig. 1 Process of Spam Detection

4. Spam Filter Blacklist

There is a need, to make difference between two levels of blacklisting: the network-level and the address-level blacklisting. The network-level blacklisting is based on creating intentional network outages. The method has the ability to detect spam letters based on its origin rather than its content. Unfortunately new spam hosts can pop up instantly. If a legitimate user was accidentally blacklisted, there is no way, to get off the blacklist, hence all mails ware rejected from the blacklisted part of the network. The address-level blacklist is an updated list of known spam sender addresses. There are on-line accessible blacklists and the user can administrate personal blacklist as well.

White list

White listing is the opposite of blacklisting. A white list is a collection of reliable contacts. If email comes from the members of this list, it should be marked automatically as legitimate letter what is also called ham. Just as the blacklisting, the white list also needs a continuous upgrade and refreshment. Rejecting all emails from unknown senders is a far too strict because there will be nothing to do when the mail is legitimate and sender is unknown.

Throttling

The throttling simply slows down the rate at which a single network or host can send traffic. Probably this is the most sensible way to fight spam. For example a legitimate mailing list may send out huge quantities of mail, but each message is addressed to different users on different networks. A spammer on the other hand may use dictionary attack, and tries to find valid email addresses on one network. In this case throttling can lead to a drawback for the spammers, but it also uses more resources from the legitimate senders. Unfortunately spammers more likely collect email addresses for sending spam letters to, and they use dictionary attack rarely what makes throttling rather a theoretical then a practical solution.

Content based filtering (KEYWORD)

One more solution is to search for keywords in the e-mail's subject. It means to scan the subject for words, related to spam letters. This is a simple language analysis, works only by match specific phrases. This method has unfortunately several problems, since the spam letters topic changes time by time. This can be handled by a keyword list that is regularly updated, but the smallest change in the words of the subject leads to mismatch (e.g. Write "softw@re" in spite of "software"). The simplicity of these spam filters led to a high false-positive rate and it had also a significantly high maintenance rate.

Bayesian filtering

To effectively combat with the above problems an adaptive new technique is needed. The answer lies in Bayesian mathematics. Bayesian filtering is based on the principle that most events are dependent and that the probability of an event occurring in the future can be inferred from the previous occurrences of that event. Bayes's theorem says that:

$$\text{posterior} = \frac{\text{prior} \times \text{likelihood}}{\text{evidence}}. \qquad (1)$$

The formula used by filter software is derived from above theorem is as follows:

$$\Pr(S|W) = \frac{\Pr(W|S) \cdot \Pr(S)}{\Pr(W|S) \cdot \Pr(S) + \Pr(W|H) \cdot \Pr(H)} \qquad (2)$$

Here

- $\Pr(S \mid W)$ is the probability that a message is a spam, knowing that the word is in it;
- $\Pr(S)$ is the overall probability that any given message is spam;
- $\Pr(W \mid S)$ is the probability that the word appears in spam messages;
- $\Pr(H)$ is the overall probability that any given message is not spam (is "ham");
- $\Pr(W \mid H)$ is the probability that the word appears in ham messages.

Particular words have particular probabilities of occurring in spam email and in legitimate 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. The user needs to generate a database with words and tokens collected from a sample of spam mail and valid mail (ham).

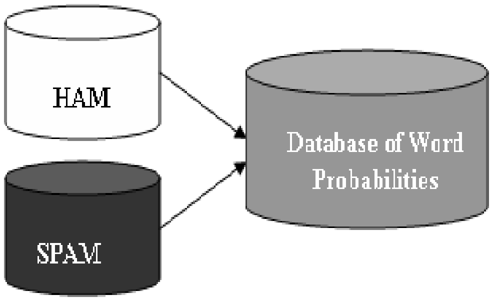

Fig. 2 creating a word database for the filter

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 word "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, once the ham and spam databases have been created, the word probabilities can be calculated and the filter is ready for use. When a new mail arrives, it is broken down into words and the most relevant words – i.e., those that are most significant in identifying whether the mail is spam or not – are singled out. From these words, the Bayesian filter calculates the probability of the new message being spam or not. If the probability is greater than a threshold, say 0.9, and then the message is classified as spam. Email marked as spam can then be automatically moved to a "Junk" email folder, or even deleted outright.

False Positives and False Negatives

There are four scenarios for an outcome when a spam filter or another countermeasure operates on an email:

- The email is ham and the filter correctly identifies the email as a genuine mail.
- The email is spam and the filter correctly identifies it as such.
- The email is not spam and the filter wrongly identifies the email as spam. This is called a false positive.

- The email is spam and the filter wrongly identifies the email as a genuine mail. This is called a false negative.

The two first items are the ones that we want to make happen. The two last items are the outcomes we do not want. To measure these we define false positive (FPR) and false negative ratios (FNR) as follows:

$$FPR = \frac{\text{Emails wrongly identified as spam}}{\text{Total emails}} \quad (3)$$

$$FNR = \frac{\text{Emails wrongly identified as ham}}{\text{Total emails}} \quad (4)$$

Measuring FNR and FPR can be difficult in a real life situation and there is trade-off between false positives and false negatives. A lower false negative will in most cases give a higher false positive. A spam filter should have as low false positive ratio as possible. This is because some spam getting through the spam filters are better than loosing genuine emails.

The working process of Bayesian Spam filter is shown below with the help of flow chart. The chart is divided in to two parts, picture1 & picture2.

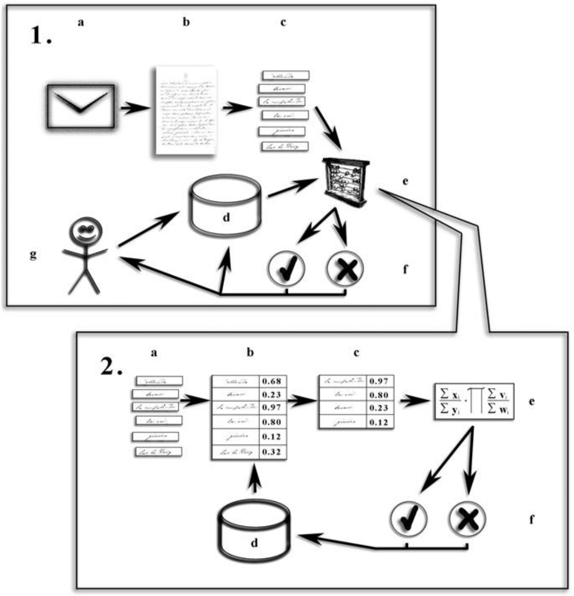
1.
a
b
c
e
d
g
f
2.
a
b
c
0.68
0.23
0.97
0.80
0.12
0.32
0.97
0.80
0.23
0.12
$\frac{\sum x_i}{\sum y_i} \cdot \prod \frac{\sum v_i}{\sum w_i}$
e
d
f

Fig. 3 Flow chart of Bayesian Spam Filter

- 1a: the letter is received.
- 1b: it is used as a plain text (including body and header).
- 1c: the letter is segmented into tokens (tokenization).
- The stored values are looked up from 1d database.
- 1e: the calculation is done.
- 1f: the result is sent to the user, while it updates the stored values in the database (1d).
- 1g: the users' feedback in case of misclassification, which helps to update the database.

  The calculation part is detailed on picture 2.
- 2a: the tokens are received.
- 2b: from the database (2d) the tokens get their stored values.
- 2c: only the most relevant tokens are used further.
- 2e: the calculation is done using the values of the most relevant tokens and other statistical data
- 2f: the final decision is made, and the database (2d) is updated relating to the result.

5. Neural Network Learning

One of the most interesting properties of a neural network is the ability to learn from its environment in order to improve its performance (measured through a predefined performance measure) over time. Learning in a Bayesian neural network stands for an iterative process of adjusting the synaptic weights and threshold values. Supervised learning networks have been the mainstream of neural model development. The training data consist of many pairs of input/output training patterns. Therefore, the learning will benefit from the assistance of the Spam Filter.

6. Conclusion

The Implemented collaborative methods of Spam filtering focused on developing techniques to identify and catch spam at the network gateway. The ultimate aim was to reduce the strain caused by spam on the network's internal infrastructure, in particular the mail servers, as they no longer have to process as many spam messages. Lowering the load on the network's mail servers leaves them with more time to perform their primary duty of forwarding legitimate emails to their intended recipients.

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. The Implemented collaborative methods of Spam filtering focused on developing techniques to identify and catch spam at the network gateway. The ultimate aim was to reduce the strain caused by spam on the network's internal infrastructure, in particular the mail servers, as they no longer have to process as many spam messages. Lowering the load on the network's mail servers leaves them with more time to perform their primary duty of forwarding legitimate emails to their intended recipients.

References

[1] Haykin S. "Neural Network" Second Edition, Pearson Education.
[2] Khorsi A. "An overview of content Based Spam Filtering Technique" Informatica 2007.
[3] Olivier G. "Bayesian Spam Filtering" September 11, 2008.
[4] Xin Jin; Anbang Xu; Bie, R.; Xian Shen; Min Yin, Granular Computing, 2006, "Spam email filtering with Bayesian belief network: using relevant words*",* IEEE International Conference on Volume 3, Issue 1, 10-12 May 2006 Page(s): 238 – 243.
[5] Le Zhang, Jingbo Zhu, and Tianshun Yao, "An Evaluation of Statistical Spam Filtering Techniques", ACM Transactions on Asian Language Information Processing (TALIP), Volume 3, Issue 4 (December 2004), Pages: 243 – 269, Year of Publication: 2004.
[6] http://www.symantec.com/searchlanding/antispam.
[7] http://www.ironport.com.
[8] http://www.spamassassin.apache.org.
[9] http://w2.syronex.com/jmr/ safe.
[10] Anders Wiehes, Master of Science in Information Security, Department of Computer Science and Media Technology, Gjovik University College, "Comparing Anti Spam Methods" 2005.
[11] Yegnanarayana B. "Artificial Neural Networks" , PHI Learning Pvt. Ltd.
[12] Anderson James A. "An Introduction to Neural Networks" 3rd Edition , The MIT Press.
[13] Paul Graham "A Plan for Spam" August 2002 http://paulgraham.com/spam.html.
[14] Better Bayesian Filter January 2003 http://paulgraham.com/better.html.
[15] Steven Hauser. ``Statistical Spam Filter Works for Me." http://www.sofbot.com.

Nibedita Chakraborty born on 3rd July 1986. She received her B.E. degree in Computer Science & Engineering from Guru Ghasidas University, Bilaspur (Chhattisgarh) in 2008. Presently she is working as a lecturer in the Department of Computer Science & Engineering in Kirodimal Government Polytechnic, Raigarh(Chhattisgarh). She is an associate member of Institute of Engineers, India having membership code- A-5533555.

Anjani Patel born on 26th June 1987. She received her B.E. degree in Computer Science & Engineering from ChhattisgarhSwami Vivekanand Technical University, Bhilai (Chhattisgarh) in 2007. Currently she is working as a lecturer in the Department of Computer Science & Engineering in Kirodimal Government Polytechnic, Raigarh (Chhattisgarh).