# Anti Spamming Techniques

Written by Sumit Siddharth

In this article will we first look at some of the existing methods to identify an email as a spam? We look at the pros and cons of the existing methods and what are the current challenges in this domain. This article also needs a special mention to Paul Graham, for his wok in this field and putting up perhaps the most comprehensive tutorials in this domain on his homepage.

I am sure that each one of us has faced this problem of spamming. Every morning when I open my inbox I spend most of the time either deleting the junk emails or reporting them as spam. The two biggest agenda which concerns developer of the anti spam filters are:
1. To differentiate between spam and legal emails (also known as ham) with maximum possible efficiency.
2. To have least amount of false positives. More is the accuracy of a spam filter more is the amount of trust a user will associate with it, and thus more often will a user not check his spam mail folders to look for a valid email which might have marked as spam due to false positives associated with the filter.  Needless to say that the false positives of a spam filter can cause more damage than making the life easier.

Here I would like to describe a few approaches on which the spam filtering engines mostly work
.
1. **Signature based:** This works somewhat like the antivirus. A kind of database of a large number of spam mail signatures is maintained and an incoming email is scanned for these signatures. If the signature matches with any of the signatures in the database then the mail is marked as a spam or else it is marked as a ham. The signature can be calculated based on different approaches.

E.g.
One way to calculate a signature for an email would be to assign a number to each character, then add up all the numbers. It would be unlikely that a different email would have exactly the same signature.

From a spammers perspective it is very trivial to get past these filters by just modifying the emails a little. One way to do it is to add random stuff to every email. Besides, there can be so many ways to bypass a particular signature. Although, using this approach it may be possible that a junk email may breach the filter and reach your inbox, however, it is very unlikely that you valid email which will be marked as spam which is much more critical. That is probably the main reason that these filters are still in existence and are often preferred.

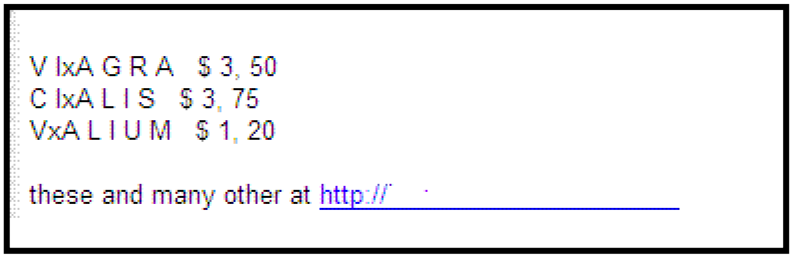
E.g. BrightMail: This is how the earlier BrigthMail used to work. It maintains a network of fake email addresses. Any email sent to these addresses must be spam. So when they see the same email sent to an address they're protecting, they know they can filter it out. Thus in this case signature happens to be the entire email which is received at the non existing email addresses.

2. **Score based**:  This is also called as a rule based approach. In this approach the spam filtering engine looks for certain words/characteristics in the entire email. Based on these individual words/characteristics a net score is calculated and attached/associated with the email. If the score crosses certain threshold then the email is marked as a spam. In addition to look for the characteristics of a spam email, some filtering engine also looks for the characteristics associated with a valid email (also known and ham) and thereby lowering the net spam score.

The Characteristics of a valid and a spam email may differ from person to person. In order to arrive at such a list of characteristics data must be collected from hundred’s of email addresses. These characteristics are mostly static in nature. Eg.

Assume a spam filter engine is looking for a keyword “Viagra” then a spammer could easily defeat this engine by modifying the word “Viagra” as “V1agra”. This can be modified in a number of ways and thus having the keyword is not sufficient and having all the modifications is also not so easy.

The screenshot below shows such an email which reached my inbox and was able to get through the gmail spam filter



VIxAGRA $3, 50
CIxALIS $3, 75
VxALIUM $1, 20

these and many other at http://

**Figure1**.

Spam email able to breach the spam filter and reach inbox. Notice that it was possible as the words Viagra, Cialis and other have been misspelled purposefully to bypass the filter.

The other problems which are associated with this approach is that the list of characteristics on which a filter is working is somewhat static in nature. Thus, it is virtually impossible to tweak these characteristics as per the individual users email preferences and also to the constantly changing spamming characteristics. Thus it is pretty easy for a spammer to bypass these filters. Not only this as spammers discover more and more new ways to modify a spam emails characteristics, more ineffective will such a filter be.

It is also worth mentioning here that the score associated with any individual characteristic of a spam or a ham [1]is although a good estimate but strictly speaking it is assigned an arbitrary number. Let me quote Paul Graham here

> [2]The problem with a "score" is that no one knows what it means. The user doesn't know what it means, but worse still, neither does the developer of the filter.

E.g.
SpamAssasin works on these principles.
This is how the rules/score mapping occurs in SpamAssasin.

| AREA TESTED | LOCALE | DESCRIPTION OF TEST | TEST NAME | DEFAULT SCORES (local, net, with bayes, with bayes+net) |
|---|---|---|---|---|
| **Body** | | Generic Test for Unsolicited Bulk Email | GTUBE | 1000 |
| **Body** | | Claims you were on a list | Excuse_11 | 1.072 0.146 1.334 0 |

To learn more about the how the rules are applied in SpamAssasin check http://wiki.apache.org/spamassassin/HowScoresAreAssigned

### 3. Bayesian Approach

This is a statistical based model and is considered to be far more effective and robust then the one described earlier. This is how it is defined

[2] *Source:* *http://www.paulgraham.com/spam.html*

[3]Bayesian spam filters calculate the probability of a message being spam based on its contents. Unlike simple content-based filters, Bayesian spam filtering learns from spam and from good mail, resulting in a very robust, adapting and efficient anti-spam approach that, best of all, returns hardly any false positives**.**

In contrast to the score based model described above which reads the spam and ham mail characteristics from a static file, the Bayesian spam filters build the characteristics themselves.

The **characteristics** a Bayesian spam filter can look at can be

- the **words** in the body of the message
- its **headers** (senders and message paths)
- other aspects such as **HTML code** (like colors)
- **word pairs, phrases** and
- **meta information** (where a particular phrase appears, for example).

Here is a brief description of how this approach works.

To start with we take one corpus of ham and another corpus of spam emails. It is generally a good idea to take an equal number of emails in both the corpuses. Then we scan all the emails in both the corpuses including all the fields highlighted above i.e. headers etc. Now here we define our Tokens. Tokens are the set of characteristics which will help us in differentiating an email between spam and ham. For example a keyword “Viagra” will have a probability of 0.9 that it is a spam where as say an email with keyword “securityfocus” will have a probability 0.9 that it is a ham. Thus we scan all the emails in both the corpuses for the number of occurrences of all such tokens. As mentioned above each token has its individualistic probability score. Thus based on bayes method we calculate the aggregate probability score.

[3] http://email.about.com/cs/bayesianfilters/a/bayesian_filter.htm

This score is the measure whether an incoming mail will be marked as a spam or as a ham. Remember a Spam probability of 0 implies the ham probability of 1 and vice versa.

Thus when new mail arrives, it is scanned into tokens, and the most interesting fifteen tokens, where interesting is measured by how far their spam probability is from a neutral .5, are used to calculate the probability that the mail is spam. Another key advantage of this filter is that the filter keeps increasing its efficiency every time a user pulls out a legal email from the spam mail folder to the ham folder. This helps the filter in identifying its mistakes and better identifying the individual characteristics of users spam and ham email. Thus it also takes into consideration that the while most people's spam may have similar characteristics, the legitimate mail is characteristically different for everybody. This is the reason why Bayesian filters are most effective and widely popular.

Bayesian filters vary in performance. As a rule you can count on filtering rates of 99%. Some, like SpamProbe, deliver filtering rates closer to 99.9%.

Quoting from SpamProbe official website about its features

> - Spam detection using Bayesian analysis of terms contained in each email. Words used often in spams but not in good email tend to indicate that a message is spam. Generally over 90% effective at detecting spam once a few hundred spams have been classified. My personal database is over 99% effective.
>
> - Automatically learns from incoming mails as they are classified. Incorporates user's feedback to tailor classification to each user's personal tastes

### 4. Challenge Response

These kinds of filters are generally for very paranoid users who just don't want to receive any spam but this comes at a cost. Every time such a user receives an email from somebody from whom he has never received an email before, an email is sent back to the sender asking him to resend by clicking the reply button in order for his email to reach the recipient. Thus although these filters are very effective in reducing spam, but because of their rude nature, it increases the work of the legitimate users.

## References

1. http://paulgraham.com
2. http://email.about.com/cs/bayesianfilters/a/bayesian_filter.htm
3. www.spamconferences.com
4. http://spamprobe.sourceforge.net
5. http://spamassassin.apache.org
6. http://www.brightmail.com/

## About the Authors

Sumit Siddharth:

Sumit leads the penetration testing team at NII consulting (www.niiconsulting.com.). Sumit is a GIAC certified intrusion analyst. He is also a graduate from IIT Kanpur.