

Why Bayesian filtering is the most effective anti-spam technology

Achieving a 98%+ spam detection rate using a mathematical approach

This white paper describes how Bayesian filtering works and explains why it is the best way to combat spam.

Introduction

This white paper describes how Bayesian mathematics can be applied to the spam problem, resulting in an adaptive, 'statistical intelligence' technique that achieves very high spam detection rates.

It also explains why the Bayesian approach is the best way to tackle spam once and for all, as it overcomes the obstacles faced by more static technologies such as blacklist checking, comparing to databases of known spam and keyword checking. These technologies are not obsolete, but cannot be relied upon without a Bayesian filter.

Introduction.....	2
Current spam detection techniques.....	2
How the Bayesian spam filter works	2
Why Bayesian filtering is better	4
About GFI MailEssentials	6
About GFI	7

Current spam detection techniques

Spam is an ever-increasing problem. The number of spam mails is increasing daily - studies show that over 50% of all current email is spam; the Radicati Group predicts this will reach 70% by 2007. Added to this, spammers are becoming more sophisticated and are constantly managing to outsmart 'static' methods of fighting spam.

The techniques currently used by most anti-spam software are static, meaning that it is fairly easy to evade by tweaking the message a little. To do this, spammers simply examine the latest anti-spam techniques and find ways how to dodge them.

To effectively combat spam, an adaptive new technique is needed. This method must be familiar with spammers' tactics 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 Bayesian mathematics.

How the Bayesian spam filter works

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. (More information about the mathematical basis of Bayesian filtering is available at Bayesian Parameter Estimation –

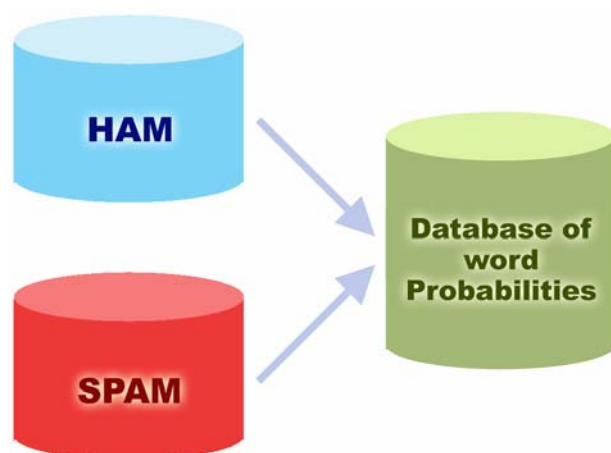
http://www-ccrma.stanford.edu/~jos/bayes/Bayesian_Parameter_Estimation.html

and An Introduction to Bayesian Networks and their Contemporary Applications - <http://www.niedermayer.ca/papers/bayesian/bayes.html>).

This same technique can be used to classify spam. If some piece of text occurs often in spam but not in legitimate mail, then it would be reasonable to assume that this email is probably spam.

Creating a tailor-made Bayesian word database

Before mail can be filtered using this method, the user needs to generate a database with words and tokens (such as the \$ sign, IP addresses and domains, and so on), collected from a sample of spam mail and valid mail (referred to as 'ham').



Creating a word database for the filter

A probability value is then assigned to each word or token; the probability is based on calculations that take into account how often that word occurs in spam as opposed to legitimate mail (ham). This is done by analyzing the users' outbound mail and by analyzing known spam: All the words and tokens in both pools of mail are analyzed to generate the probability that a particular word points to the mail being spam.

This word probability is calculated as follows: If the word "mortgage" occurs in 400 of 3,000 spam mails and in 5 out of 300 legitimate emails, for example, then its spam probability would be 0.8889 (that is, $[400/3000]$ divided by $[5/300 + 400/3000]$).

Creating the ham database (tailored to your company)

It is important to note that the analysis of ham mail is performed on the organization's mail, and is therefore tailored to that particular organization. For example, a financial institution might use the word "mortgage" many times over and would get a lot of false positives if using a general anti-spam rule set. On the other hand, the Bayesian filter, if tailored to your company through

an initial training period, takes note of the company's valid outbound mail (and recognizes "mortgage" as being frequently used in legitimate messages), and therefore has a much better spam detection rate and a far lower false positive rate.

Note that some anti-spam software with very basic Bayesian capabilities, such as the Outlook spam filter or the Internet Message Filter in Exchange Server, does not create a tailored ham data file for your company, but ships a standard ham data file with the installation. Although this method does not require an initial learning period, it has 2 major flaws:

1. The ham data file is publicly available and can thus be hacked by professional spammers and therefore bypassed. If the ham data file is unique to your company, then hacking the ham data file is useless. For example, there are hacks available to bypass the Microsoft Outlook 2003 or Exchange Server spam filter.
2. Such a ham data file is a general one, and thus not tailored to your company, it cannot be as effective and you will suffer from noticeably higher false positives.

Creating the spam database

Besides ham mail, the Bayesian filter also relies on a spam data file. This spam data file must include a large sample of known spam and must be constantly updated with the latest spam by the anti-spam software. This will ensure that the Bayesian filter is aware of the latest spam tricks, resulting in a high spam detection rate (note: this is achieved once the required initial two-week learning period is over).

How the actual filtering is done

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, then the message is classified as spam.

This Bayesian approach to spam is highly effective – a May 2003 BBC article reported that spam detection rates of over 99.7% can be achieved with a very low number of false positives!

Why Bayesian filtering is better

1. The Bayesian method takes the whole message into account - It recognizes keywords that identify spam, but it also recognizes words that denote valid mail. For example: not every email that contains the word "free" and "cash" is spam. The advantage of the Bayesian method is that it considers the most interesting words (as defined by their deviation from

the mean) and comes up with a probability that a message is spam. The Bayesian method would find the words "cash" and "free" interesting but it would also recognize the name of the business contact who sent the message and thus classify the message as legitimate, for instance; it allows words to "balance" each other out. In other words, Bayesian filtering is a much more intelligent approach because it examines all aspects of a message, as opposed to keyword checking that classifies a mail as spam on the basis of a single word.

2. A Bayesian filter is constantly self-adapting - By learning from new spam and new valid outbound mails, the Bayesian filter evolves and adapts to new spam techniques. For example, when spammers started using "f-r-e-e" instead of "free" they succeeded in evading keyword checking until "f-r-e-e" was also included in the keyword database. On the other hand, the Bayesian filter automatically notices such tactics; in fact if the word "f-r-e-e" is found, it is an even better spam indicator, since its unlikely to occur in a ham mail. Another example would be using the word "5ex" instead of "Sex". You would probably not have a word 5ex in a ham mail, and therefore the likelihood that its spam increases.
3. The Bayesian technique is sensitive to the user - It learns the email habits of the company and understands that, for example, the word 'mortgage' might indicate spam if the company running the filter is, say, a car dealership, whereas it would not indicate it as spam if the company is a financial institution dealing with mortgages.
4. The Bayesian method is multi-lingual and international - A Bayesian anti-spam filter, being adaptive, can be used for any language required. Most keyword lists are available in English only and are therefore quite useless in non English-speaking regions. The Bayesian filter also takes into account certain languages deviations or the diverse usage of certain words in different areas, even if the same language is spoken. This intelligence enables such a filter to catch more spam.
5. A Bayesian filter is difficult to fool, as opposed to a keyword filter - An advanced spammer who wants to trick a Bayesian filter can either use fewer words that usually indicate spam (such as free, Viagra, etc), or more words that generally indicate valid mail (such as a valid contact name, etc). Doing the latter is impossible because the spammer would have to know the email profile of each recipient - and a spammer can never hope to gather this kind of information from every intended recipient. Using neutral words, for example the word "public", would not work since these are disregarded in the final analysis. Breaking up words associated with spam, such as using "m-o-r-t-g-a-g-e" instead of "mortgage", will only increase the chance of the message being spam, since a legitimate user will rarely write the word "mortgage" as "m-o-r-t-g-a-g-e".

Bayesian filters or updated keyword lists?

Some types of anti-spam software regularly download new keyword files. While this is, of course, better than not updating keyword lists, the fact is a rather patchy approach that is easily circumvented. Downloading updates makes it a little bit harder, but the principal system is

flawed compared to a Bayesian filter.

What's the catch?

Bayesian filtering, if implemented the right way and tailored to your company is by far the most effective technology to combat spam. Is there a downside? Well, in a way there is one downside, but this can easily be overcome: Before you can use and judge the Bayesian filter, you have to wait for it to learn for at least two weeks – that or create the ham or spam databases yourself. This task can be quite complex, so it is best to wait until the filter has had time to learn. Over time, the Bayesian filter becomes more and more effective as it learns more about your organization's email habits. To quote the old saying, good things come to those who wait.

It is important, therefore, to keep this in mind when evaluating anti-spam software. If the product has advanced, customized Bayesian analysis, then it can only be judged after a few weeks. It is probable that basic anti-spam software might perform better initially, but after a few weeks the Bayesian filter catches up and well outperforms the conventional anti-spam filters once and for all.

About GFI MailEssentials

GFI MailEssentials for Exchange/SMTP offers spam protection at server level and eliminates the need to install and update anti-spam software on each desktop. GFI MailEssentials offers a fast set-up and a high spam detection rate using Bayesian analysis and other methods - no configuration required, very low false positives through its automatic whitelist, and the ability to automatically adapt to your email environment to constantly tune and improve spam detection. It also enables you to sort spam to users' junk mail folders. GFI MailEssentials also adds key email tools to your mail server: disclaimers, reporting, mail archiving and monitoring, server-based auto replies and POP3 downloading. More information and a full evaluation version are available at <http://www.gfi.com/mes/>.

About GFI

GFI Software Ltd. is a leading provider of network security, content security and messaging software. Key products include the GFI FAXmaker fax server software for Exchange and SMTP servers; GFI MailSecurity email security software for Exchange and SMTP servers; GFI MailEssentials server-based anti-spam software; GFI LANguard Network Security Scanner (N.S.S.) security scanning and patch management software; GFI Network Server Monitor network management software; and GFI LANguard Security Event Log Monitor (S.E.L.M.) that performs network-wide event log management and auditing. Clients include Microsoft, Telstra, Time Warner Cable, Shell Oil Lubricants, NASA, DHL, Caterpillar, BMW, the US IRS, and the USAF. GFI has offices in the US, the UK, Germany, Cyprus, Romania, Australia and Malta, and operates through a worldwide network of distributors. GFI is a Microsoft Gold Certified Partner and has won the Microsoft Fusion (GEM) Packaged Application Partner of the Year award. For more information about GFI, visit <http://www.gfi.com>.

© 2005 GFI Software Ltd. All rights reserved. The information contained in this document represents the current view of GFI on the issues discussed as of the date of publication. Because GFI must respond to changing market conditions, it should not be interpreted to be a commitment on the part of GFI, and GFI cannot guarantee the accuracy of any information presented after the date of publication. This White Paper is for informational purposes only. GFI MAKES NO WARRANTIES, EXPRESS OR IMPLIED, IN THIS DOCUMENT. GFI, GFI FAXmaker, GFI MailEssentials, GFI MailSecurity, GFI LANguard, GFI Network Server Monitor, GFI DownloadSecurity and their product logos are either registered trademarks or trademarks of GFI Software Ltd. in the United States and/or other countries. All product or company names mentioned herein may be the trademarks of their respective owners.

