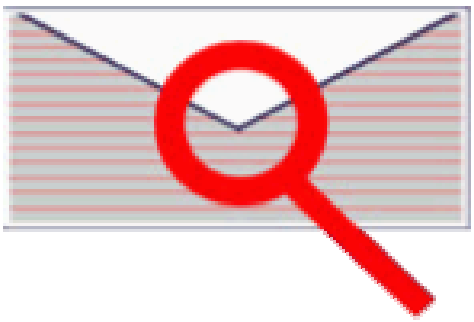
Judging A Spam Filter

The Mail-Filters Way



June 11, 2003

Judging a Spam Filter – The Mail-Filters way
Document version 1.02

Mail-Filters.com, Inc.
205 De Anza Blvd #200
San Mateo, CA 94402
650-655-7700

How To Judge A Good Filter

In the final analysis a spam filter should do two things well - with minimum user and system administrator effort.

1. It should catch a very high percentage of the spam that passes through it.
2. It should seldom mistake a legitimate message as spam – almost no false positives.

The Mail-Filters solutions detect over 95% of the spam they filter while generating less than one in 100,000 false positives. They do it by building into the filters the human ability to really identify spam; the user or the system administrator is not expected to do it.

The Challenge For A Good Spam Filter

A spam filter and an airport metal detector face many of the same challenges. At an airport, if the detector is too sensitive, everyone gets stopped. If it is not sensitive enough, undesirable objects get through. The airport uses human intervention to handle this problem. The detector sensitivity can be set fairly high to catch as much metal as possible, while an inspector using a metal-detecting wand detects false positives - innocent people with hip replacements, for example, who set off the alarm.

Just as there are some objects that only a human can recognize for sure, spam also requires human recognition. What did the US Supreme Court justice say? “I can’t define pornography, but I know it when I see it.”

It follows that spam filters must rely on some level of human intervention, “to know spam when they see it”. There are typically two ways of doing this: have the user recognize it, or build the human recognition into the filter.

<u>Have the user recognize spam</u>

Users can intervene by doing one or more of the following:

1. Set the filter to be fairly lenient so some spam gets through, then delete what they don’t want from what did get through.,
2. Set the filter to be rigorous where it misidentifies some legitimate e-mail, then search through the folder where the suspected spam has been diverted to retrieve the false positives.
3. Select the type of filter where they have to learn enough about spam and spammers to be able to program the filter to recognize what is spam.

The problem with any of these three user interventions is the amount of work they require of the user. Deleting uncaught spam or retrieving misidentified good messages is a load that will only increase in the future as the number of spam messages increase and spammers become more sophisticated, As for teaching the filter to recognize spam, only a devoted anti-spam user is willing to treat the anti-spam filter like a high performance but temperamental car to be worked on constantly to keep it running. To most users the spam filter is a tool, not a hobby.

Build the user recognition into the filter

The second option, building the human recognition into the filter, not only reduces the work a user has to do, it actually improves the filters ability to catch spam while avoiding false positives. Here is how this is done in the Mail-Filters solutions.

Human Recognition Built Into The Mail-Filters Spam Filter.

One only has to read a hundred or so spam messages to realize that spam comes in many different forms. Moreover, it also becomes clear that those forms change constantly as the spammers come up with new ways to defeat spam filtering techniques. These two facts lead to two conclusions:

1. It takes a large database of spam characteristics to identify most of the spam received by most users.
2. That large database must be updated regularly, just like an antivirus filter, to keep up with the spammers' constantly evolving tricks.

A Spam Recognition Database

The Mail-Filters solutions are based on databases comprised of tens of thousands of spam signatures. Editors trained to see patterns in spam generate these databases by reading actual spam messages. They include information based on key phrases or words, source of the spam, or specific action requested of the recipient.

These databases are at the heart of the Mail-Filters filtering system. In fact, the system applies 11 categories of tests. Each category is run at three levels: global, domain and mailbox. In each test we match the sending e-mail information against databases that are kept for each level. The tests include Spam signature checks, White List checks, header checks and content checks for misleading information

The tests are as follows:

1. IP Address Spam Signature – Match the sending IP address against known spammer IP addresses

2. IP Address White List – Match the sending IP address against known non-spammer IP addresses. These lists are usually customer requested.
3. Reverse DNS Spam Signature – Match the reverse DNS of the server that sent the e-mail against RDNS that we have determined to be from known spammers.
4. Reverse DNS White List – Match the reverse DNS of the server that sent the e-mail against RDNS that we have determined to be from known non-spammers (usually determined by the customer).
5. HELO Name Spam Signature. – Match the HELO name of the sending server against HELO names that we have determined to be from known spammers.
6. HELO Name White List. – Match the HELO name of the sending server against HELO names that we have determined to be from known non-spammers. (usually determined by the customer)
7. Mail from Spam Signature – Match the sender's mail address (mail from) against our database of known spammers or recipients that the receiver does not want to receive mail from. We can also block all mail from a particular domain that a customer requests.
8. Mail From White List – Match the senders e-mail addresses (mail from) against our database of senders that the receiver wants to get mail from. This list is usually customer requested and can be used for major customers. We can also white list entire domains in the 'mail from' category.
9. Subject Spam Signature – Match words in the subject of the e-mail against phrases we believe would only be in a spam e-mail. This is an area where customers can get more aggressive (to stop a particular type of message) than the global rules.
10. Body Spam Signature – Match words in the body of the e-mail against phrases we believe could only be in a spam e-mail. This is another area where customers can get more aggressive (to stop a particular type of message) than the global rules.
11. Public Spam Signatures – Matching against these lists is optional. Because these lists are not maintained by Mail-Filters, we default this test to be off unless requested otherwise by the customer.

Once a test has identified a piece of spam the testing process is discontinued and the message is marked as spam. If there is no match the message is deemed not to be spam.

Updating The Database

Spammers know that a good filter with built-in human knowledge will catch practically all their spam while generating few false positives, so they constantly change the form of that spam. To keep up with these changes the Mail-Filters editors who created the signature databases update them constantly, usually within minutes of new spam hitting the internet. These new signatures are

incorporated continuously into customer e-mail testing to provide up-to-the-minute protection from spam.

Judging A Spam Filter – A Final Word

The best spam filters have human knowledge of spam built into them. Unfortunately it is cheaper for a filter maker to get the user or the customer's system administrator to provide the human intervention to make their filter really effective. Don't be confused by claims of spam filter performance. It is essential to judge a filter by asking very specific questions, such as:

- "What percentage of spam does your filter catch?"
- "How many false positives are generated per 100,000 messages?"
- "Can it achieve both of these on the same setting?"
- "How is the spam detected?"
- "How is the filter updated to catch constantly evolving spammer tricks?"
- "What exactly must I or the system administrator do to maintain this performance?
- How much of our time will it take?

Ask that all these questions be answered for the filter as it processes a sizable, unbiased sample of e-mail. As said earlier, many filters can be adjusted to catch a lot of spam, but at the expense of high false positives. To repeat, catching most of the spam, while generating few false positives, with little use of the user's and administrator's time is the real test.