Detecting Spam

May 4, 2004
By Gabrielle Gagnon

More than 60 percent of Internet traffic is spam, so it's not surprising that people are turning to the science of big numbers for help. Many hope that Bayesian filters, based on algorithms that use probability to block unwanted messages, will stop this cyberpestilence.

Bayesian filters are more or less based on the Bayes rule, a theory of conditional probability that estimates the likelihood of an event (hypothesis) given the certainty of another event (evidence). Basically, the rule says that the likelihood of an event occurring in the future can be inferred from the number of times it occurred in the past.

Applied to spam, this means that if you break a message down into discrete elements (words, HTML tags, URLs, whatever) and you find that particular elements recur frequently in spam and not in ordinary mail, you can be reasonably confident that messages containing them are spam.

A typical Bayesian filter is a client-side, e-mail plug-in with a built-in database for collecting messages (evidence) and an inference engine for assigning probability (confidence) ratings. As messages arrive, it rates them, vetting individual elements and assigning a composite rating to each message as a whole, and copies them to the database. If a given rating indicates that a message is probably spam, the filter blocks it from the in-box. Users can flag any spam that gets through.

Because Bayes's theorem requires looking at all the evidence, Bayesian filters look at an entire message and compare it with both spam and nonspam samples to arrive at their numbers. This not only helps prevent false positives but also catches flags that don't occur to people. (Surprisingly, ff0000, the HTML code for bright red, is as much of a spam indicator as any pornographic term.)

The filters aren't fooled by cheap tricks, either. Extra characters inserted into known spam words (as in "S*E*X!!!") will not pass, because such things do not appear in ordinary correspondence.

Bayesian filters can also be tuned for individuals—a plus if you're a mortgage broker and your regular mail looks more like spam than most people's—but training them takes time. To achieve an accuracy rate better than 99 percent, you might have to classify thousands of messages. Fortunately, filters often come pretrained.

Does this mean Bayesian filters will put spammers out of business? No. It doesn't even mean they will keep your mailbox spam-free. You'll always see some spam, and if you train your filter from scratch you'll see lots of it. That's partly because Bayesian filters are designed to prevent false positives—blocking good mail is a greater sin than letting spam through—but it's also because they need training.

Someone—either you or the vendor—has to tell the filter what is and isn't spam. Once this happens, the filter can do an excellent job of evaluating items similar to those it has seen before. But when it sees something new, it rates the unknown either as neutral (50/50) or as slightly more likely to be not spam. Without other evidence, innovative spam gets through.

You have to be conscientious. If you don't identify spam as such, or if you flag good mail as spam by mistake, the filter's rating system could become corrupted. It can also become corrupted if disguised spam fools you into classifying it as good mail. Finally, just because a filter is labeled Bayesian doesn't guarantee that it applies all of Bayes's precepts or that it takes a purely scientific approach. To minimize the chance of losing good mail, designers make simplifying assumptions and give varying weights to different evidence, and these variations affect performance.

Bayesian filters are a useful weapon in the war on spam, but they're not perfect. The best solutions employ other techniques as well, such as white- and blacklists, honeypots, and community filtering. For more information on antispam techniques and how effectively antispam tools use them, see "Spam Blockers" at www.pcmag.com/article2/0,1759,1514410,00.asp .

Gabrielle Gagnon is a longtime contributor to PC Magazine.