

Commentary

Best Practices for Implementing Spam Control

Many of the eager new anti-spam vendors in the market are promising to catch 100 percent of the spam. Don't believe it. Even if this were achievable, you don't want them to try.

Spam-catching is a very tricky game. It's a dimension more difficult than anti-virus. An anti-virus laboratory makes a binary decision: Is it a virus? Is it not a virus? Yes. No. And we want the lab to be right every time.

The anti-spam game, however, is triage. There are messages that can be clearly identified as spam, and messages that can be clearly identified as not-spam, and then there is a pile of messages left over that are neither black nor white, they are gray. What are we going to do with the gray pile? What is in it, and why is it so tricky?

"Good Guys" vs. "Bad Guys": There are four categories of spam (see "Keeping Spam Out of Your E-Mail"):

1. Confidence games, pornography and unethical senders
2. Chain letters, hoaxes and urban legends
3. Legitimate offers from legitimate senders
4. "Occupational spam" from your colleagues

The job at the boundary is to sort out the good guys (No. 3 and No. 4) from the bad guys (No. 1 and No. 2). If it's from one of the bad guys, you can request removal from their mailing list ("unsubscribe"), but your request will only serve to validate your address, and you will receive more spam. If it's from one of the good guys, they should be practicing ethical permission-based marketing: In other words, you can safely unsubscribe and the sender should politely stop sending additional messages.

Best Practices for Controlling Spam

Tuning for the Business: You can't just look for specific banned words and expect to make the right decisions every time. For example, if you have a department working on a cure for prostate cancer or designing bicycle seats, there may well be some anatomical terms that will come up in the normal course

Gartner

of doing business that might also be on the dirty-words list. Rules that perform some contextual analysis can help: if the word "breast" is in the same sentence or paragraph with the word "cancer," then it's okay.

Spicy Language: It's one thing to ask your employees not to use spicy language when corresponding with a client, but it's hard to ask your clients not to use it. If an angry client writes a fiery message full of inappropriate language, should this be identified as spam and quarantined, or should it be handled in some other way? The answer is a business decision.

Foreign Words and Phrases: One group of attorneys was experiencing rejection of their messages sent to a number of clients, until they finally realized it was because they had a lot of smart people on staff who had graduated magna cum laude. There's a Latin word in there that also happens to appear on the dirty-words list. Another client with offices in Sweden discovered that the word "slut" in Swedish means to complete the project. The enterprise had to adjust its filters to permit this word.

Rate of False Positives: Some clients are asking what the rate of false positives is for a given product. A "false positive" is a message "captured" or flagged as spam that was a bona fide message that you don't want to lose. You should be able to tune the capture rate of a product and work with it so as not to have any false positives. But that will mean you will have "false negatives." Because this is not a perfect science, it may well require human review to make the final determination.

- **Formula to Calculate Spam Capture:** The capture rate (CR) and false positive rate (FP) are related by a quotient based on the efficiency of the filter system. As CR increases, so does the FP. No product can achieve $CR=100$ and $FP=0$. If the vendor makes this claim, it does not fully understand the complexity of the game. You can approach $FP=0$ by tolerating a lower CR, or you can maximize CR by tolerating an elevated level of false positives, but you can't have both.

The best anti-spam products have the highest rates of confidently identifying spam (for example, we've seen it before). No matter how elegant, systems that examine the header and body will not always be right, but they should tell you how confident they are that this is spam (for example, 85 percent sure vs. 55 percent sure this is spam). When dealing with the "gray pile," you will need tools to tune the system for your business. Based on the point scale or confidence rating, you can participate in deciding how a given message should be handled. Through 2003, e-mail spam capture rates in excess of 85 percent will result in false positive rates of 5 percent or greater (0.8 probability).

Developing a Comfort Level: When you implement any anti-spam product or service, you should never begin deleting flagged messages on Day 1. You should stage the implementation, and enroll your users in the validation process. Keeping an open and courteous channel of communication with your employees is an essential part of any spam-control program.

- **Reporting:** You can use reporting to help size the problem and possibly test the rules. If you were to turn on this rule set today, how many messages would have been deleted? How many would have been quarantined? How many would require human review? How many staff would be required to do the human review?
- **Marking headers:** The second step is to turn on the rules, but only mark the headers. At this point, you can see the decision-making in action, and can evaluate whether you agree. Again, you are not yet stopping messages, but are evaluating in combination with the reporting what level of effort would be required if you were to stop these messages.
- **Quarantine:** As a third step, quarantine the most egregious messages, those with the highest point score or confidence rating (most probably spam). For example, quarantine all messages with confidence ratings greater than 85 percent. Let that run for a week or so and validate with the users

how that looks. Gradually increase the number of messages quarantined by using a lower confidence threshold: 80 percent confident, then 75 percent confident, then 70 percent confident, possibly putting the newest 5 percent into a separate pile so that you can review it closely, watching for false positives. Perhaps, by creating an additional rule, you can prevent the false positives you find, as in the examples above.

As you approach midrange, you need to decide what approach to use for the messages with confidence ratings of 45 percent to 65 percent. What do we do with these? Clearly, the automated system cannot make the final decision. You need to make the decision that's right for your environment. Often, our clients simply mark the header, note that there is a possibility this is spam, show the reasons for that concern, and let the end user decide. At this point, the most-offensive messages should have been deleted, and the possibility that this is a legitimate message is also high.

Choices for this group are:

- Mark and pass through
- Mark, pass through, and send a copy to a collection for review. You may find that with some additional rules you could change the odds
- Quarantine and review. Once you have quarantined a message, especially one with such a low probability that it is spam, you have an obligation to review these messages promptly and send them on their way or declare them spam, so you have to be prepared to deal with the quantity of messages in this category.

Bottom Line: No one knows your environment better than you. No canned set of rules can know everything about your business. You will need to work with the anti-spam product or service to train it for your environment, and decide how to deal with the "gray pile" of messages that can't be clearly judged as spam or not-spam.