



# Why Spam Filters Fail

By Roger Matus, Chief Executive, Audiotrieve, LLC

© October 2003 Audiotrieve, LLC

**Spammers easily beat traditional filtering methods. Filters that implement the latest language-based technologies can help you succeed.**

---

Any good spam filter must eliminate the email you don't want to see while making sure you don't miss what you do need to see. If you still wade through significant amounts of spam or if you need to check all of the messages that have been categorized as spam (to make sure that there are no mistakes), then your filter has failed. You are wasting time and may miss something important.

## ***The Three Most Common Failures***

Unfortunately, no matter how good traditional spam filters may become, they are ultimately doomed to failure. That is because the overwhelming majority of the more than 100 spam filters on the market today depend on one of three basic methods to stop spam:

- Challenge-Response Systems
- Rules and List Systems
- Community Voting Systems

All of the anti-spam approaches listed above can be beaten by spammers. These methods either allow too many spam messages or cause unintended consequences, such as blocking important messages that you need to see.

Why are these traditional systems doomed to failure? To answer the question, it is important to understand how these traditional spam filters work.

## ***Challenge-Response Systems***

Challenge-Response systems are simple to operate and understand even for naïve users. Therefore, they became very popular and widely deployed. Unfortunately, the trade-offs are significant and the risk of losing an important message is quite high.

The premise behind Challenge-Response is simple: Good email comes from people or companies that you know or from people who are willing to justify sending you mail. Bad email comes from bulk mailers and people who will not justify themselves. Therefore, the theory is that if you create a list of good email sources (a so-called "white list") and challenge everyone else, you will eliminate spam.

How do you build a white list? The most popular approaches are easy, too. One way is to add the email addresses of everybody to whom you have sent a message. The concept is that if you sent a message to someone, then you are willing to receive a message from that person. If your email product, such as Microsoft Outlook, has a contact list, then all the names in the contact list can be added. This assumes

that you are willing to receive an email from any contact, even if you never sent that person a message.

What if a person is not on the list? That is where the "challenge" part comes in. The software sends an email message to the sender asking the sender to respond. Sometimes the response involves visiting a web site. Often, the system will display a special code in a distorted type font that a human can decipher, but a machine will not recognize. The theory is that people will reply with the correct code and that machines cannot.

The spam removal rate is extraordinary - sometimes 100%. People see an immediate impact without any complicated set-up. Some challenge-response systems win reviews with high marks for spam removal and ease-of-use.

Unfortunately, 100% spam removal comes at an extremely high price - many good messages will be removed as well. This is a significant problem for individuals and makes challenge-response systems impractical for enterprise deployment.

What kind of good messages will be removed (also called "false positives")? All of the following are at risk:

- **Subscriptions:** You probably receive newsletters, automated updates and promotions from companies and organizations that you care about. You might therefore get messages from a competitive newsletter or from an industry association. It might also come from a local school or community center. Generally, the automated bulk mailer will not be able to respond to challenges. None of these addresses will be in your "sent" folder or contact list. So, they must be entered manually. Unfortunately, the address used to send the subscription can change and messages that were once OK will be blocked.

- **Automated Mailing Systems:** Messages from automated systems are more common than you may think. For example, if you use an on-line service to book an airline flight, the confirmation needed for security processing is usually mailed from an automated system. Bank notices and other important messages may also be blocked.
- **Wireless Device Users:** Some users may not be able respond to a challenge because their device does not support web browsing. For example, suppose the system blocks an urgent message sent from the Blackberry wireless email device. The user may be challenged to visit a web site in order to enter a code. These users will not be able to reply and the message will be unnecessarily delayed.
- **Visually and Physically Challenged PC Users:** Some challenge-response systems display a code in an unusual font so that they cannot be deciphered by a bulk mailing system. The problem is that this technique prevents systems that support challenged users from working as well. Newer techniques enable challenged users to hear a recording of the code or to call a special operator for authentication. Such systems require extra work.

### **Rules and Lists**

Some filters use rules and lists to define spam. Early spam filters created simple rules by scanning each message for the presence of a trigger word or phrase. For example, if a pornographic term appeared within the text, the message would be automatically deleted or marked as spam.

Unfortunately, the weakness of this approach became immediately apparent as spammers quickly got around the simple rules by changing the text and even by deliberately misspelling words. For example, the word "lover" could be changed to "l0ver" or "1lover" or "l-o-v-e-r", which a human would under-

stand, but a computer would not recognize as a match.

Naturally, the rule makers responded by updating the rule list stored on each client PC. Unfortunately, spammers probably own all of the major anti-spam products and receive the updates at the same time as you. The spammers can easily modify their message so that the next salvo will bypass even the newest filters. This creates a never-ending cycle of creating, modifying, and maintaining filters.

New kinds of spam, including financial scams and messages that promote products for your body, made the rule list impossibly large to manage. The filters started to block valid messages about financial issues from your bank or health issues from your doctor.

To counteract these problems, the next generation of anti-spam software assigned point values to trigger words or phrases. Messages were flagged as spam when the total value of all the points in a message passed a threshold. However, this technique did not solve the problem of changing text or misspelling words.

Over time, the rules became more complicated as new tricks were used by the anti-spam companies to block messages by detecting flaws in the message headers. For example, a well-known error in the StealthMailer software always set the time and zone to "-0600 (EST)". A simple rule to block messages with that time in the header would block StealthMailer messages. Of course, such a rule could also block legitimate mail sent at that time if other steps were not taken.

Lists were also added as a way to identify spammers. These lists may contain the email or network addresses of spam sources and relay sites that they have identified. When you receive a message from these sources, the spam blocking software prevents delivery of the message.

*"One-size fits all" methods can't differentiate between the mail that an engineer might want to see versus the mail a market researcher might need to see.*

This technique is always one step behind the worst spammers. They can easily change locations in order to bypass the block, or impersonate a real site with a fake "from" address.

One side effect is that legitimate email sites have been blocked by these lists because the legitimate sender might have been impersonated or they might use a relay site that had been used by a spammer. Many senders will not even know that their message has been blocked. The lists are often widely spread across the network. Some of them may be copied for free from public sites. Therefore, it is often difficult and costly for the legitimate email sender to correct mistakes to these lists.

In the final analysis, systems based on rules and lists are fairly easy for spammers to defeat.

Rules and lists are created in response to an initial flood of messages. Developers can only update the anti-spam product after the harm is already done, identifying only messages that use the same technique or source.

Enterprise network-level spam solutions tend to be either rules or list based and have all of the same problems. These "one size fits all" methods can't differentiate between the mail that an engineer might want to see versus the kind of mail a market researcher might need to see. They tend to either be coarse filters, which still allow considerable spam to get through to the desktop, or they are so restrictive that wanted mail is often blocked. In addition to being relatively costly, they can require considerable daily administration.

## **Community Voting Systems**

Community voting is a clever idea that gets around the need to create rules and lists. When a message arrives, users click a button to identify a message as spam. That click is sent as a vote to a central server. If enough votes are cast, the central server identifies the message as spam and blocks it. Sophisticated voting systems weight the response from each voter by a credibility score earned over time. (The more times a person identifies spam that is eventually removed from a system and does not identify a message as spam which is not removed from the system, the more credible the user.)

This approach has several weaknesses.

As in the case with rules and lists, the spam message is already delivered to many people before it is identified as spam and the block is distributed from the central server to the client.

To fool these systems, many spammers use an automated process to make several changes in the spam message so that a computer will not consider them to be identical. This constant morphing of the message is a significant problem for anti-spam filters.

There have been many cited circumstances where a community could block a message that an individual would want to see. For example, you may be an avid golfer and you would like messages about the latest courses, equipment and discounts. Unfortunately, if the community does not share your interest, you may find that message may be blocked as spam. Legitimate newsletters can be blocked by community filters because some users no longer wish to receive them.

Finally, some versions of community-based filters require ready access to the Internet in order to process the high volume of information that comes to and from the central server. In some cases, the software requires an "always open" port to the Internet,

which may be vulnerable to security attacks. In all cases, the user should find out from their community voting anti-spam product vendor exactly how much Internet traffic is created by each PC (inbound and outbound) and what steps are taken to foil hackers.

---

## **How Spam Filters Can Succeed**

U.S. Supreme Court Justice Potter Stewart once commented that he could not define pornography, but he explained, "I know pornography when I see it." In other words, he could only define pornography by example.

The same could be said about defining spam. In this case, it is pornography, adult content, medical enhancements, and financial scams that cannot be defined by a set of rules that would be accepted by everyone. However, each individual can define spam by example.

The newest trend is for anti-spam filters that learn from examples. Many of them, like InBoxer, use Bayesian mathematical analysis to filter spam.

## **Bayesian Analysis**

The term Bayesian, when used to describe a spam filter, generally means two things:

- The filter uses examples of actual email and spam messages, as well as other information, to classify new mail into categories.
- The filter uses Bayesian mathematics to calculate how likely it is that each received message fits into the category.

**What to look for in Bayesian filters**

- *Well-chosen spam database for immediate good results, using both messages from the user's own Inbox and carefully selected data.*
- *Each user defines spam for him/herself with minimal tinkering time and the filters immediately adapt.*
- *Filters constantly learn and adjust for increasing success.*
- *Users identify good messages, as well as bad.*
- *Review folder for quick evaluation of uncertain mail, without the requirement to review all messages.*
- *No automatic message deletion.*

Thomas Bayes (1702-1761), a Presbyterian minister from Tunbridge Wells, 35 miles southeast of London, had no idea that his work would be used to fight spam more than 300 years after his death. He demonstrated how to categorize items by analyzing examples and other information.

Here is an example. If traditional rules were used to categorize playing cards, a rule for a "picture card category" might be "if a card is greater than 10 and less than an ace, it is a picture card." These rules depend upon your knowing something about the value of cards and all of the kinds of cards that might be available.

In contrast, the Bayesian approach would be to make observations about each card. For instance, if the observation was "the number of colors on the card," then in general picture cards would have a high number of colors, while other cards would be associated with low numbers. If the card to be categorized has more than one color, then the probability that it is a picture card would be expected to be very high. In this case, if a new card with an unknown value appeared, a Bayesian analysis could be used to determine whether the card is probably a picture card.

Bayesian analysis are not restricted to one type of observation. You can use other information, not just the colors on the card, to help categorize the card.

For example, you may note that picture cards all have images on them or that none of them have a number on them. Any observation can be used to make the estimate even more accurate.

It is easy to see that a Bayesian analysis enables you to make decisions based upon new types of information.

**Do All Bayesian Filters Work the Same Way?**

There is no industry standard for a Bayesian anti-spam filter. While the Bayesian mathematical formulas are well-known and understood by mathematicians, the way in which they are implemented is critical to success. It is similar to having a recipe from a famous chef. The formula may be well known and printed in a book, but the quality of the ingredients and the technique used to prepare them are critical for success.

Some of the key critical success factors for Bayesian filters are the source of your examples and the observations you can make on the examples.

**SOURCE OF THE EXAMPLES**

If spam can only be defined by example, then the quality of the examples used to create the definition is very important. If the samples are not representative of the type of mail that you receive, as can be the case with some filters, you may not get accurate results.

Generally, there are two factors to consider: the initial start-up and continual learning.

- **Startup:** The results of Bayesian systems vary significantly at start-up, based on the initial definition of spam. Some filters come with a pre-programmed set of "spam models" that are the same from user to user. If they are not modified,

such systems always result in a high error rate, as each person's definition of spam is different.

For other types of filters, generally those created by Bayesian purists, you must collect hundreds of examples of good messages and also hundreds of examples of spam messages before you can create the spam models. Filters based on your own email folders do create the best results. However, it can take some time to collect and properly categorize hundreds of messages. In addition, those messages must be truly representative of messages that might be received year-round. For example, if no pornographic messages arrive during the collection period, then the initial filters will not block them.

A third kind of system combines the two techniques to create the most accurate "out of the box" experience for most users. These systems base the filters on your own email messages, but also add a database of carefully selected statistics if there are not enough spam messages. This lets the filters cover a broad range of spam collected over a long period.

- **Continual Adaptation:** While all Bayesian filters are based on examples, the best continually update the filters because these are able to learn from their mistakes. In this way, they do not need to wait for an update from a central source and they can immediately adapt to new types of spam.

The process for continual learning is as follows: If the spam filter mistakes a spam message as good, the user has an easy way to identify it. The best filters then immediately update all of the relevant information used to analyze all future messages. In this way, the best messages are not likely to make the same mistake in the future.

Even further, the best spam filters can become more accurate without user interaction just by processing new spam as it arrives. For example, a new type of spam message may arrive. Since it is

new, a Bayesian filter may calculate that it is 90% similar to spam messages it has already seen. In this case, it would identify the message as spam because it is so much like known spam. To take it one step further, the filter can also learn about the new type of spam from the new message as it enters the spam folder. This enables the best filters to automatically adapt to new types of spam without updates from a centralized server.

### OBSERVATIONS ABOUT THE EXAMPLES

While the source of the examples used to make the Bayesian filters is very important, it is also critical to organize them correctly. The examples must be correctly sorted into spam and not-spam categories. Then, the correct observations must be made about each category. In the playing cards example, the sorting observations might include the number of colors on the cards, the letters printed on the cards and more. Choosing the right observations is critical for success.

In language analysis, the process begins with "tokenization." Each message must be broken into smaller "tokens." Depending on the vendor, a token could be a word, punctuation mark, proper name, a string of text that mix words and numbers or words and punctuation marks, email addresses, or the time the message was sent. The same word could be part of different tokens depending on usage. For example, the InBoxer spam filter considers a word in the subject to be a different token than the same word in the message. Building ideal tokens requires significant study and experimentation.

The final stage is to sort the incoming mail using the definitions and observations above. Many Bayesian filters sort mail into two categories, "spam" and "not spam". Messages in the "not spam" category are usually left alone and "spam" messages are quarantined. Sorting into two categories is done because it is easy to explain and similar to other types of spam filters.

However, the real power of a Bayesian filter comes when there are three categories: "spam", "not spam" and "likely spam". Why? Bayesian filters have the power to be finely tuned in ways that other types of filters cannot. This solves the problem mentioned at the beginning of the article - if you need to review the entire "spam" folder to make sure that you do not miss a message, and then the filter does not save time. If you only need to review the "likely spam" category to check for mistakes, good messages are never deleted or lost in a huge list of spam. In this way, the anti-spam filter can save time. (InBoxer calls the likely spam category: InBoxer - Review.)

Copyright 2004 by Audiotrieve, LLC. All Rights Reserved.  
Not responsible for errors. Subject to change without notice.  
INBOXER and AUDIOTRIEVE are registered trademarks of Audiotrieve, LLC.  
All other marks are property of their respective owners.  
Contact Audiotrieve, LLC at [info@Audiotrieve.com](mailto:info@Audiotrieve.com) or [info@inboxer.com](mailto:info@inboxer.com).