



Avoiding False Positives

PROCESS[™]
SOFTWARE

Executive Summary

Most sites install an anti-spam filtering solution for one reason: to get rid of all that spam! Soon after the filtering solution goes live, sites discover they've introduced a new problem in the form of false positives. A false positive is a legitimate message that has been incorrectly identified as spam by an anti-spam filtering solution. There are several reasons why an anti-spam filter might think a legitimate message is spam, the most common being that the message has several spam-like qualities. Newsletters, promotions from e-commerce sites, and joke lists all contain content that is virtually identical to spam.

Organizations consider legitimate messages incorrectly identified as spam a much larger problem than the occasional spam message that manages to sneak by the filter. The rare spam message that makes it to an end user's inbox can be quickly deleted, while false positives can cost from \$25 to \$110 per user each year¹. That cost only reflects the cost of the time lost while the user or system administrator tries to retrieve the message - the cost of lost business can be much higher.

The quality of an anti-spam filter is measured by a combination of its spam catch rate and false positive rate. Generally, a 90% spam catch rate (90 out of 100 spam messages are correctly identified as spam) and a false positive rate of less than 1% (less than 1 legitimate message out of a hundred incorrectly identified as spam) is considered good. Less developed filters that have a high spam catch rate tend to also have a high false positive rate, and vice versa.

In a recent NetworkWorld test, Process Software's PreciseMail Anti-Spam Gateway ranked in the top ten products (out of 41) when spam catch rates and false positive rates were combined. PreciseMail Anti-Spam Gateway tied for 2nd place in spam catch rate and the reviewer noted that PreciseMail "offers dozens of adjustments that can be used to drop the false positive rate while keeping the spam catch rate at 98% to 99%."

This document describes several techniques that modern anti-spam solutions support to minimize or even eliminate false positives, and includes examples of how those techniques are implemented in PreciseMail Anti-Spam Gateway.

Allowlisting

The first line of defense against a message being improperly marked as spam is an allowlist (alternately known as a whitelist). An allowlist lets messages that meet a certain criteria, such as being sent from a particular address or containing a certain word in the Subject line, pass through the anti-spam filter without being identified as spam. These allowlisted messages are always delivered to the recipient, regardless of their content. In effect, the allowlist provides a way to short-circuit the anti-spam solution to handle special cases.

¹ "Cost of Spam False Positives", Ferris Research

At almost every site, the most common use of allowlists is for automated mailing lists. Most providers of free mailing list services, such as Yahoo, recoup their costs by placing advertisements for third-party products in every email sent through the mailing list service. These advertisements are indistinguishable from spam - in fact, by some widely accepted definitions, their inclusion in the message turns the message into spam. Since the recipients' desire to receive the message outweighs their desire to avoid unsolicited marketing messages, they can use an allowlist to instruct the anti-spam solution to let the message through.

In addition to mailing lists, most email users set up allowlist entries for certain addresses that they know they will frequently send them content that might be identified as objectionable by an anti-spam solution. For example, end users will allowlist their personal real estate agent and mortgage broker so their messages can get through, while the never-ending stream of mortgage refinancing spam will still be caught by the anti-spam filter.

Any anti-spam solution designed for use in the enterprise supports both system-wide allowlists and personal allowlists. System-wide allowlists let a site's email administrator define criteria for messages that should always be let through the anti-spam filter, regardless of who the end recipient at the site is. Personal allowlists are set up by each individual user at the site, and the entries apply only to that user.

On occasion, it may be desirable to allow every message from a particular domain to bypass the anti-spam filter. Most anti-spam solutions allow wildcards inside allowlist entries to enable this functionality. For example, if a user wanted to allow every message from the domain example.com, they could add an allowlist entry that looked something like `*@example.com`.

PreciseMail Anti-Spam Gateway provides several different ways for end users to set up and manage their personal allowlists. Most users prefer to enter the addresses that allowlisted messages will be sent from in the web-based user interface. The screenshot in Figure 1 shows this interface in use.

In addition, the PreciseMail Anti-Spam Gateway web interface gives users the option to add the sending address of messages released from the message quarantine area to their personal allowlist. This way, future messages from that particular sender will never be quarantined. The screenshot in Figure 2 is a sample page that appears after an end user has released one or more messages from the quarantine area.

Alternatively, users who don't wish to or aren't able to use the PreciseMail Anti-Spam Gateway web-based user interface can use PreciseMail's email-based user interface. To add one or more addresses to their allowlist, users send a message to their site's PreciseMail processor address (usually `precisemail@domain.com`). The message contains entries like:

```
ALLOW frank@thechickenfarm.com
ALLOW agent@cheaphouses4u.com
ALLOW *@example.com
```

Users can also send commands to get a listing of all entries currently in their allowlist or to remove entries from their allowlist.



Figure 1: PreciseMail Allowlist Interface



Figure 2: PreciseMail Allow after Quarantine Release Option Page

Bayesian Filtering

Several anti-spam filters, including PreciseMail Anti-Spam Gateway, incorporate an artificial intelligence engine based on Bayesian text filtering. A Bayesian engine “learns” the difference between spam and legitimate messages sent to your site. A properly trained Bayesian engine can help prevent the anti-spam solution from misclassifying legitimate messages that contain some objectionable content as spam.

For example, suppose a user receives an overly chatty email from a friend who mentions that he’s started taking Viagra. Even worse, the friend accidentally adds an extra space so the word is spelled “Via gra”. Spammers often try to hide “bad” words like Viagra from anti-spam filters by breaking

them up with whitespace characters. Most spam filters would immediately classify the friend's message as spam because of the misspelled Viagra.

But the Bayesian engine has "seen" all of the previous messages sent to the user by his friend, so it notes other words in the message that indicate the message probably isn't spam. In the same message the friend might also include an invitation to a backyard barbecue, and discuss a new quarterback on their favorite football team. The Bayesian engine "knows" that a message containing the words "backyard", "barbecue", "quarterback", and "football" most likely isn't spam, so it tells the anti-spam solution to discount the misspelled Viagra. The specifics of this example would be different for every site, but the Bayesian filter learns the specifics of each site's mail content.

Modifying Filtering Thresholds

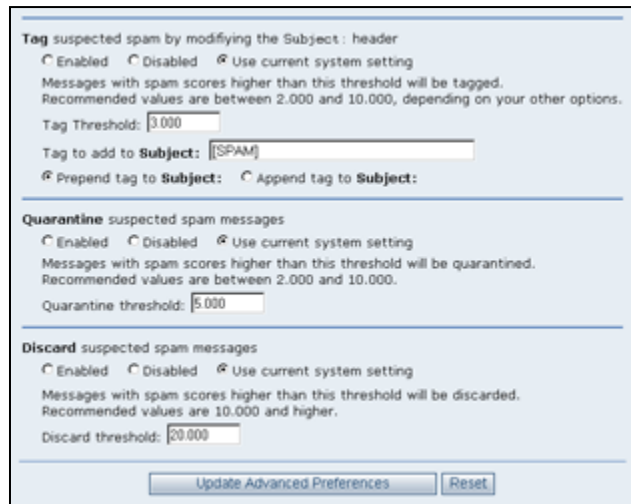
Many anti-spam filtering solutions, including PreciseMail Anti-Spam Gateway, assign every message they filter a numerical score. The higher the score, the more likely it is that a message is spam. Both the system administrator and end users can decide how high a message has to score before a certain action is taken. Example actions include tagging the message as spam in the Subject line, quarantining the message, or discarding the message.

The defaults for the numerical thresholds that a message's score has to cross before it is treated as spam are general settings that may not be appropriate for all sites. For sites that mainly receive messages that never contain spam-like features, such as manufacturing companies, the default thresholds may be too lenient. For other sites that receive a wide variety of message content, such as ISPs and universities, the default thresholds may be too aggressive.

Anti-spam solutions should let the system administrator change the numerical thresholds at which certain actions occur to site-appropriate values. In addition, the system administrator should be able to give end users the option to set their own personal account thresholds. For example, a university might leave the default threshold settings in place, but allow end users to modify the threshold settings for their personal account. A psychology professor whose field of interest is deviant behavior would set her threshold settings very high to allow objectionable content to reach his account, while an executive assistant to the university's president would set his threshold settings low to allow only work-related content in his inbox.

If the anti-spam solution allows the end user to choose a combination of actions to perform on messages identified as spam, the customization possibilities are endless. For example, the psychology professor in the above example may hate spam as much as the executive assistant, but she knows that she has to be able to receive email that contains very objectionable content at the same time. The professor might choose to set the numerical threshold that a message has to cross to be quarantined very high, but leave the threshold a message has to cross to be tagged as spam set to a middle-of-the road value. With those settings messages that contain all but the most objectionable content will be delivered to her account, but she can easily separate messages that are possibly spam from messages that are definitely not spam.

PreciseMail Anti-Spam Gateway lets system administrators change the default threshold settings that apply to all users, and lets them give end users the option to change the thresholds for their personal accounts. The screenshot below shows an example of the preferences web page that allows end users to customize their personal threshold settings.



The screenshot displays a web interface for configuring spam filter preferences. It is divided into three sections: Tag, Quarantine, and Discard. Each section has radio buttons for 'Enabled', 'Disabled', and 'Use current system setting'. Below each section is a text input field for a threshold value and a 'Reset' button. The 'Tag' section also includes a text input for a tag to add to the subject line and radio buttons for 'Prepend tag to Subject' and 'Append tag to Subject'. The 'Update Advanced Preferences' button is located at the bottom of the form.

Action	Enabled	Disabled	Use current system setting	Threshold	Tag to add to Subject	Prepend tag to Subject	Append tag to Subject
Tag suspected spam by modifying the Subject : header	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	3,000	[SPAM]	<input checked="" type="radio"/>	<input type="radio"/>
Quarantine suspected spam messages	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	5,000			
Discard suspected spam messages	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	20,000			

Tuning Rules

Heuristic filtering engines use a large set of rules to look for features of an email message. Some features indicate that a message may be spam, while other features indicate that a message probably is not spam. By carefully weighing the combination of features found in a message, the heuristic engine determines if the message is spam or not.

The rules used by the heuristic engine are designed to effectively filter spam for virtually every site. Even so, some sites may find that a small number of rules incorrectly identify a message as spam. A common example is a pharmacy in a large hospital that frequently emails lists of drug inventories. Several drugs on the list are commonly advertised in spam messages, so the heuristic engine incorrectly identifies the message as spam.

A flexible anti-spam filter allows the system administrator to de-activate certain rules, or reduce their importance in determining if a message is spam or not. In the example of the hospital pharmacy, the hospital's system administrator simply told the anti-spam filter not to use the appearance of those drug names as evidence that a message might be spam.

Likewise, a flexible heuristic engine allows system administrators to write their own site-specific rules that look for features that they know will most likely appear only in spam or non-spam messages. For example, a hand tool manufacturer might add a rule that identifies messages containing words like "hammer", "screwdriver", and "wrench" as not likely to be spam. System administrators with advanced knowledge of the pattern matching language that the rules are written

in (usually Sieve² or regular expressions³) should also be able to modify rules to fit the needs of their site.

PreciseMail Anti-Spam Gateway allows system administrators to add, modify, delete, and change the weights of rules inside its heuristic filtering engine. These changes are preserved across rule updates and software upgrades, so the system administrator only has to customize a rule once.

Quarantine Release

The last line of defense against the loss of legitimate email messages is the quarantine area of an anti-spam solution. Messages that the filter believes are most likely spam are removed from a site's mail stream and quarantined. Usually, this means that the messages are written out to a special area of a disk on the system that the anti-spam solution is running on. Messages that have been stored in the quarantine area longer than a fixed period of time are automatically deleted by the anti-spam solution.

Each user can get a listing of messages in their quarantined area, and choose to release messages that should not have been quarantined. Messages that are released are delivered normally to the user. The screenshot below shows a listing of quarantined messages in PreciseMail Anti-Spam Gateway's web-based user interface.



Ideally, users should check their quarantine every day or so to make sure legitimate messages have not been accidentally classified as spam. Some anti-spam solutions, including PreciseMail Anti-Spam

² Sieve is defined by *RFC 3028, Sieve: A Mail Filtering Language*

³ Regular expressions are a specialized regular language designed for text pattern matching. For more information, visit <http://etext.lib.virginia.edu/helpsheets/regex.html>

Gateway, send automated mailings to each user who has new quarantined messages to remind them to check their personal quarantine area. These mailings are only sent a few times a day to prevent them from becoming an annoying distraction. Users can preview the contents of quarantined messages and release accidentally quarantined messages from inside the notification email.

Conclusion

While most sites' only concern when installing anti-spam solutions is to reduce the incoming flow of spam, legitimate messages incorrectly identified as spam can quickly become a major problem. This whitepaper has provided explanations of several methods that can be used to minimize the impact of these false positives, and shown examples of how they are implemented in PreciseMail Anti-Spam Gateway.

Keep Legitimate Messages from Being Filtered – Advice For End Users

To help prevent your email messages from accidentally being classified as spam by PreciseMail Anti-Spam Gateway, try performing the following actions:

1. Add allowlist entries for all of your mailing lists and email newsgroups, especially those that are provided by a third party that inserts advertisements in each message.
2. Add allowlist entries for e-commerce sites you wish to receive mailings from (such as Amazon, eBay, and travel sites).
3. Add allowlist entries for friends and business associates.
4. When you first begin using PreciseMail Anti-Spam Gateway, change your settings to tag messages with a score greater than 5 and quarantine messages with a score greater than 15. Gradually lower those threshold settings until most spam is quarantined and messages that are likely spam are tagged.
5. When you release a message that has accidentally been quarantined, add the sender's address to your personal allowlist so future messages from that sender aren't quarantined.
6. Opt to receive the daily summaries of your quarantined messages so you don't forget to check for legitimate messages in your quarantined message area.
7. If legitimate mail from a certain sender or host is consistently quarantined, report it to your mail administrator so the system-wide rules can be tuned.

About PreciseMail Anti-Spam Gateway

PreciseMail Anti-Spam Gateway is an enterprise software solution that eliminates spam, phishing and virus threats at the Internet gateway or mail server. It has a proven 98% spam detection accuracy rate out-of-the-box without filtering legitimate messages. PreciseMail Anti-Spam Gateway has a highly sophisticated filtering engine is based on a combination of proven heuristic, DNS blacklisting, and Bayesian artificial intelligence technologies, which automatically learn how to separate spam messages from legitimate email. As a result, PreciseMail Anti-Spam Gateway can determine whether email is spam instead of passively reacting to known spammers by creating rules that block them after a spam attack occurs.

About Process Software

Process Software has been a premier supplier of communications software solutions to mission critical environments for twenty years. We were early innovators of email software and anti-spam technology. Process Software has a proven track record of success with thousands of customers, including many Global 2000 and Fortune 1000 companies.



U.S.A.: (800) 722-7770 • International: (508) 879-6994 • Fax: (508) 879-0042
E-mail: info@process.com • Web: <http://www.process.com/>