# SpamSieve 1.1 Manual

Michael Tsai

September 19, 2002

# 1 Introduction

## 1.1 What Is SpamSieve?

SpamSieve is an application that integrates with e-mail clients to filter out unsolicited mass mailings, commonly known as "spam." Previously, most people just ignored spam messages or created simple rules in their e-mail clients to filter it out. In recent years and months, the spam problem has gotten worse. Today's spam is harder to detect, and there is more of it. People have turned to a wide variety of software solutions to help regain control over their inboxes. This software differs primarily along two dimensions: how it identifies spam messages; and how it reduces their burden on the user.

## 1.2 Identifying Spam

*Simple rules* that you create in your e-mail client generally look for common spam words. If a message's subject contains "fix your credit" or its body contains "free porn" it's probably spam. You can look at the spam you receive and recognize some patterns. Unfortunately, spam is constantly evolving, so you will have to keep working on your rules. Eventually you get to the point where you don't know what else to add, afraid that adding more words or phrases to your "blacklist" would start filtering out legitimate mail. You can somewhat get around this by creating a "whitelist" that accepts every message from the people you regularly correspond with. However, this doesn't help for messages received from new people,

and it's not uncommon for spam messages to contain a forged return address of someone at your own company.

Commercial anti-spam software often combines rules with the notion of a *score*. Rules can look for patterns that make a message more or less likely to be spam. Each rule either increases or decreases the message's spam score. If the score is above a certain threshold, the message is considered to be spam. The flexible nature of the score means that this approach is often an improvement over simple rules. However, the user typically has little or no control over how the spam score is calculated. If a spam message gets through, the user has no recourse but to hope that the next update corrects the problem. Another problem is that spammers can buy this software, too. They can tailor their spam to get through the rules.

SpamSieve uses a statistical technique known as *Bayesian analysis*[1]. It combines the good properties of the above two approaches and adds some of its own. First, you *train* SpamSieve with examples of your good mail and your spam. When you receive a new message, SpamSieve looks at how often its words occur in spam messages vs. good messages. Lots of spammy words means that the message is probably spam. However, the presence of words that are common in your normal e-mail but rare in spam messages can tip the scale the other way. This "fuzzy" approach allows SpamSieve to catch nearly every spam message yet produce very few false positives[2].

Because you train SpamSieve with your own mail, you have full control. If SpamSieve makes a mistake, you can train it with the message in question so that in the future it will do better. Further, since spammers don't have access to the messages you trained SpamSieve with, they have no way of knowing how to change their messages to get through. Whereas other spam filters become less effective as spammers figure out their rules, *SpamSieve becomes more effective over time* because it has a larger corpus of your messages to work from.

## 1.3 Filtering It Out

Anti-spam software that runs on mail servers filters out spam before you ever see it. This means that unless the filter is perfect, either some spam messages will get through or a few legitimate messages will not. In the first case, you may want additional, client-side, anti-spam software. The second case troubles many people so much that they prefer that the server do no filtering at all.

Other client-side anti-spam software connects to your mail server to delete spam messages before your e-mail client can download them. This works similarly to above, except that to catch all the spam messages you have to run the program right before your regular e-mail program checks for mail. This is difficult to time properly if you check your mail often, and even so you may download some messages that weren't filtered. The anti-spam software may let you see the messages that it filtered out, so that you can verify that there were no false positives. However, you have to do this using its interface, not your e-mail program's (which

[1] For a more in-depth treatment of Bayesian analysis applied spam, see the article by Paul Graham at http://www.paulgraham.com/spam.html and the papers it references.

[2] A *false positive* is a good message mistakenly identified as spam. Most users consider false positives to be much worse than *false negatives* (spam messages that the user has to see).

is typically nicer). And if there was a false positive you then have to transfer it into your e-mail program so that you can file and reply to it.

E-mail clients like Entourage have their own spam filters built-in. This is convenient and makes it easy to manually scan for false positives, but there is typically little you can do when the filter makes mistakes.

Clearly, the best solution is user-configurable anti-spam software that works directly with your e-mail client. Apple has added exactly this to its Mail program in Mac OS X 10.2. By most accounts this works great. . .unless you prefer a client other than Mail.app. This is where SpamSieve comes in. It brings powerful spam filtering to other popular e-mail clients such as Mailsmith and Entourage.

# 2 Requirements and Installation

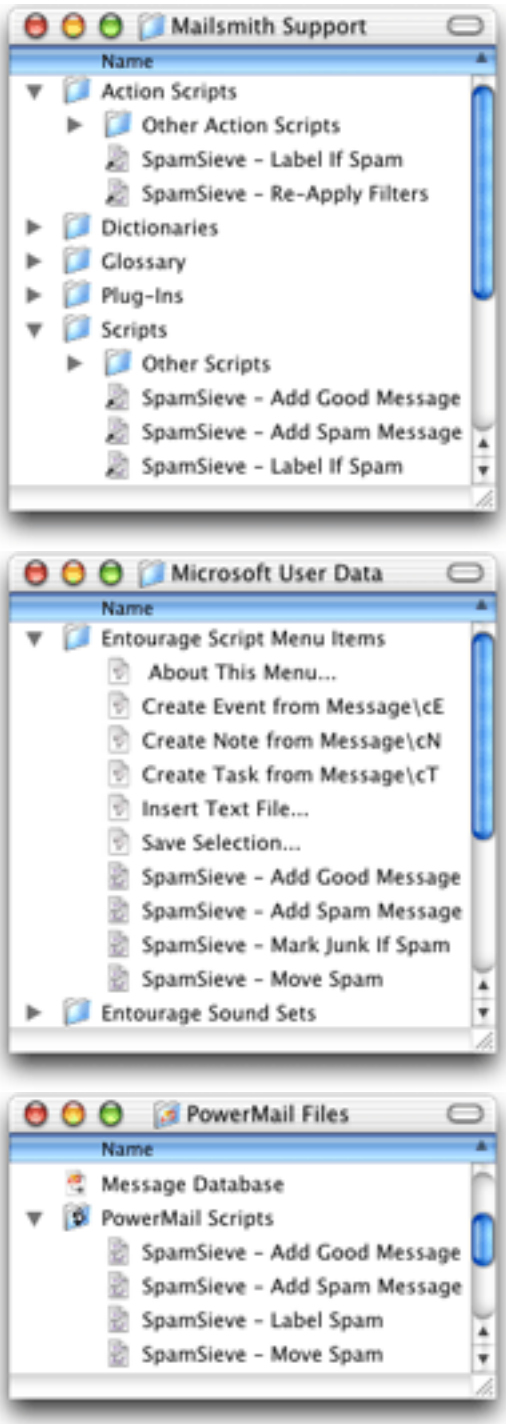
SpamSieve has been developed and tested on Mac OS X 10.1.5 and 10.2. I do not have the resources to test it on older systems, although I suspect it will work fine on Mac OS X 10.1 or later.

SpamSieve is designed to work with the following e-mail clients:

- Mailsmith[3] from Bare Bones Software
- Entourage[4] from Microsoft
- PowerMail[5] from CTM Development

To install, copy the SpamSieve application to your hard disk, e.g. into `/Applications`. Then find the folder on the disk image that pertains to your e-mail client (e.g. `For Entourage Users`). The names of the other folders inside it tell you where to copy their contents. The AppleScripts in these folders allow you to interact with SpamSieve from within your e-mail client.

[3] http://www.barebones.com/products/mailsmith.html
[4] http://www.microsoft.com/mac/entouragex/default.asp?navindex=s4
[5] http://www.ctmdev.com/powermail4.shtml




There's no need to copy this manual to your hard disk. A copy of it is built into SpamSieve, and you can access it by choosing **SpamSieve Help** from the **Help** menu.

# 3 Training SpamSieve to Recognize Your Spam

Before you can use SpamSieve, you must give it some examples of messages you consider to be spam, and ones which you do not. Make sure that you have installed the appropriate integration AppleScripts for your e-mail client (see Section 2.

Select some spam messages and then use your e-mail client's Scripts menu to run the `SpamSieve - Add Spam Message` script. Then select some good messages and run the `SpamSieve - Add Good Message` script. You can run these scripts at any time, e.g. whenever SpamSieve makes a mistake. The more messages you train SpamSieve with, the better its accuracy it will be. For best results, you should train it with *at least* 600 messages. It is important to train SpamSieve with both spam messsages and good messages. If you can, train it with the same ratio of these two types of messages as you normally receive.

If SpamSieve marks a good message as spam, you should run the `SpamSieve - Add Good Message` script on that message. This lets SpamSieve know that it made a mistake, and also adds the message to the corpus to improve future accuracy. Likewise, if SpamSieve marks a spam message as good, you should run the `SpamSieve - Add Spam Message` script on that message.

# 4 Filtering Messages With SpamSieve

## 4.1 How SpamSieve Works With E-Mail Clients

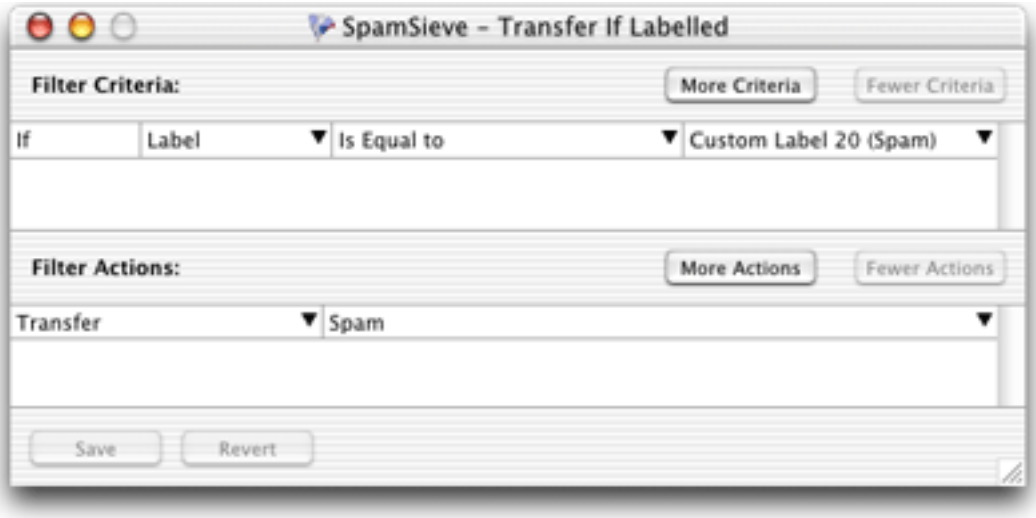
Once you've trained SpamSieve, you can begin using it to identify spam messages. The general procedure is that you configure your e-mail client to run an AppleScript each time you receive a message. The AppleScript asks SpamSieve if the message is spam. If it is, it marks the message. You can then set up your own rules in your e-mail client to process the marked messages. You might move them directly to the trash, or into a "Possible Spam" mailbox if you want to manually check for false positives.

All the e-mail clients that SpamSieve supports let you control the order in which their rules process mail. How you order the SpamSieve rule is up to you. If you get a lot of spam that happens to match the rules you use to organize your mail, you might want to run the SpamSieve rule first. If you'd rather deal with spam manually than have any false positives, then you might want to run the SpamSieve rule last, after all your other rules have been given a chance to match.

## 4.2 Setting up a Mailsmith Filter

Create a filter that looks like this:


The two criteria ensure that the filter's action will be applied to every message. The action of the filter sets the label of spam messages to Custom Label 20. You can then use a filter such as the following to transfer spam messages to a particular mailbox:




On some Macs, Mailsmith is slow at applying filters while downloading mail, if the filters have AppleScript actions. If you observe that Mailsmith is taking a long time to download and filter your mail with SpamSieve installed, there are a few steps you can take to improve performance. First, go to Mailsmith's **Filtering** preferences and uncheck **Apply Filters while Downloading**. Next, go to Mailsmith's **Notification** preferences. Check the **Run Script** checkbox and from the popup menu select the `SpamSieve - Re-Apply Filters` script. The effect of these two steps is that Mailsmith will download your mail quickly, and then filter it. This should be much faster than filtering the messages as they are downloaded.

## 4.3 Setting up an Entourage Rule

Create a rule that looks like this:

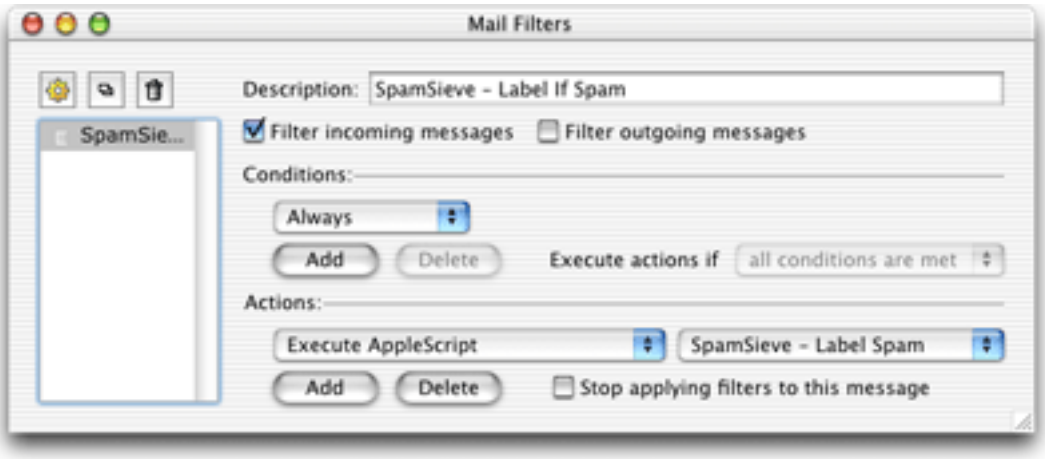

The rule applies to all messages and sets the category of spam messages to Junk.

Alternatively, you can set the rule to use the `SpamSieve - Move Spam` script. This script will move move spam messages to an Entourage folder named **Junk** (if you have one).

One complication is that after Entourage runs an AppleScript it cannot apply any more rules to the message. Therefore, you will probably want Entourage to execute the SpamSieve rule after all your other rules. If you want to further process messages that SpamSieve has identified as spam, you will need to use AppleScript in the manner of the `SpamSieve - Move Spam` script.

## 4.4 Setting up a PowerMail Filter

Create a filter that looks like this:




This filter applies to all messages and sets the label of spam messages to 7.

Alternatively, you can set the filter to use the `SpamSieve - Move Spam` script. This script will move spam messages to a PowerMail folder named **Spam** (if you have one).

# 5 The Corpus Menu

## 5.1 Show Corpus

This command opens the Corpus window so that you can examine the words that SpamSieve has found in your e-mails. You can click on the name of a column to sort by that column. The meanings of the columns are as follows:

**Word** A word in the corpus. Note that the Corpus window does not show words that are in the corpus but have not occurred a significant number of times.

**Spam** The number of times the word has occurred in spam messages.

**Good** The number of times the word has occurred in good messages.

**Total** The total number of times the word has occurred.

**Prob.** The probability that a message is spam, given that it contains the word (and in the absence of other evidence).

## 5.2 Prune Corpus

This command removes from the corpus words that are not contributing to spam recognition. This can dramatically decrease SpamSieve's memory use and launch time. You can prune repeatedly to reduce the corpus size further. However, excessive pruning will increase the training needed to recognize new types of messages.

## 5.3 Show Statistics

This command opens the Statistics window, which displays the following items:

**Good Messages** The number of non-spam messages that SpamSieve has filtered.

**Spam Messages** The number of spam messages that SpamSieve has filtered.

**False Positives** The number of good messages that SpamSieve identified as spam.

**False Negatives** The number of spam messages that SpamSieve identified as good.

**% Correct** The percent of messages that SpamSieve identified correctly.

**Good Words** The number of unique words appearing in good messages.

**Spam Words** The number of unique words appearing in spam messages.

**Unused Words** The number of words that are not yet being used to identify spam messages, because they have not occurred enough times in your messages.

**Total Words** The total number of unique words in the corpus.

# 6 Contact Information

The SpamSieve Web site is located at http://www.c-command.com/spamsieve/. Questions about SpamSieve may be sent to mailto:support@c-command.com. I'm always looking to improve SpamSieve, so please feel free to send any feature requests to that address.

To make sure that you have the latest version of SpamSieve, you may wish to subscribe to the SpamSieve Announcements mailing list. The traffic on this list is very low, only one message per new version of SpamSieve. You may sign up using the form at http://www.c-command.com/spamsieve/support.shtml.

# 7 Registering

SpamSieve is shareware. If you find yourself using SpamSieve beyond a reasonable trial period, you must register it. Registration costs $10 (US) and entitles you to free updates and support.

To register, go to http://store.eSellerate.net/s.asp?s=STR804431608. Soon after paying, you'll receive an e-mail with your serial number. Enter it in the Registration window to personalize your copy of SpamSieve.

This is the honor system. If you use SpamSieve without registering, I probably won't know. However, registering will give me an incentive to continue updating and enhancing SpamSieve, and to write more Mac software. And you won't have to look at the "Unregistered" window anymore.

# 8 Version History

## 1.1—September 19, 2002

- E-Mail Client Integration
  - Added support for PowerMail.
  - Added instructions and an AppleScript for making Mailsmith download and filter mail faster.
  - Added an AppleScript for Entourage that moves spam into a Junk folder.
- Performance
  - Launches about 60% faster than 1.0.
  - You can now prune the corpus to remove words that are taking up memory without contributing to spam recognition. This can also dramatically decrease SpamSieve's launch time.
  - Recalculating spam probabilities is about 10% faster and uses less memory.
  - Quitting is faster because SpamSieve now writes corpus changes to disk during idle time.

  - Saving the corpus is slightly faster.

- Displays statistics about the number of messages filtered, SpamSieve's accuracy, and the types of words in the corpus.
- SpamAssassin's X-Spam-Status headers are now treated as single words. This means that if SpamAssassin is running on your mail server, SpamSieve will learn to respect (or ignore) its judgement.
- Does a better job of ignoring e-mail attachments, thus reducing corpus bloat.
- Installs the eSellerate Engine if it's not present, thus enabling "Instant Registration" for more users.
- Asking SpamSieve to categorize a message now forces an update of all the word probabilities. Previously, the update only happened during idle time.
- Highlights the sorted column in the Corpus window. The columns themselves have shorter names. There's a new "Total" column. Auto-resizing of the columns works better. You can now manually resize any column, and manual resizings and reorderings are saved between launches.
- Shows fatal errors as alert panels rather than just printing them on the console.
- The Corpus.plist data file is now sorted by word. This makes it easier to examine the corpus manually, and to compare it to other users' corpuses.

### 1.0—September 10, 2002

- First public release.

# 9 Legal Stuff