ImproveAccuracy

How to Improve SpamAssassin Accuracy

Run a recent version

Regular updates of SpamAssassin 3.2.x rules stopped in 2008. Accuracy depends on more recent rules. Upgrade to 3.3.0 or newer.

Run sa-update daily

This is often included in SpamAssassin packaging, but sa-update should be run from cron daily, to get the latest SpamAssassin rules which are generated every day.

(On Debian based systems, set "CRON=1" in /etc/default/spamassassin - this is not the default.)

Enable network rules

This is the default, but disabling network rules (including DNS rules) causes SpamAssassin to be wrong on about 3 times more emails. Network tests may have been disabled by running spamassassin or spamd with the command line arguments `-L` or `--local`. DNS rules may have been disabled with "`dns_available no`" in local.cf. You should run a local caching DNS server for efficiency.

As of 2011-12-21, without network tests, SpamAssassin is wrong 2.58 times as often on non-spam, and 3.40 times as often on spam.

Install Pyzor and Razor

These are two helper applications with useful (network) rules. If they're installed correctly, the debug output of SpamAssassin will include:

```
Apr 14 16:24:37.315 [4709] dbg: plugin: loading Mail::SpamAssassin::Plugin::Pyzor from @INC
Apr 14 16:24:37.318 [4709] dbg: pyzor: network tests on, attempting Pyzor
Apr 14 16:24:37.318 [4709] dbg: plugin: loading Mail::SpamAssassin::Plugin::Razor2 from @INC
Apr 14 16:24:37.381 [4709] dbg: razor2: razor2 is available, version 2.84
```

Trusted Networks settings

Ensure that internal_networks and trusted_networks are set correctly. Often, spamassassin will intelligently do the correct thing by default. But if you're receiving a significant portion of your email via a trusted relay, it needs to be listed in one of these manually, otherwise the wrong hop will be used for things like DNS blacklist tests. More info at TrustPath.

Verify AWL and the Bayesian classifier aren't poisoned

The AutoWhitelist, and Bayesian classifier when automatically trained, can get trained incorrectly, resulting in scoring email wrong. Verify they are providing useful scores - positive scores for spam, and negative scores for ham (AWL and BAYES_* tests). If they are causing counterproductive scores, the only solutions are to delete the relevant databases and start over training them, or disable them with:

```
use_auto_whitelist 0
use_bayes 0
```

To only disable automatic training of the Bayesian classifier:

```
bayes_auto_learn 0
```

To remove all existing Bayesian tokens to start training over:

```
rm ~/.spamassassin/bayes_*
```

Remove any SARE rules

SARE rules have not been updated in years, and are therefore actively harmful. They are not included in SpamAssassin by default, but often have been added to local configurations.

Remove Sought rules

SoughtRules ***was*** a custom rule set generated from spam 4 times a day by a SpamAssassin developer. This rule set is no longer maintained and due to its design may have randomly harmful effects on scoring.

Only accept email in specified languages - TextCat

In /etc/spamassassin/local.**pre** add:

```
loadplugin Mail::SpamAssassin::Plugin::TextCat
```

In /etc/spamassassin/local.cf add:

```
ok_languages en es
```

Where "en es" is a list of codes for languages you wish to accept. The full list is in the TextCat documentation.

It is very important that the loadplugin line be added to a .pre file not a .cf file so it is loaded *before* the rules files are loaded, otherwise those rules will not get enabled.

You may also want to increase the score from the default of 2.8:

```
score UNWANTED_LANGUAGE_BODY 5
```

Use a local, caching, non-forwarding DNS sever

CachingNameserver.

Some DNS Blacklists and Whitelists will block queries from DNS servers issuing what they consider too many queries. This is often avoided by running a local DNS server. Also, it's good for performance.

Run SPF at your MTA

SPF is intended to operate on the envelope sender (SMTP protocol MAIL FROM) which is not available in a standard way by the time the email gets to SpamAssassin. The solution is to run SPF at your MTA (Message Transfer Agent, such as Postfix, Exim, Qmail, Sendmail, etc.). This is, of course, dependent on what software you're using, but it should insert a Received-SPF: header for use by SpamAssassin. If you do not run SPF at your MTA, you really should set `ignore_received_spf_header 1` so you don't end up honoring headers inserted by spammers.

An option for Postfix: https://launchpad.net/postfix-policyd-spf-perl/

Use sa-learn to manually train the Bayesian classifier

If it's worth the time to increase the accuracy of filtration of your own personal email, you can manually sort it into ham and spam folders, and then use sa-learn to train it. This can be used for a group effectively if the group is well trained (not to classify mailing lists they've subscribed to but lost interest in as spam).

Pick a useful threshold

The default threshold is 5, which is used to calculate the scores of all of the tests. Higher numbers will result in fewer emails considered spam - both reducing false positives, and increasing false-negatives. Reducing the threshold below 5 is not recommended. This is configured with:

```
required_score 5
```

Filtration at your MTA

While outside the scope of SpamAssassin, it is common to do some configuration at your MTA to reject invalid mail.

(For postfix, some settings you might want to look into: reject_non_fqdn_hostname reject_non_fqdn_sender reject_invalid_hostname reject_unknown_client reject_unknown_sender_domain reject_unauth_destination check_sender_access check_helo_access check_sender_mx_access)

Mass-Check

Participating in NightlyMassChecks means that the scores for many of the SpamAssassin tests take your own emails into account, which is likely to increase your accuracy. You don't even upload your emails, unless you want to, just the test hit rates.

Find other missing perl modules

You may be able to find other perl modules you can install to help SpamAssassin by running:

```
spamassassin -D --lint 2>&1 | grep -i failed
```

Writing Rules

WritingRules - when existing tests are not sufficient.

SpamTips.org setup guide

http://www.spamtips.org/p/ultimate-setup-guide.html

IRC / Mailing List

If you have done everything above, and are still not having satisfactory results, the next step is to provide some examples of spams being missed, with complete headers, to people who can provide suggestions, via something like pastebin.com - don't paste this stuff in the bodies of emails or include it as attachments. The IRC channel #spamassassin on irc.freenode.net may help, but it is unfortunately not very active. The relevant mailing list is the spamassassin users list.