VirusScannerTypeUpdates

UPDATE: this page is pretty much obsolete nowadays; we *are* doing updates. See RunningSaUpdate.

It's quite common for people new to SpamAssassin to try to view it as being like a virus scanner. Virus scanners typically release signature updates every few days, and it seems a logical extension to try to do this with SpamAssassin too.

Unfortunately that's not possible in practice because of how SpamAssassin rules work. This is because SpamAssassin works absolutely nothing like any virus scanner.

In a virus scanner each signature operates independent of the other signatures. It's simply a "this is a virus" assertion. This allows you to add new signatures and update old ones without having to change every signature in the database. This feature facilitates "quick turn" updates to the signature database. By comparison updates to the code take a very much longer time frame, and the code is also more-or-less independent of the signatures as well.

SpamAssassin on the other hand has a "score tally" system where large numbers of inter-related rules fire off and total up a score to determine if a message is spam or not.. In this system each rule affects the proper score of *every* other rule in the ruleset. The whole system operates by trying to balance the most spam and nonspam each on the right side of the 5.0 mark. Adding rules, changing their hits, etc tends to shift the "center line" and requires you to rebalance the entire system. To make matters more complex, many aspects of the rules are dependent on the code as well.

This "rebalancing" is done by the mass-check tests and scoring run of SpamAssassin. This process takes about 4 weeks to complete for a new ruleset. This process heavily bogs down the computers of all the corpus submitters for days on end, particularly for the runs with network checks. The last time this was done 1,690,967 messages were processed (between the 4 sets combined), the hits of each and every message (not just aggregate hit counts per rule) were fed into an error minimisation program, and the 4 scoresets were evolved.

And that's 4 weeks just for the mass-check and scoring process, and that's done after a set of rules are decided to be good based on some smaller-scale dry runs, developer debate, testing, tweaking, and hand analysis of various bits of spam and nonspam.

While this makes the whole mass-check and scoring process sound bad, those very aspects of SpamAssassin are what makes it so effective in the first place. They are what provides the "real world" feedback into the whole system. The reality of email is a complex knotted mess of human behaviours, something not easily characterized by simple "anything with this word/phrase must be spam". Using a system that analyzes real world email, and generates a set of scores that fit reality is very powerful. In some ways, it's a lot like crossing the best parts of human generated filter rules, with the statistical measurements of a bayes system (and making it even more deeply analyzed than simple probabilities can represent).

The reality is that a new version of SpamAssassin gets released every time a new ruleset can be churned out. And typically, because rulesets take so long, at least a couple of additional releases occur before a new ruleset is ready.

Daily and/or weekly updates aren't practical, because it takes weeks to evolve a scoreset for a release.