

Re: OT: writing resumes with VT100 for a Lisp job

Source: <http://coding.derkeiler.com/Archive/Lisp/comp.lang.lisp/2004-08/0710.html>

From: rem642b_at_Yahoo.Com (*RobertMaas_at_YahooGroups.Com*)

Date: 08/12/04

Date: Wed, 11 Aug 2004 17:25:53 -0700

I've been reading the online documents and composing this response for several hours, and it's currently up to 383 lines, 21k bytes, so I'm going to break it into appx. 10k segments and upload each piece successively unstead of trying to send it all at one time when I finally finish it...

- > *From:* Thomas Schilling <tjs_ng@yahoo.de>
- > *You know of Paul Grahams nice paper, don't you?*
- > <http://www.paulgraham.com/spam.html>

Unfortunatley that document doesn't say clearly whether the filtering occurs inside the SMTP server, whereby anything determined to be likely spam is rejected with a 5yz code, or whether the filtering occurs after the message has already been received and acknowledged by the SMTP server and it's too late to refuse it and you're stuck with it forever until you delete it. From my reading of the following text:

The more spam a user gets, the less likely he'll be to notice one innocent mail sitting in his spam folder. And strangely enough, the it appears his filter diverts already-received e-mail to a side folder. As I've pointed out many times, this has the disadvantage that the sender of legitimate e-mail never learns his/her e-mail was diverted and will never been seen by the recipient. If the e-mail was urgent, lost time waiting for the recipient to see the e-mail before giving up and contacting the recipient some other way, can cause major cost, even loss of life.

One question that arises in practice is what probability to assign to a word you've never seen, i.e. one that doesn't occur in the hash table of word probabilities. I've found, again by trial and error, that .4 is a good number to use. If you've never seen a word before, it is probably fairly innocent; spam words tend to be all too familiar.

There's a very simple way any spammer can defeat such a filter: Include thousands of pseudo-random "words". Each one biases the message by 0.4 toward being non-spam, but thousands of such tiny biases cause the spam to be clearly recognized as non-spam. (But see later below.)

To beat Bayesian filters, it would not be enough for spammers to make their emails unique or to stop using individual naughty words. They'd have to make their mails indistinguishable from your ordinary mail. Not correct. All they'd have to do is increase the fraction of random pseudo-words so those .4 factors combine to swamp out any actual Bayesian information. I think brand-new words never seen before should be scored 0.51, so they weigh very slightly toward an indication of spam, so that thousands of brand-new pseudo-words strongly indicate spam. (But see later below.)

Spam is mostly sales pitches, so unless your regular mail is all sales pitches, spams will inevitably have a different character. Except that if the sales pitch uses a different random mis-spelling of each key word each time, none of those words will be recognized as indications of spam, and per the .4 factor currently used will in fact slightly bias the filter against considering it non-spam. For example: to un4subscrib75894e cl43ick h23re you can 32en43larg43e your 5p32enis43 with 5vi34gra432 The lack of even one clearly spam-indicating word spelled correctly, hence not a single factor greater than .5, and all those randomly mutated words plus thousands of totally random "words" at the end of the message, each with that .4 factor, would falsely show the message to be solidly non-spam. (But see just immediately below.)

But I see the filter ignores all but the 15 or 20 most significant words, which I presume means it uses only the 15 or 20 that are most distance from exactly 0.5 probability. If so, then the randomized spam would have 15 or 20 factors of 0.4, and no other factors at all, so the spam would be mistakenly recognized as non-spam, but not with certitude. Still I see the filter putting all that randomized spam into your inbox, so the spammer wins, and the filter is nearly worthless.

I don't know enough about the infrastructure that spammers use to know how hard it would be to make the headers look innocent, but my guess is that it would be even harder than making the message look innocent. The key header line which can't be forged is the last-relay Received line. If spam is sent through trojaned personal computers, each such line will be different from any other a given victim has received (except in a very rare chance event where the same trojaned machine remains under control of spammers for a long time and the same part of the 30-million e-mail address happens to be farmed out to that same machine a second time). Except for those particular IP numbers and host names belonging to dynamic address pools, there's nothing suspicious about them. Given that worms have recently been calling Google and other search engines to collect more e-mail addresses to spam, there's nothing to stop a worm from calling Google Groups to find out what topics a given victim has posted about and making the Subject line be Re: whatever that victim posted about. (I've never seen any spam that does that yet, but give the spammers a couple more weeks. It was just a

couple weeks ago when they started the search-engine method to a large enough degree to be noticeable. Victim-customized Subject fields in spam is the logical next step. When I start getting spam to rem642b@Yahoo.Com with Subject:

Re: Salaries for Lisp engineers
in place of the current Subject fields unrelated to anything I posted:
About your home in Mt. View
Bubby Mees Do you still need help
Bubby We Need Your PERMISSION
Bubby, Work with eBay ... call 1-866-622-9987 x 8030
Bubby, You'll enjoy this!
Bubby, luck is in the air
Complimentary Instructions To Remove Spyware/Adware Infections
Current Critical Update
Don't Be The Last
Find other singles in your area
GREETINGS TO YOU AND YOUR FAMILY
HELLO AND GOOD DAY
Hey I just read your email..
Hi, you there? 3
Kindly assist
Mail Delivery (failure rem642b@yahoo.com)
Network Pack
Please contact us Bubby
Regular update and verification of the accounts (from: eBay)
The Challenges of Our Time (from: Vice President Cheney)
The Real "Heart and Soul" of America (from: President George W. Bush)
Undelivered Mail
Upgrade your career
Urgent Reply on (Business)
advice
eBay Workers Needed...Call 1-866-621-2387 x1661
i was thinking..
undeliverable mail
undeliverable message: user unknown
I'll know this has happened)

> (or other papers from <http://www.paulgraham.com/antispam.html>)

Let me take a look at them now:

--> <http://www.paulgraham.com/spamfaq.html>

Is there anything that can protect my company's server?

Most commercial server-level spam filters are still rule-based. But there are starting to be some that use Bayesian filtering. The way to find them is probably to search in Google.

The question to ask the salesman is, does the filter learn to recognize spam based on the spam and nonspam mail we receive? If it doesn't learn, it isn't Bayesian.

Bayesian filtering seems nice for single users, but for a company's server there's a problem: It filters based on a consensus of what the average user likes and dislikes, so that any recipient whose

preferences are different from the norm would not be filtered appropriately, special topics only that one person likes would be regarded as spam and he wouldn't be allowed to receive them, while topics most people like but he hates would be crammed down his throat.

Lisp's symbol type is useful in manipulating databases of words, because it lets you test for equality by just comparing pointers. That is false advertising. When raw text is taken in, each word must be "interned" i.e. convert to upper case then entered into a hash table, and *then* any later use of the same word is found by hashtable lookup, but EVERY instance of that word must be hashed again, not just the first instance, because it's only after hashing that the word can be recognized as a repeat of an already-hashed word. Since there's extra overhead involved in creating a symbol, and problems with interning in a specific package where the sudden presence of a particular symbol might have semantic consequences, it'd be better to just convert to upper case and enter into an EQUAL hash table and not make an actual symbol out of each different word.

--> <http://www.paulgraham.com/better.html>

There are two kinds of spams I currently do have trouble with. One is the type that pretends to be an email from a woman inviting you to go chat with her or see her profile on a dating site. If the filter can't block that, it's pretty much worthless.