Subject: Is the learner working?
Posted by InforMed Direct on Tue, 18 May 2004 13:25:43 GMT
View Forum Message <> Reply to Message

We're using No Spam Today! with Exchange 5.5 and SpamAssassin and our hit rate with spam is only about 50%, sometimes less. But this is better than nothing!

I think we're using the learner correctly via this process:

1. We asked users to move spam emails that sneak through into a shared mailbox
2. Periodically, we collect these messages into Outlook Express
3. When then run DBXtract (freeware) to export the inbox into EML files
4. Execute the following command on the No Spam Today server:

sa-learn --spam s:\Temp\Spam*.eml -C ruleset

We've put thousands of spam emails through there. Couple of questions:

1. SpamAssassin says it needs ham as well as spam. Am I correct in assuming that SpamAssassin automatically passes emails that it processes through the learner thus passing ham (non-spam) and known spam?

2. Assuming this is correct, consider the case where a spam email passes through undetected. This will be processed as above when the user reads it, i.e. it'll then be passed through the learner as spam. Is this okay?

3. Is there anyway to tell if the learner is actually learning anything?

Thanks, Rob.

Subject: Re: Is the learner working?
Posted by support on Tue, 18 May 2004 14:19:44 GMT
View Forum Message <> Reply to Message

> I think we're using the learner correctly via this process:

[...]

> 4. Execute the following command on the No Spam Today
> server:
>
> sa-learn --spam s:\Temp\Spam*.eml -C ruleset

I see several problems here:
- in the default setup of NoSpamToday!, the Bayes DB path is relative to the working directory.

The mails you learn may end up in a different DB than the DB used by SpamAssassin.
- if I'm not mistaken, it is not possible to invoke sa-learn for multiple mail files like this. You could write the file names to a file, and use the -f option.
- the docs (sa\doc\sa-learn.htm) indicate that the sa-learn options should be before the mail file name.

Try this:

cd NoSpamTodayInstallationDirectory
sa\sa-learn -C sa\ruleset --spam s:\Temp\Spam\somemail.eml

Look into this howto for more info:

http://www.byteplant.com/support/nospamtoday/howtolearnspam.html

> 1. SpamAssassin says it needs ham as well as spam. Am I
> correct in assuming that SpamAssassin automatically passes
> emails that it processes through the learner thus passing ham
> (non-spam) and known spam?

Yes, a low spam probability decreases the overall spam score.

> 2. Assuming this is correct, consider the case where a spam
> email passes through undetected. This will be processed as
> above when the user reads it, i.e. it'll then be passed through
> the learner as spam. Is this okay?

Yes.

> 3. Is there anyway to tell if the learner is actually
> learning anything?

You can dump the database (--dump) , and look at the tokens (words) learned and the number of tokens.

Subject: Re: Is the learner working?
Posted by bjsvec on Tue, 18 May 2004 17:34:57 GMT
View Forum Message <> Reply to Message

I have found that Mozilla can export Outlook email to mbox format in a single file.
I have a public Outlook folder that I have users move spam too, then I run Mozilla to 'import' the email whn I get about 1000 spams. Then I find the mbox mail file and run sa-learn on it.

However, I get way too much email still and I even have my limit set to 0.

When installing updates to I lose the spam training that was done?

It sounds like it is not necessary to run sa-learn on ham. Is this correct?

Feature idea-

specify a special email address to forward spam too. Then SA can automatically run sa-learn against it. This was users could simply be instructed to forward spam to this address and there would be less admin to get a good trained system working. Any comments by the developers as to what it would take to implement this?

brandon

Subject: Re: Is the learner working?
Posted by InforMed Direct on Tue, 18 May 2004 20:53:22 GMT
View Forum Message <> Reply to Message

> - in the default setup of NoSpamToday!, the Bayes DB path is
> relative to the working directory. The mails you learn may end

Sorry, I omitted that we run this from a batch file where we change to the SA folder first.

> - if I'm not mistaken, it is not possible to invoke sa-learn
> for multiple mail files like this. You could write the file

It appears to work fine - it reports processing multiple messages.

> - the docs (sa\doc\sa-learn.htm) indicate that the sa-learn
> options should be before the mail file name.

Tweaked the batch file...

> You can dump the database (--dump) , and look at the tokens
> (words) learned and the number of tokens.

I've piped the output to a text file and counted the number of lines. Hopefully, after we've run the learner a few more times it'll contain a different number of lines, showing it's working.

Cheers, Rob.

Subject: Re: Is the learner working?
Posted by InforMed Direct on Tue, 18 May 2004 20:58:05 GMT
View Forum Message <> Reply to Message

> I have a public Outlook folder that I have users move spam
> too, then I run Mozilla to 'import' the email whn I get about
> 1000 spams. Then I find the mbox mail file and run sa-learn on
> it.

Similar idea except you're cutting out the step of using DBXtract to convert Outlook Express inbox into messages.

> However, I get way too much email still and I even have my
> limit set to 0.

That's the same problem we've got - okay, so No Spam Today is far cheaper than the subscription services but it appeared to work great when we first installed it but has steadily got worse. I'm assuming this is because SA hasn't been updated recently so we're still running off relatively old rules.

> When installing updates to I lose the spam training that was done?

The latest version had a option to retain the current training.

> It sounds like it is not necessary to run sa-learn on ham.
> Is this correct?

I think so - SA passes all messages that it classes as non-spam through as ham so there's a stream of ham messages passing to the learner.

> Feature idea-
>
> specify a special email address to forward spam too. Then SA
> can automatically run sa-learn against it. This was users
> could simply be instructed to forward spam to this address and
> there would be less admin to get a good trained system working.
> Any comments by the developers as to what it would take to
> implement this?

That would be neat! We could simply configure (say) spam@informed-direct.com into No Spam Today and forward to that. Hmm, thinks - would Exchange actually pass the email back out to the Internet or route it internally? I suspect the later considering that if I send a large email to myself internally using my Internet email address, it arrives instantly.

Cheers, Rob.

Subject: Re: Is the learner working?
Posted by bjsvec on Tue, 18 May 2004 21:41:01 GMT

View Forum Message <> Reply to Message

Good point about how Exchange handles emails to your own domain.. I suppose that could be worked around though. Just forwaring a free webmail account back in might do the trick. What threshhold do you set SA to? I have noticed that almost all legit email I get is -4.9 for some reason.. I set the threshold now to -1.0, but I still get at least 20 spams out of about 60-80 emails on my account alone..

Subject: Re: Is the learner working?
Posted by InforMed Direct on Wed, 19 May 2004 08:20:36 GMT
View Forum Message <> Reply to Message

> What threshhold do you set SA to? I have noticed that almost

We're still on the default of 5.0. How are you display/working out the scores of legit email?

Cheers, Rob.

Subject: Re: Is the learner working?
Posted by bjsvec on Thu, 20 May 2004 00:36:31 GMT
View Forum Message <> Reply to Message

I am set to -1.0 for the last day or so and am getting better results, still too much spam though.. Surprisingly I have no false positives yet. We probably get about 200 or so legit emails a day, so it's not a big sample, but still ok. I've been looking at the headers of my good email lately and it seems amazingly that 90% of them have a score of -4.9. So I am thinking of moving my threshold closer to that point until I get some false positives..

The real trouble is it is just too time consuming to monitor and tune the system all the time. Not sure about you, but I have a real job besides trying to block spam at my office ;)

brandon

InforMed Direct wrote:

> > What threshhold do you set SA to? I have noticed that
> almost
>
> We're still on the default of 5.0. How are you
> display/working out the scores of legit email?
>
> Cheers, Rob.
>

Subject: Re: Is the learner working?
Posted by InforMed Direct on Thu, 20 May 2004 10:47:50 GMT
View Forum Message <> Reply to Message

> The real trouble is it is just too time consuming to monitor
> and tune the system all the time. Not sure about you, but I
> have a real job besides trying to block spam at my office ;)

True - the whole aspect of virus and spam really p****s me off. We spend a lot of money and time combating annoying little oinks who think it's cool to write this stuff in the first place.

Rob

Subject: Re: Is the learner working?
Posted by fbennett on Wed, 26 May 2004 17:05:11 GMT
View Forum Message <> Reply to Message

I use SA on both Windows and Linux platforms and curiously, I get >90% detection on the Linux platform, but like you, only about 50% on the Windows platform. Not sure why, but I have learned a few things about SA along the way. One of the most important is that you don't want to *forward* messages and then use them to train SA. The reason is, anything that changes the message header (like forwarding) will cause SA to learn the message incorrectly. You'll actually be making your database worse instead of better. Also, MS products have an annoying habit of modifying the message headers regardless of what you might do to try and preserve them. I also heard that Mozilla was a good client to use, but I found that it creates multiple mailbox files and makes subtle changes in the headers too, making it less than ideal. Instead, I've adopted and highly recommend the Thunderbird client for this purpose. It's free, based on the Mozilla code and actually comes from www.mozilla.org, but it downloads all the messages into a single mbox compatible file. That file can then be used to train SA correctly. To get around the MS quirks, I have users move (note: that's *move*, not copy, not forward) their spams to a public folder on the exchange server, which I periodically download using Thunderbird and then train SA with the contents. The public folder is writable by all, but I'm the only one who can read it. I manually scan it for mistaken submissions before downloading it. There are a couple of other "gotchas" to watch for. One is that you must use the --mbox option with sa-learn. I just set up a simple batch file to do this on the Windows server and a similar script on the Linux box. I also compress and archive the spams in case I ever need to retrain from scratch. And with the Thunderbird client, you must set it up so that the folder you wish to download has the "offline" option set. Thunderbird will by default only download the message headers, but if you have this option set, you can right click the folder, go to Properties>Offline and tell it to Download Now. It will then download all the messages into its single mbox file which you will find in your user profile directory. I find that the SA training needs to be done at least weekly in order to maintain the highest hit rates. I also use the bigevil.cf file to augment my training. There's much more info available on SA at http://wiki.apache.org/spamassassin/ and elsewhere for those that wish to learn. After reading the other posts here, I recommend reading about bayesian classification and what SA does and does not do automagically.

Subject: Re: Is the learner working?
Posted by InforMed Direct on Fri, 28 May 2004 08:42:21 GMT
View Forum Message <> Reply to Message

fbennett wrote:

> about 50% on the Windows platform. Not sure why, but I have
> learned a few things about SA along the way. One of the most

Since upgrading to the latest version (with URL blocking) and reducing the threshold down from 5.0 to 3.0, it's doing better. Yesterday it caught 59 and missed 15. Before it was catching about 40 and letting 25 through.

We'll keep reducing down the threshold until we get too many false positives.

> important is that you don't want to *forward* messages and then

We don't - we move them to a shared mailbox so the header information is intact.

> to a public folder on the exchange server, which I periodically
> download using Thunderbird and then train SA with the contents.

We use a similar process except we go download the folder into Outlook Express and then use DBExtract to write the messages into standard (!) EML files.

Cheers, Rob.

Subject: Re: Is the learner working?
Posted by Heidner on Sun, 06 Feb 2005 00:08:06 GMT
View Forum Message <> Reply to Message

I've been working on a Perl script that opens an IMAP4 connection to exchange and reads/exports ham&spam from public mailboxs into two mbox flat files. The script also deletes the existing entries in the specified public folders. The mbox flat file can then be used as input into learn-spam.bat and learn-ham.bat.

I've been working with Exchange 5.5 on NT4.0 would anyone else be interested in trying it? The goal is to be able to setup a batch file that runs a couple of times a day to empty the ham & spam and update the SA database with the minimal amount of my attention.
