Internet Level Spam Detection and SpamAssassin 2.50

Matt Sergeant
Senior Anti-Spam Technologist
MessageLabs

MessageLabs Details

- Scan email mainly for companies but also personal scanning
- >10m emails/day
- Originally anti-virus only. Anti-spam is main focus in US
- Anti-virus scanning is 100% solution (with guarantee)
- Anti-spam is about 95% accurate
- Different problem scale though - 1:200 emails vs 1:3 emails

MessageLabs
the message is cle

How we work

- MX Records point to us
- Outgoing points to us too
- 20+ email processing racks worldwide
- Spam and Viruses stopped before they enter your network

MessageLabs
the message is cle

Technology

- Started off with SpamAssassin
  - It was the only decent spam scanner at the time
- Extended it, changed it, submitted patches, added custom code
- Became a lead SpamAssassin developer in the process
- Scared the pants off our business people in the process ;-)

SpamAssassin Intro

- SpamAssassin is a rules based heuristic engine combined with a genetic algorithm blah blah blah...
- Reality: *SpamAssassin is a framework for combining spam detection techniques*
- Probably around 30m users

MessageLabs
the message is cle

Spam Detection Stats

	DNSBLs	Phrase Matching	Heuristics (SA)	Statistics
Accuracy	0 - 60%	80%	95%	99%+
False Positives	10%	2%	0.5%	0.1%

Why Not Just Use Statistics Then?

- 99% is only true for personal email
- Statistical techniques learn what your personal email looks like
- Doesn’t work quite as well when you have users with dissimilar inboxes
- Live data testing accuracy about 80 - 95%

Further Details

- Some users (bless them) like to receive stock reports, marketing reports, sales leaflets, offers, deals of the century, HTML, and every piece of junk you can imagine, all via email.
- Yes, these people do exist, and they are our customers!
- Their statistics db entries tilt the database - and they are often right to do so

What Can We Do?

- Statistics don't consider all the details
  - Feature extraction is *hard*
- SpamAssassin examines a lot more of the email
  - Finer details of the headers
  - Regexps in the body text
  - Eval tests do things like "HTML tag percentage"
- So lets combine the two

Aside: How Statistics Works

- Extract *features* from the email
- Look up how many times we've seen that feature before in Spam and Ham
- Create probability for that feature
- Combine all the probabilities for all the features into an overall probability

Possible Method

- Store SpamAssassin results as features
  - e.g. P(ADVERT_CODE) = 0.95
- This works, but not very well compared to current scoring mechanism
- Reason: SpamAssassin doesn't have enough non-spam indicators to correctly weight against the spam features

Chosen Method

- Assign scores to probabilities
- Use GA to assign those scores

```
score BAYES_00          -4.000
score BAYES_01          -2.000
score BAYES_10          -0.500
score BAYES_20          -0.100
score BAYES_70           0.100
score BAYES_80           0.500
score BAYES_90           2.000
score BAYES_99           4.000
```

- Add this to the total along with everything else

Results

- With threshold at 7 to reduce FPs
- Using customer live emails as training and test data (split half and half)
- Accuracy: 99%
- FPs: 0.1% (all in the “hard_ham” folder - newsletters and other HTML mail)
- Overall, better than we could ever expect with pure SpamAssassin (pre 2.50) or pure Bayes

Questions?

MessageLabs
the message is cle

Extra Slides

Alternate Possible Schemes

- Decision Trees
  - Reduce number of rules run by a factor of 50
  - Speeds up SpamAssassin by an enormous amount
  - Not quite as accurate, though lots of work left to do in this area
- Boosting/ADABOOST
- Neural Nets
  - Each rule becomes a node in the network
  - Slow to learn (even compared to the GA)
  - Single layer perceptron may be comparable to the GA

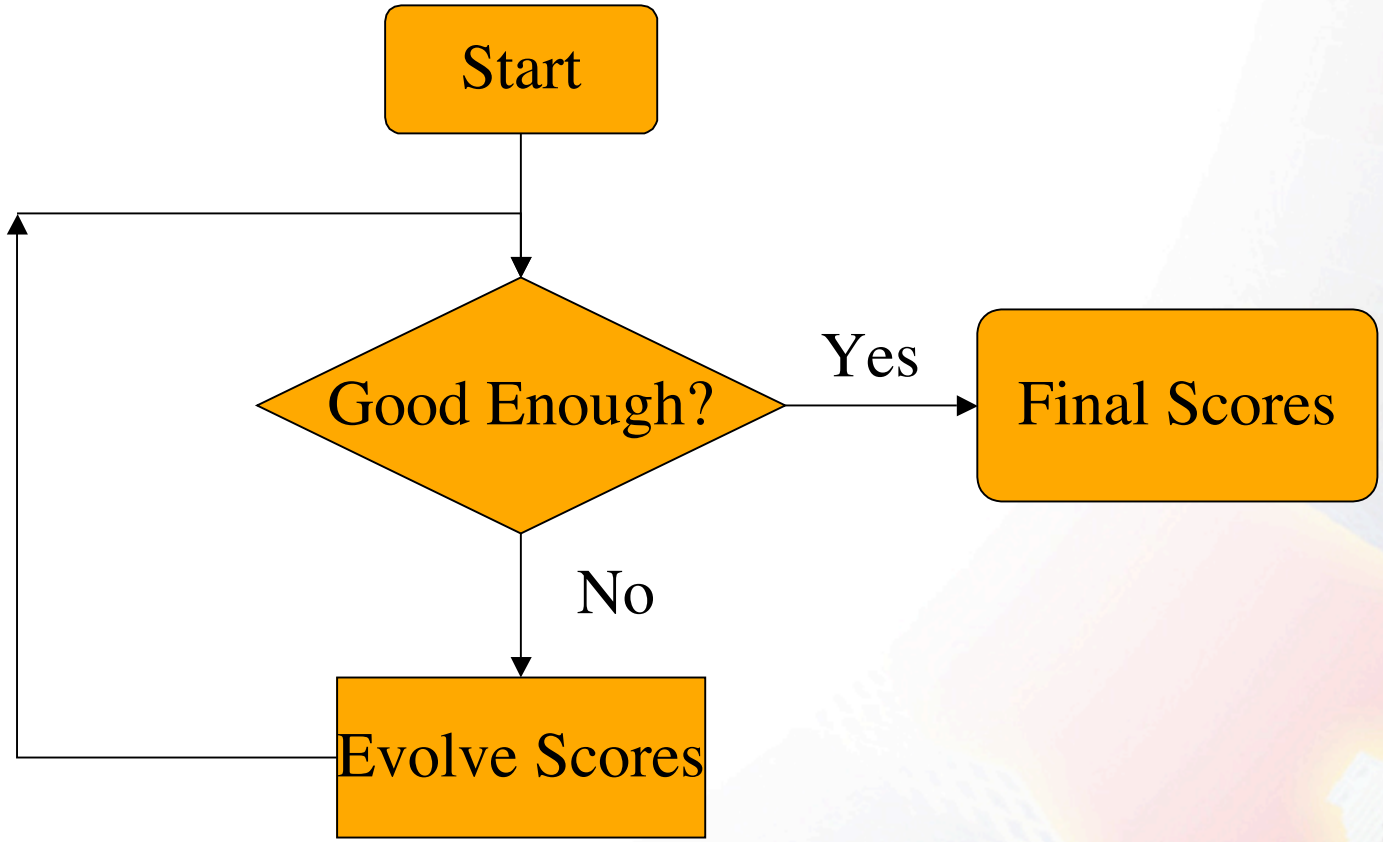
Future SpamAssassin Developments

- Auto-learn - trains bayes db continually on the email that gets scanned.
- Spam Signatures
  - Some effort by Razor, but massively inaccurate
  - Other work by Brightmail is proprietary
  - Must work like anti-virus signatures - human element?
  - Is it possible to make it open source?

SpamAssassin Retraining

- Most people install SpamAssassin and forget about it
- This is why Bayes kicks butt for personal installations
- But… If you treat SpamAssassin like a Bayes system - i.e. train on your own email, it gets much more accurate
- The training uses a genetic algorithm
- Achieving >99% accuracy via the GA isn't unheard of

SpamAssassin GA



GA In Action

```
Read test results for 3948 messages (7011 total).
Read scores for 1015 tests.
Iter #     Field      Value
1          Best       2.179058e+03
           Average    2.210501e+03
1234567890123456789012345678901234567890123456789012345678901234567890123456789
Pop size, replacement: 50 33

Mutations (rate, good, bad, var, num): 0.0066970 3 3 4744 0

Adapt (t, fneg, fneg_add, fpos, fpos_add): 0 0 0 0 0
Adapt (over, cross, repeat): 0 0 0

# SUMMARY for threshold 5.0:
# Correctly non-spam:   3268  46.61%  (99.60% of non-spam corpus)
# Correctly spam:       3381  48.22%  (90.64% of spam corpus)
# False positives:        13  0.19%  (0.40% of nonspam,    756 weighted)
# False negatives:       349  4.98%  (9.36% of spam,   1052 weighted)
# Average score for spam:  14.4    nonspam: -3.6
# Average for false-pos:   5.1  false-neg: 3.0
# TOTAL:                7011  100.00%

300        Best       7.093876e+02
           Average    7.119296e+02
```

MessageLabs
the message is cle

Full DNS Results

	OVERALL%	SPAM%	HAM%	S/O	RANK	SCORE	NAME
	20084	8138	11946	0.405	0.00	0.00	(all messages)
	100.000	40.5198	59.4802	0.405	0.00	0.00	(all messages as %)
	7.469	18.4198	0.0084	1.000	0.95	3.18	RCVD_IN_SBL
	5.377	13.2711	0.0000	1.000	0.95	2.66	DCC_CHECK
	4.212	10.3957	0.0000	1.000	0.94	0.30	X_OSIRU_SPAMWARE_SITE
	3.177	7.8152	0.0167	0.998	0.93	2.25	RCVD_IN_ORBS
	1.369	3.3792	0.0000	1.000	0.93	3.25	RCVD_IN_DSBL
	0.981	2.4207	0.0000	1.000	0.93	3.91	RAZOR2_CHECK
	1.797	4.4237	0.0084	0.998	0.93	1.00	RCVD_IN_OPM
	0.199	0.4915	0.0000	1.000	0.93	0.01	RAZOR2_CF_RANGE_01_10
	0.144	0.3564	0.0000	1.000	0.93	0.01	RAZOR2_CF_RANGE_11_20
	0.139	0.3441	0.0000	1.000	0.93	0.01	RAZOR2_CF_RANGE_21_30
	0.030	0.0737	0.0000	1.000	0.93	0.01	RAZOR2_CF_RANGE_31_40
	0.030	0.0737	0.0000	1.000	0.93	0.01	RAZOR2_CF_RANGE_41_50
	0.025	0.0614	0.0000	1.000	0.93	2.80	ROUND_THE_WORLD
	0.015	0.0369	0.0000	1.000	0.93	0.01	RAZOR2_CF_RANGE_71_80
	0.010	0.0246	0.0000	1.000	0.93	0.01	RAZOR2_CF_RANGE_81_90
	0.010	0.0246	0.0000	1.000	0.93	0.01	RAZOR2_CF_RANGE_51_60
	5.686	13.8486	0.1256	0.991	0.92	3.09	NO_MX_FOR_FROM
	2.285	5.5665	0.0502	0.991	0.91	0.61	RCVD_IN_RELAYS_ORDB_ORG
	0.393	0.9585	0.0084	0.991	0.90	0.01	RAZOR2_CF_RANGE_91_100
?	6.652	15.5935	0.5609	0.965	0.86	2.28	RCVD_IN_RFCI
?	17.865	40.4030	2.5113	0.941	0.83	0.01	RCVD_IN_NJABL
?	9.500	21.6269	1.2389	0.946	0.81	2.00	X_NJABL_OPEN_PROXY
X	12.005	26.3578	2.2267	0.922	0.76	0.77	RCVD_IN_UNCONFIRMED_DSBL
X	30.980	58.2453	12.4058	0.824	0.60	0.38	RCVD_IN_OSIRUSOFT_COM
X	1.917	3.3915	0.9124	0.788	0.46	0.36	X_OSIRU_DUL_FH
X	1.598	2.5928	0.9208	0.738	0.38	0.62	X_OSIRU_DUL
X	2.111	3.3669	1.2557	0.728	0.36	0.50	X_NJABL_DIALUP
ok	0.025	0.0369	0.0167	0.688	0.30	0.01	RAZOR2_CF_RANGE_61_70
?	0.274	0.1843	0.3348	0.355	0.25	-10.00	RCVD_IN_BONDEDSENDER
X	2.290	2.8508	1.9086	0.599	0.20	0.81	RCVD_IN_MULTIHOP_DSBL
?	0.070	0.0246	0.1005	0.197	0.01	2.29	BAD_HELO_WARNING