

jgc's spam and anti-spam newsletter

Feature Article: **Understanding Spam Filter Accuracy** November 16, 2004

jgc's spam and anti-spam newsletter is a biweekly email newsletter written by POPFile author John Graham-Cumming. Each newsletter contains a feature article on spam or anti-spam techniques.

To subscribe to the newsletter visit www.jgc.org/

This article is Copyright © 2004 John Graham-Cumming. If you want to use all or part of this article in your own publication or presentation please drop an email to antispam@jgc.org.

Claims

Brightmail¹ claims on their web site that their anti-spam solution has “95% effectiveness”; my personal POPFile² installation states that it is working with “99.84% accuracy with 5 false positives”; a review³ of ProofPoint's product says that it “stopped 94% of spam but has 26 false positives”.

The question is which of these products is the best? Is 5 false positives better or worse than 95% effectiveness? Without knowing how much email was analyzed and what the mix of spam and ham⁴ it's not possible to tell.

This article shows why these numbers are useless and then goes on to propose a way to measure the “spam filter batting average”. The batting average gives a uniform way to measure spam filter accuracy that makes filters comparable just by looking at the numbers.

The Mixture Matters

Suppose you receive 10,000 emails, of which 9,000 are spam and 1,000 are ham. If your spam filter claims “95% accuracy”, it's hard to know what that actually means in terms of the number of spams that will make it to your inbox, or the number of legitimate messages that will be misidentified and lost.

Suppose that the 95% is completely uniform, both spam and ham are handled with 95% accuracy: 5% of your mail is going to get handled incorrectly. That means that 5% of spam is going to make it to your inbox (that's 450 messages) and 5% of ham is going to be removed (that's 50 messages).

But perhaps 95% accuracy means that 95% of the spam is going to be removed (letting 450 spams through to your inbox) and at the same time doesn't tell you anything about how your ham is going to be handled. In the worst case the spam filter could remove 100% of your legitimate messages let through 450 spams and still claim “95% accuracy”.

A third scenario is that 95% accuracy is the overall number: 95% of your total messages will be sorted correctly, but that there's a difference between how spam and ham are filtered. If 95% accuracy means 100% accurate at removing spam and 90% accurate at handling ham, then you wont see a single spam, but you'll lose 100 legitimate emails.

A comic scenario is a spam filter with 100% effectiveness: it simply deletes all emails. Sure, it's 100% effective at removing spam, just happens to be 100% effective at removing all your legitimate messages too.

The other problem with being told a single number for the accuracy of a spam filter is that

1 <http://www.brightmail.com/>

2 <http://getpopfile.org/>

3 http://www.proofpoint.com/downloads/InfoWorld_Proofpoint_Appliance_Review_May2004.pdf

4 Ham is the term commonly used for non-spam email

that number means different things to different people. If your mixture of spam and ham was as presented above (9,000 spam and 1,000 ham) then a filter that uniformly handles 95% of messages correctly lets through 450 spams and loses 50 hams. But if your email is skewed the other way: you receive 9,000 hams for every 1,000 spams then you are losing 450 legitimate messages while seeing only 50 spams.

The only way to know how well a spam filter performs is by looking at two numbers: the false positive rate and false negative rate.

False Positives and Negatives

Legitimate messages that are misidentified as a spam are known as a **false positives**. **False negatives** are spam messages that make it through to your inbox.

False positives are the bane of any spam filter. People are far more tolerant of the occasional spam that makes it to their inbox, than legitimate messages that are incorrectly removed.

One problem with the terms false positive and false negative is that they aren't quickly understood by the average reader. What's needed are different terms that make the numbers clearer, and a uniform way of reporting accuracy numbers so that spam filters can be measured and compared.

The Batting Average⁵

So what I propose is that claims about the accuracy of a spam filter be given in the form "spam hit rate / ham strike rate". The spam hit rate is the fraction of spam correctly identified; the ham strike rate is the fraction of ham that is misidentified. You can think of the first number as measuring how well the filter will remove the spams, and the second number as measuring how badly the filter handles legitimate messages. The bigger the first number the better, the smaller the second the better.

So a spam filter that correctly handled 95% of spam messages, but loses 10% of hams would be said to have a batting average of .950/.010: it managed to hit 95% of the spams and get rid of them, but in doing so it struck out on 10% of hams and lost them.

Another way that .950/.010 can be read is that out of every 1,000 spams, 950 will be removed and out of 1000 hams, 10 will be lost.

My POPFile configuration was claiming 99.84% with 5 false positives. Digging further into the actual number of messages I've received in each category shows that I'm actually getting a spam hit rate of 99.8% and a ham strike rate of 0.1%. so we can say that my POPFile installation is batting .998/.001.

⁵ Yes, I realize that I'm pushing the baseball statistics analogy a bit far here, but, hey, I'm British :-)

As spam filters get better it may be necessary to extend beyond three digits of accuracy, for POPFile I have more digits of accuracy and can say that its batting average is .9982/.0008: or for every 10,000 spams I receive 9,982 will be removed and I'll lose 8 out of every 10,000 real messages.

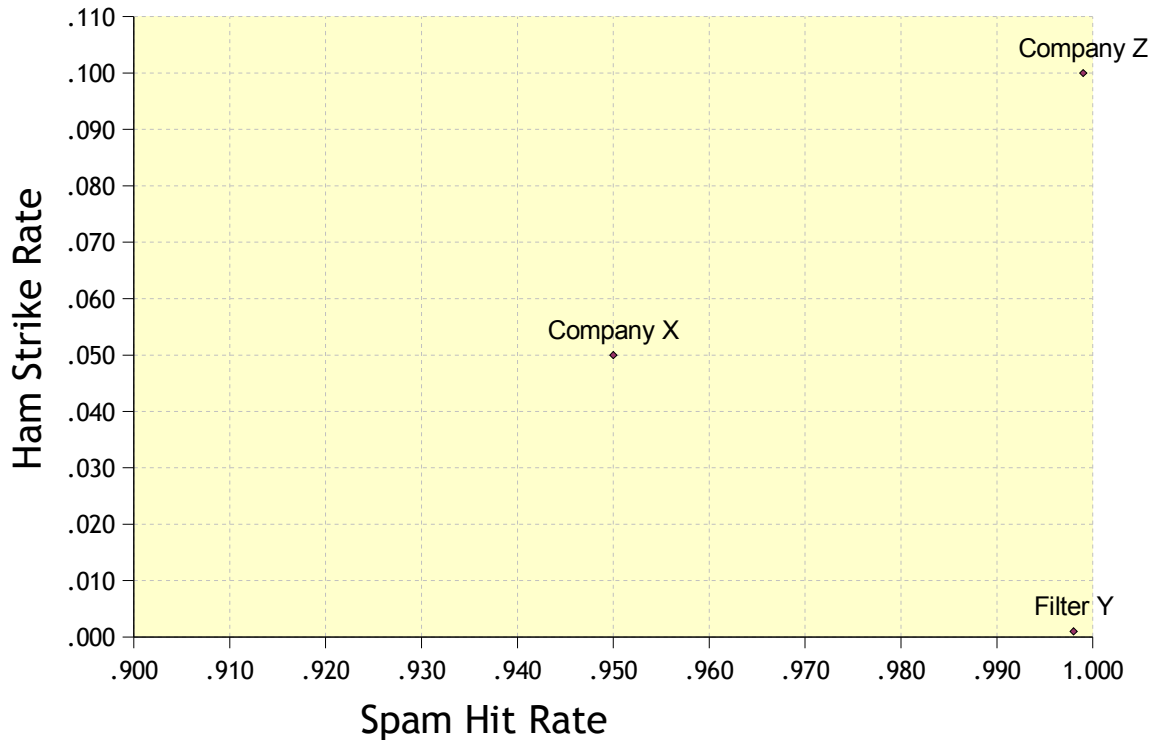
To calculate these two numbers you work out the following:

```
spam hit rate = total number of spam messages correctly removed /
                total number of spam messages received.
ham strike rate = total number of ham messages incorrectly removed /
                  total number of ham messages received.
```

Armed with just those two numbers you can compare how well a spam filter will remove spam (the first number) and how many legitimate messages will end up being collateral damage (the second number)

Another way to consider these two numbers is to draw a graph. Filters are "better" if they are towards the bottom, right corner of the chart. In this example Filter Y is both better at removing spam and better at handling ham than Company X or Company Z. Company X isn't very good at either. Company Z's filter is good at filtering spam, but has a fairly high ham strike rate.

Comparing Three Filters



Next time you're comparing spam filters try asking about the filter's "spam filter batting average".

If you can't live without a single number to measure spam filter accuracy that try taking the batting average and performing the division it implies: e.g. In my POPFile example I had a spam filter batting average of .998/.001 which yields a combined score of 998 when the division is done. A less efficient spam filter with a batting average of .950/.010 yields a combined score of 95.

The combined score decreases as the spam hit rate decreases and decreases as the ham strike rate increases. It's still not as informative as the two number batting average, but at least makes a handy rough estimate. The bigger the number the better.

HOW DO THEY KNOW?

A final thing to consider when looking at spam filter accuracy numbers is "how does the vendor know that's the figure?". After all, someone has to go through and manually identify whether a message is a spam or a ham before a filter can be measured.

If you see a vendor claiming that their filter is 99.999% accurate, that means it made a mistake on 1 in 100,000 emails. Ask them how they check that figure?