

August 20th 2005

# **Fighting spam blogs**

a hypothesis

**Kailash Nadh**

# The hypothesis..

This paper discusses certain methods which could be used to determine whether a blog is spam or not. Please understand that this article has nothing to do with comment spam (which can be easily and effectively tackled)

## What is a spam blog?

A blog that is created with the sole purpose of reaping ad revenues and capturing search engine results can be called a spam blog. In most of the cases, a spam blog is created/run and maintained by an automated program, which we commonly call 'bots'. Some spam blogs disguise so well that you sometimes end up reading it for hours !

## How a spam blog works?

As I said, a spam blog is created (usually) by a bot. For the time being, consider the term 'Mortgage refinance'. A spam blog's primary objective being getting search traffic, it's domain name would be something like this *mortgage-refinance-info.com* or *mortgage-refinance.some\_blog\_host.com*. To keep itself alive, it'll crawl directories, search engines, rss feeds etc.. and collect information on 'Mortgage refinance' thus preparing a neat collection of information. The next thing obviously, is posting this info regularly. I have seen blogs listing live news feeds related to a certain subject (via XML feeds obtained from the news publishers). An interesting fact is that 'Poker, Slot machine' spam blogs are very rare. They seem to be limited to comment spam.

## The challenges

Earlier, it was easy to track a spam blog (for a computer program) because of its peculiar nature, i.e. loads and loads of ppc listings, affiliate links. But now things have changed. Spammers got better and now you rarely see any ppc listings or affiliate links (of course, except for the popular Contextual ads) on a spam blog. This makes it even more difficult for a computer program to determine the blog's true nature. There are even blogs that have no ads, no banners, nothing, but leeches content.

## How to possibly tackle?

Now the question is how to track/detect a spam blog? We could employ the tactics that anti-spam (email) systems use. The very popular SpamAssassin is the best example. It works by performing a series of tests on an email, assigning scores and then using the Bayesian theorem to predict the probability of an email being spam.

Before performing any tests or arriving at any conclusions, I started off by collecting around a hundred spam blogs and closely studying their characteristics. I found that 90% of the spam blogs have certain features in common, which could be well used against them.

# The hypothesis..

First off, I prepared a list of keywords, that could possibly be spam. These were divided into two sets, High priority and low priority. Here's a few of the keywords I experimented with.

High Priority - *Looking for, drug, viagra, mortgage, debt..*

Low priority - *Cheap, deal, invest..*

( Surprised to see my keywords ? )

## 1. The URL

Every spam blog has a spammy looking url. So the first step should be examining the target's url. A possible trigger could be the number of '-' in the url. Most of the spammy urls have at least 2 hyphens. But even with 4 hyphens, the blog needn't be spam. Then next thing would be running a keyword analysis on the url (both high and low). Presence of a high priority keyword means more probability. Presence of a low priority keyword can be considered as the presence of 1/4th of a high keyword.

## 2. Ads

The next step would be examining the presence of ads on the blog page. A simple search for the number of occurrences of an ad keyword could yield valuable results. Searching for the number of occurrences of 'google\_ad\_client' will give the number of Adsense units on that page. Same way, a search could be done for the presence of popular ad programs on the page. More number of ad means, more possibility of spam.

## 3. Page content

Now the most important step. We run a keyword analysis on the page, low and high. The results are recorded then. As discussed ('The URL'), certain score can be assigned for the presence of each high/low keyword. (The low keywords may be found in non-spam blogs too) To get a proper result, the result of the keyword analysis could be compared to the total number of words on that page (After stripping html, special characters and removing common words like 'the, when, what' .. )

TIP: If a page has the work 'Looking for' or the word 'Information' in abundance, its chance of being spam is very very high.

## 4. The title

The title says it all ! No 'mortgage' spam blog in the world would give the title 'Books', but 'Mortgage' itself. So taking words out of the title and counting the number of occurrences of those words on the page body could produce surprising results. ( I did this test on a 'Dance teaching' spam blog. The word 'dance' was found 948 times on the page. The total number of words on the page was around 2048. That means 50% of the page was 'dance' ). Comparing this number with the total number of words will give a good idea about the genuinity of the blog.

## 5. Words

Again, this should be under the 'Page content' section. Even if the title test doesn't work out properly, we could still cover up that part. Get all the words from the page (After stripping html, special characters and removing common words like 'the, when, what' .. ). Out of those words, check for word redundancy, similar to the title test, if a word's occurrence is very high, it could possibly be spam.

## 6. Links

This test could prove really valuable. Get the total number of hyperlinks on the target page. Now count the number of hyperlinks to external sites and the number of local links. If the number of external links is very high, it could be.. (again, this is a blunt test). A good idea is to do our good old keyword test (high/low) on the hyperlinks. After all, a spam blog will naturally have spam links (Scanning the hyperlinks against the title words is an excellent idea)

# The hypothesis..

## 7. \$Money\$

This is a very interesting test, but a vague one. Check for the presence of 'money', i.e. occurrences of strings in the \$0, \$.0, \$0.0 format. I used a regex to get the number of 'money' entities on a typical spam page (On a 'pet supplies' blog page, I found 59 occurrences)

## 8. Compression

This is highly experimental. The idea struck me when I was experimenting with gzip routines in PHP. The gzip compression (and all other) algorithm works by searching the content (to be compressed) for similarities, and replacing them with a reference. That is, when compressing the string 'Goooogle', all the o's will be replaced by a reference, thus compressing the system. Doing this simple test would make it more clear. I created a plain text file with 99999 'a' in it. The filesize was around 9.5 KB. After compression, the filesize reduced to 146 bytes. Then I took a random html page (9.46 KB) and compressed it. The resultant filesize was 4KB !

So, if a spam blog contains certain words over and over, again and again, the compression ratio of that page compared to a non-spam page of the similar size would produce a big difference. This is completely theoretical, but might prove effective. (This worked on 6 spam blogs out of the 10 I tested)

## Conclusion

As the title of this paper says, everything I said here is hypothetical. If done properly and combined with something like the Bayesian theorem to predict the probability out of the results obtained, it might just work magic! I have experimented with all the tests mentioned and they all worked fine, well almost.

**Kailash Nadh**  
**August 20th 2005**

Email : [mail@kailashnadh.name](mailto:mail@kailashnadh.name)  
Url : <http://kailashnadh.name>

**The end**