umbria
303.217.8299

SPAM in the blogosphere

umbria
303.217.8299

Key Findings

According to Umbria Inc.,

- Spam blogs currently comprise between ten to 20 percent of the blogs in the blogosphere growing from two percent in March 2005. For the week of October 24, 2005, Spam blogs comprised 13 percent of the blogs in the blogosphere (2.7 million out of 20.3 million blogs).
- Blog search engines, which utilize keyword search exclusively, are rife with Spam blogs. In a limited test of three blog search engines, an average of 44 of the first 100 blog search listings were Spam blogs.
- Keyword search approaches don't appear to effectively filter Spam blogs due to their real-time nature.
- Because Umbria uses a data mining approach for blog analysis, Umbria is able to filter out up to 95% of Spam blogs before performing analysis, yielding cleaner data and more accurate results than other blog analysis services.

Abstract

According to Technorati, in October 2005 there were more than 20 million blogs (or web logs) with over 80,000 new web logs created each day (or one new blog created every second of every day). While Technorati estimates the blogosphere is doubling in size every 5.5 months, estimates of the overall percentage of Spam blogs (or Splogs) contribution to the growth rate vary.

Spam blogs – which are fake blogs designed to fool or "spoof" search engines and drive traffic to sites peddling everything from consumer electronics to online gambling to pornography -- are proliferating in the blogosphere at an astronomical rate. Umbria's analysis of the blogosphere indicates that between 10%-20% of the blogosphere is now comprised of these Spam blogs.

This paper also offers insight into the magnitude of the problem of Spam blogs, and provides results from three blog search engines to demonstrate the impact of Spam blogs on search results. Also included are techniques used by Umbria to filter out Spam blogs in order to provide for cleaner and more accurate analysis of the blogosphere.

Introduction

Spam blogs (or "Splogs") are an ever-increasing problem in the blogosphere. The presence of Spam blogs makes finding relevant search engine results challenging. Worse, when marketers use blog analytics to ascertain key insights about brands, products, and competition, Spam blogs can significantly skew blog analysis results. Umbria is one of the few companies currently addressing the Spam blog problem for blog analysis. Using a combination of leading edge technologies and

manual intervention processes, Umbria can eliminate up to 95% of Spam blogs from the data used to analyze the blogosphere.

Umbria is in a unique position to address the problem due to our approach to blog analytics. While most blog search tools and companies use real-time keyword search as the approach for analysis, Umbria uses a data mining approach, allowing for data cleansing prior to analysis -- resulting in the removal of "dirty" data, such as Spam blogs. This paper will proceed to demonstrate the impact of Spam blogs on analysis, and provide insights into how Umbria is addressing the issue to ensure cleaner data analysis – yielding more accurate analysis for Umbria clients.

Spam Blogs Defined

Spam blogs, according to Wikipedia.org, are defined as "Web Log (or "blog") sites which the author uses only for promoting affiliated websites. The purpose is to increase the PageRank of the affiliated sites, get ad impressions from visitors, and/or use the blog as a link outlet to get new sites indexed. Content is often nonsense or text stolen from other websites with an unusually high number of links to sites associated with the Spam blog creator which are often disreputable or otherwise useless Web sites."

"The term 'splog' was popularized around mid-August 2005 when it was first used by some high profile bloggers but appears to have been used a few times before then to, dating Spam blogs back to at least 2003."

Spam blogs are quickly becoming a problem in the blogosphere; particularly for blog search engines and other blog or consumer generated media analysis platforms. Recent Spam blog analysis results by Umbria in October 2005 indicate that between 10%-20% of the 20+ million blogs in the blogosphere are comprised of Spam blogs. Spam blogs pollute search engine results, leaving users unsatisfied with the search results.

Several Spam blog reporting services have been created for good-willed users to report Spam blogs with plans of offering these Spam blog URLs to search engines so that they can be excluded from search results. Splog Reporter was the first service of this kind.

These results are problematic on several fronts:

- **Spam blogs are compromising blog search engine results.** Because blog search engines (e.g., Technorati, Intelliseek's BlogPulse, Ice Rocket, etc.) operate in near real-time, there is not adequate time to filter out Spam blogs, and as a result they are not eradicated from search results.

umbria
303.217.8299

- **The prominence of Spam blogs in the blogosphere skews blog search and analysis results**. Analysis of "dirty" data means marketers have a false sense of the magnitude of discussion and the topics of discussion regarding their brands, products and competition. Informed business decisions require high quality data.

- **Spam blogs leave users of blog search engines dissatisfied with results**. Because Spam blogs are established to elevate search rankings, Spam blogs typically appear in the first few pages of search results relegating legitimate and relevant search results to lower ranking pages which are not typically viewed by search engine users.

Let's take a look at a case study highlighting the magnitude of the Spam blog problem among blog search engines.

Case Study

To test the magnitude of Spam blogs in the blogosphere, Umbria initiated searches on six leading brands and products using three commonly used blog search engines. The first 100 results were then analyzed for the occurrence of Spam blogs. This test was conducted during the week of October 24, 2005 using Intelliseek's BlogPulse, IceRocket, and Technorati. Though the results of this test are directional only, they do give an indication as to the scale and scope of the Spam blog problem. This test was replicated three times for validation with little to no variation (no more than +/- 3). All search terms were entered in quotations to ensure that words would appear together in the search results. The results per each search engine are averaged below:

Figure 1: Number of Spam blogs Per First 100 Blog Search Results for Selected Brands and Products

Keywords	BlogPulse	IceRocket	Technorati	Average
Apple iPod	75	80	71	75
Starbucks	28	18	7	18
Gap Kids	46	55	28	43
Sprint Wireless	91	92	90	91
McDonalds	20	5	8	11
Pepsi	26	24	19	23
Total	**286**	**274**	**223**	**261**
Average	**48**	**46**	**37**	**44**

On average, 44 of the first 100 blog search results (across all three blog search engines) were Spam blogs.

Some categories and brands appear to have higher penetrations of Spam blogs than others. For example, consumer electronics and services sold by resellers and affiliates online like Sprint Wireless and Apple iPod have a much higher proportion of Spam blogs in the first 100 results versus brands and products not sold online, like Starbucks, McDonalds and Pepsi.

In summary, Spam blogs are pervasive and compromise search results, diminishing the value of search engines to end users, while misleading those who rely on their results for analysis and research.

How Search Engines Work and Why It Is Difficult to Address the Issue

A keyword or phrase search triggers a search engine to employ technology to crawl the internet/blogosphere to find, categorize, index, and rank web sites, web pages, blogs and other online sites, by time and/or relevance. Internet search engines like Google or Yahoo, or blog specific search engines like BlogPulse, IceRocket, or Technorati, typically use similar approaches to provide search results to users. In general, most of the blog search engines find, categorize, index and rank blogs that are new or modified on average six hours after a change is made.

The primary objective of any search engine is to deliver timely and relevant results. The more relevant the results, the greater the value provided to the user. However, the advent of online advertising and affiliate programs created a competing objective for content creators which have led to the creation of a whole industry centered on trying to improve search engine rankings for their sites in order to incite click-throughs by users. It is this phenomenon that is partially responsible for compromising the relevance of search results using today's search engines.

In addition, blogging is exploding. Technorati estimates the size of the blogosphere is doubling every 5.5 months. Blogs are extremely easy and inexpensive to create, thus lowering the barriers to entry for anyone who wishes to have an online presence. Technology exists to automatically create blogs at little or no cost, and to do so in a way that creates hundreds and even thousands of blogs in very little time. Since blog hosting is a nascent industry with little technological protection against the proliferation of Spam blogs, thousands of Spam blogs are created daily to drive traffic to a specific blog and incite click-throughs or e-commerce transactions.

Since search engines are designed to find, categorize, index and rank blogs for presentation in search results, many people have found ways to exploit search algorithms to get their blogs and sites listed higher in search engine rankings. While search engines want to eliminate this practice, creative-minded Spam bloggers, have learned how to "spoof" search engine algorithms by homing in

on frequently searched keywords and phrases. Given the low cost barriers to entry, Spam bloggers can afford to key in on search phrases for just about any search conducted on the Internet and find creative ways to make their sites appear at the forefront of search engine results.

A Closer Look: Example of How a Spam Blog may Generate Revenue

Two common ways for web publishers and spam bloggers to earn revenues are by placing small boxes of text ads on their sites or blogs using a traditional Affiliate Program (highlighted in the figure below in the red box on the right), or search engine advertising programs (highlighted in the figure below in the red box under "Links" on the left).

Example of a Spam Blog

The affiliate programs (right side of the box, the one marked "cellular one ringtone") are widely used as a method for web-based businesses to expand their sales channel and drive traffic to their web sites in order to conduct a financial transaction, sign up or register prospective customers.

If a purchase is made on the Affiliate owner's web site, the web-based business normally compensates (usually as a percentage or fixed fee) the blog site owner that originally drove the paying customer to the business' web site.

Advertisers fund search engine advertising programs. In the example above, if a visitor to this Spam blog clicks on the "Ringtones" link in the box (on the left side), then the search engine and the owner of this Spam blog divide the advertising revenue paid by www.faster-results.com, the advertiser who paid the search engine advertising program to place this ad. By focusing on highly sought after keywords that tend to have high costs per click placement fees, owners of Spam blogs make money by driving traffic to advertisers.

The Pieces of the Puzzle

Establishing a blog is very easy. Software is readily available that allows users to simultaneously create many blogs at once. Once the blogs are established using targeted keywords, blog spammers sign up for search ad and Affiliate Programs. (Because it is easier and less time consuming to create new blogs vs. update or alter old blogs, the creation of Spam blogs continues to increase).

Blog search engines then crawl and post search results based upon relevancy of terms and time frame.

What are the search engines and blog hosting companies doing to address the problem? Unfortunately, due to the real-time nature of search and the trade-offs associated with ease of use for blog creators, it's very difficult for these companies to effectively address the Spam blog problem.

To date, these are some of the approaches taken by search engines and blog hosting companies to prevent Spam blogs:

1. Black lists – central repositories for submitting known Sploggers (e.g. splogreporter.com)
2. CAPTCHA systems, yet these can be compromised using software
3. Standardization efforts – the facilitation of information sharing among blog search engines and attempts to develop standards to address the problem.

Unfortunately, these efforts have provided limited effectiveness to date.

umbria
303.217.8299

Umbria's Quality Assurance: Identifying and Removing Splogs from Analysis

Umbria's mission is to enable companies to derive meaningful analysis insights from the unprompted, unsolicited, and timely opinions and perceptions of the online community – that are free from dirty data rife with Spam blogs.

Since its inception, Umbria has grappled with the issue of Spam blogs. Unlike some market intelligence companies that may rely largely on keyword search technology, Umbria collects and analyzes blog information using a data mining approach. Utilizing three levels of filtering, Umbria attacks the Spam blog problem and constantly improves on Spam blog detection to ensure our clients receive the best possible results. We constantly refine and modify the techniques, which is necessary to keep up with the rapidly evolving techniques of Spam bloggers.

Umbria's three-pronged approach eliminates up to 95% of Spam blogs from data prior to analysis. Umbria's approach is as follows:

1. Automated Machine Learning algorithms, detect up to 80% of Spam blogs
2. Blacklist approach, detects another 5 to 10% of Spam blogs
3. Manual inspection and review, eliminates final 1 to 5% of Spam blogs

Utilizing Machine Learning techniques and indicators from the content (e.g., URLs, links, etc), Umbria analyzes all aspects of blogs and determines whether it's a Spam blog.

Umbria also maintains a "blacklist" of known Spam blogs and continually updates this list with known Spam blog URLs. The blacklist enables the filtering out of an additional level of Spam blogs based on our knowledge of specific Spam bloggers and their techniques.

The last step is the manual inspection of blogs during our Quality Assurance analysis. Spam blogs that may have made it to this step are identified and used to improve the detection and elimination processes in steps 1 and 2 by using them to hone the machine learning algorithms and entering them into the blacklist.

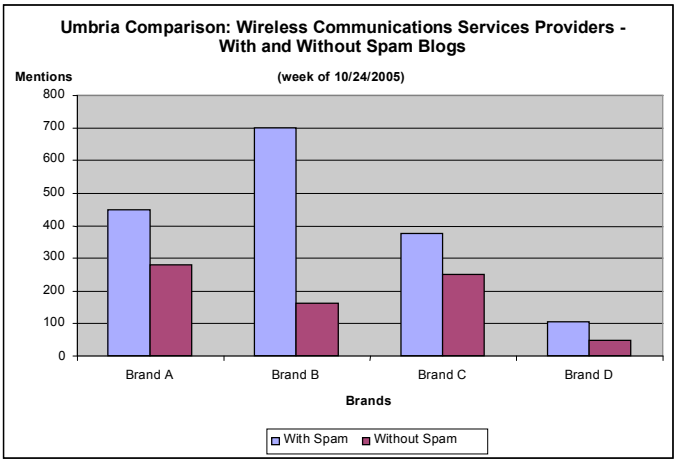
This process is continually repeated over time providing deeper knowledge and better protection against Spam blogs going forward.

umbria
303.217.8299

Umbria Case Study

Here is an example of how Spam blogs can skew the results and misrepresent the actual amount and quality of conversations in the blogosphere.

The chart below represents 4 wireless communications services providers tracked in the blogosphere the week of October 24, 2005. The bar on the left is a representation of the data *before* our 3-step Spam blog filtering process while the bar on the right represents the same 4 brands *after* our 3-step Spam blog filtering process.



Source: Umbria Inc.

With Spam blogs included in the results, one might have deduced that Brand B was the most often mentioned brand. Yet, after removing Spam blogs it's apparent that Brand A and Brand C were more often mentioned. Results for Brand B were over-inflated by over 300% if Spam blogs were included for analysis. Filtering for Spam blogs ensures against erroneous conclusions being drawn from the data. Like all other marketing research, brands and companies must rely upon clean data as a basis for analysis to drive business decisions.

umbria
303.217.8299

Conclusion

Spam blogs are a growing problem. Every day Spam blog creators innovate new methods and techniques to spoof or fool blog hosting companies and search engines into elevating their position in search results to increase the probability of click-through. An analysis of six brands and three blog search engines during the week of October 24, 2005 found that on average 44 of the first 100 listed blogs were Spam blogs.

Umbria's research shows that in October 2005 between 10% and 20% of the blogosphere is comprised of Spam blogs. Umbria insures that our blog data analysis is largely free of Spam blogs that can greatly skew analysis results. Our three stage de-spamming approach insures greater accuracy in analysis and reporting of results.

As companies increasingly look to derive insights and analysis from the unprompted, unsolicited, and timely opinions and perceptions of the blogosphere, it's imperative that data be cleansed from Blog spam to ensure that business decisions aren't derived from erroneous results.

Umbria is committed to ensuring that blog analytics are free from impure data sources and will continue to lead the market in addressing the problem.

About Umbria

Umbria is a consumer intelligence company that collects and distills commentary of the online community to help clients make more informed business decisions. From millions of blogs, message boards, opinion sites and other public forums, Umbria excavates the opinions of consumers who write about their experiences, preferences, likes and dislikes regarding the products or services they buy and use.

For more information about deriving consumer intelligence from online sources including blogs, message boards, opinion sites and other public forums, please contact Umbria at **303.217.8299** or go to www.umbrialistens.com.