Online Spam-Blog Detection Through Blog Search

Linhong Zhu, Aixin Sun
School of Computer Engineering
Nanyang Technological University
{zhul0003,axsun}@ntu.edu.sg

Byron Choi
Department of Computer Science
Hong Kong Baptist University
bchoi@comp.hkbu.edu.hk

ABSTRACT

In this work, we propose a novel post-indexing spam-blog (or splog) detection method, which capitalizes on the results returned by blog search engines. More specifically, we analyze the search results of a sequence of temporally-ordered queries returned by a blog search engine, and build and maintain blog profiles for those blogs whose posts frequently appear in the top-ranked search results. With the blog profiles, 4 splog scoring functions were evaluated using real data collected from a popular blog search engine. Our experiments show that the proposed method could effectively detect splogs with a high accuracy.

Categories and Subject Descriptors

H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval—Information filtering

General Terms

Experimentation

1. INTRODUCTION

Formally defined in [2], a spam blog is “a blog created for any deliberate action that is meant to trigger an unjustifiably favorable relevance or importance, considering the blog’s true value”. As the presence of splogs could greatly reduce the quality of search results and waste network resources, among others, splog detection has recently attracted much attention from research.

Most existing work on splog detection focuses on systems that could be used as filters that protect blog search engines’ index from spam, also known as pre-indexing splog filters, illustrated in Figure 1. In contrast, we study the problem of post-indexing splog detection where we aim at detecting those splogs that have already successfully slipped through pre-indexing splog filtering, e.g., [4]. Our proposed technique exploits user search queries and their top-ranked blog posts returned by search engines, shown in the right-hand side of Figure 1. The splogs detected could then be fed into either the search engine index, the pre-indexing splog filters, or subsequent post-indexing splog detection. We remark that our technique can be implemented on top of any pre-indexing splog filters.

Copyright is held by the author/owner(s).

ACM ’08, October 26–30, 2008, Napa Valley, California, USA.
post graph constructed with all posts in $R_{d,t}$. Each vertex in the post graph is a post $p_i \in R_{d,t}$. An edge from post $p_i$ to post $p_j$ is created if and only if $p_j$ is among the $k$-nearest neighbors of $p_i$, where $k$ is a user-defined parameter ($k=20$ in our experiments). The similarity between two posts is computed using Kullback-Leibler (KL) divergence based on the language models constructed using the words in both title and content of the posts. Based on the constructed graph, we define weighted out-going degree of each post to be the sum of the weights of outgoing edges of the post. The clustering coefficient of a post is given by $\frac{2\sum w_{j,i}}{|N_i|(|N_i|-1)}$; where $N_i$ is the set of immediate neighbors of $p_i$; $p_j \in N_i$; $w_{j,i}$ is the similarity between $p_j$ and $p_i$; and $|N_i|$ denotes the number of elements in set $N_i$.

Based on the assumption that posts from splogs are likely to be dissimilar to those from legitimate blogs, the spam-post score of a post $p$, $p$.spam score, is derived using a distance-based outlier detection algorithm [3], where the distance between $p_i$ and $p_j$ is defined by $\text{dist}(p_i, p_j) = 1 - \cos(p_i, p_j)$. The spam-post score of a post is assigned by the outlier detection algorithm if it is an outlier. If a post is not an outlier but is very similar to the spam seeds, then the post is also assigned a spam-post score. Otherwise, the spam-post score of a post is 0. All newly detected spam-posts are considered as part of the new spam-post seeds. To avoid maintaining a large pool of spam-post seeds as well as to reflect the recent spammer’s tactics, the out-dated seeds that are similar to the newly-added seeds will be discarded.

2.2 Spam-blog Detection

Given a blog profile, we present three (obviously non-exhaustive) scoring functions based on the heuristics stated below, denoted by $SF_1$ to $SF_3$. Each of them independently attempts to estimate the likelihood of a blog being a splog. For the ease of discussion, all state tuples in a blog profile $b$ is denoted as $ST$. A blog profile consists of the blog’s URL and a sequence of blog state tuples, each of which is denoted as $(t, \ell, p$.spam score$)$.

$SF_1$: Inblog increment over time. Based on a recent study on evolving graph data, links among legitimate blogs are expected to follow a densification law [1]. $SF_1$ is therefore proposed to detect sudden shrink$^2$ in inblog increment. Let $\ell'$ be the estimated number of inblog links by a linear regression model during $W$ time, $\ell$ be the actual number, and $\ell$ be the average derived from ST; the likelihood of $b$ being a splog is given as $(\ell' - \ell)/\ell$.

$SF_2$: Correlation between the number of posts appearing in top search results and inblog increment. It is expected that a blog would attract more inblog links if its posts frequently appear in top-ranked search results, as those posts were expected to reach more readers. Hence, the blogs whose posts frequently appear in top-ranked results but do not attract inblog links are more likely to be splogs. $SF_2(b) = 1 - p$, where $p$ is the correlation coefficient between the two values derived from $ST$.

$SF_3$: Average spam-post score. As a legitimate blog often contains relatively fewer spam posts, the average spam-post score derived from $ST$ can be used to detect splogs: $SF_3(b) = \frac{\sum p$.spam score$(ST)}{|ST|}$

$SF_4$: To aggregate the efforts of the three proposed scoring functions in splog detection, we take the average of the three scoring functions as the aggregated splog score of a blog: $SF_4(b) = \frac{1}{3} \sum_{i=1}^{3} SF_i(b)$.

3. EXPERIMENTS

The proposed splog detection framework was evaluated using real data collected from Technorati$^3$. We collected the top-15 popular queries published by Technorati every 3 hours from Nov 06 to Mar 08 [5]. For each query, the top-50 blog’s posts were retrieved in XML format through the Technorati API.

To measure the accuracy of the splogs detected, we manually labeled the top-200 blogs with the highest spam scores, and report the Precision@N where $N$ varies from 20 to 200. The precisions of top-N detected splogs of the 4 scoring functions are reported in Figure 3. It is clear that the aggregated scoring function $SF_4$ outperformed all the rest. Among the other three scoring functions, both $SF_2$ and $SF_3$ were fairly effective. $SF_1$, based on the assumption that the inblog links increase along time, was worse than $SF_2$ and $SF_3$. One possible reason could be the slow inblog increments among most legitimate blogs.

Our experiments have shown that splogs can be detected from search results with a very high accuracy. It is worthwhile mentioning that the splogs detected are those that have already successfully passed the pre-indexing filter.

4. REFERENCES


$^2$We have also tried the detection of sudden bursts as opposed to sudden shrinks. Nevertheless it was not effective in our experiments. Due to page limit, we choose to report the scoring function involving detection of shrink only.

$^3$http://www.technorati.com/pop/