SPLOG DETECTION USING CONTENT, TIME AND LINK STRUCTURES

Yu-Ru Lin *Hari Sundaram*

Arts Media Engineering, Arizona State University

Tempe, AZ 85281

Email: {yu-ru.lin, hari.sundaram}@asu.edu

Yun Chi *Jun Tatemura* *Belle Tseng*

NEC Laboratories America

Cupertino, CA 95014

{ychi, tatemura, belle}@sv.nec-labs.com

ABSTRACT

This paper focuses on spam blog (splog) detection. Blogs are highly popular, new media social communication mechanisms and splogs corrupt blog search results as well as waste network resources. In our approach we exploit unique blog temporal dynamics to detect splogs. The key idea is that splogs exhibit high temporal regularity in content and post time, as well as consistent linking patterns. Temporal content regularity is detected using a novel autocorrelation of post content. Temporal structural regularity is determined using the entropy of the post time difference distribution, while the link regularity is computed using a HITS based hub score measure. Experiments based on the annotated ground truth on real world dataset show excellent results on splog detection tasks with 90% accuracy.

1 INTRODUCTION

This paper addresses the problem of spam blog (splog) detection using temporal and structural regularity of content, post time and links. Splogs are undesirable blogs meant to attract search engine traffic, used solely for promoting affiliate sites. The splog detection problem is important – blogs represent highly popular, content-rich (text, audio, video, images) new media for communication, and the presence of splogs degrades blog search results as well as wastes network resources.

Splogs are relatively new phenomena; however there has been work on web spam detection. While there are critical differences between the two, a review is useful. Prior work to detect web spams can be categorized into content and link analysis. In [8] the authors distinguish web spams from normal web pages based on statistical properties in the content, such as number of words, average word length, or term frequencies in title, anchor text, tokenized URLs, etc. In order to combat spam blogs (or splogs), [7] extracts a set of content and link features and compare features in terms of classification performance by SVM classifier. In their work, blog pages are considered as web pages. As examples of web spam detection using link analysis, [4,3] detect web spams by propagating the trust score from a good core of web pages to the rest web pages, following the citation links of web pages. It classifies a webpage as spam by estimating the *spam mass*—the amount of PageRank score contributed by other spam pages. In [9] they detect the web (link) spams using temporal link information extracted from two snapshots of link graphs.

Blogs have unique features. Unlike web spam where the content is usually static, a splog needs to have fresh content to drive traffic to it continuously and often the content is generated by an automatic framework. Therefore, extracting and using temporal dynamics is critical to detecting splogs. Relying purely on content features is not sufficient because spammers copy content from normal blogs to avoid detection. Trust propagation will work poorly due to the editable nature of the blog – a spammer can easily create links to point to the splog. Finally due to the blog temporal dynamics, we cannot rely on snapshots alone—the content and link creation mechanisms used by blog spammers are different from web spam. The changing behavior of splogs is more evasive than that of web spam and cannot be easily captured by a set of snapshots.

We have developed new technique for detecting splogs, based on the observation that a blog is a dynamic, growing sequence of entries (or posts) rather than a collection of individual pages. In our approach, splogs are recognized by their temporal and link properties observed in the post sequence, including: (a) temporal content regularity (self-similarity of content), (b) temporal structural regularity (regular post times), and (c) regularity in the linking structure (frequent links to non-authoritative websites).

We extract a content based feature vector from different parts of the blog – URL's, post content, etc. The dimensionality of the feature vector is reduced by Fisher linear discriminant analysis. We define a novel autocorrelation function on blog post data, and temporal content regularity feature obtained by the autocorrelation vector. The temporal structural regularity is computed using the entropy of the post time difference distribution. We use a variant of the HITS algorithm [5] to compute normalized hub scores that serve as the link regularity measure. We develop an SVM based splog detector using all three features. We have tested our approach on real world datasets with excellent results.

The rest of this paper is organized as follows. In the next section, we provide background to splogs. Then in sections 3 we present our regularity-based features extracted from content, post time and linking patterns. In section 4 we discuss the experimental results and we finally present our conclusions.

2 WHAT ARE SPLOGS?

The motive for creating a splog is solely for driving visitors to affiliated sites that have some *profitable mechanisms*, such as *Google AdSense* or pay-per-click (ppc) affiliate programs [1]. Spammers increase splog visibility by getting indexed with high rank on popular search engines, which usually involves schemes such as non-sense keyword stuffing or content duplication. As blogs have become increasingly mainstream, the presence of splogs has a detrimental effect in the blogosphere [6,10].

Temporal and link structures in splogs. In a typical splog, the content is usually algorithmically generated with specific commercial intent. As a consequence, splogs tend to have repetitive patterns in the post sequence, such as identical posting times, post content, and links to affiliated websites. As human blogs rarely have such repetitive patterns, they can be used as splog indicators. We develop an approach that captures the repetitive splog structural properties.

3 REGULARITY-BASED DETECTION

We have developed new techniques for splog detection based on temporal and linking patterns. We expect that splogs can be recognized by their stable temporal and link patterns observed in entry sequences. In a splog, the content and link structures are typically algorithmically generated (possibly copied from other blogs / websites). The link structure is focused on driving traffic to a specific set of affiliate websites. In the next section we present extraction of baseline content features. In section 3.2, we introduce novel features capturing the temporal regularity, and in section 3.3, we present features based on link structure.

3.1 Content-based features

We shall now discuss the content based features used in this work – these will serve as the baseline feature set as they are widely used in spam detection. We use a subset of the content features presented in [8]. These features are used to distinguish between two classes of blogs – normal and splogs, based on the statistical properties of the content.

We first extract features from five different parts of a blog: (1) tokenized URLs, (2) blog and post titles, (3) anchor text, (4) blog homepage content and (5) post content. For each category we extract the following features: word count (w_c), average word length (w_l) and a tf-idf vector representing the weighted word frequency distribution (w_f). In this work, each content category is analyzed separately from the rest for computational efficiency.

3.1.1 Fisher Linear Discriminant Analysis (LDA)

We need to reduce the length of the vector w_f as the total number of unique terms (excluding words containing digits) is greater than 100,000 (this varies per category, and includes non-traditional usage such as "helloooo"). This can easily lead to over fitting the data. Secondly, the distribution of the words is long-tailed – i.e. most of the words are rarely used.

We expect good feature subsets contain features highly correlated with the class, but uncorrelated with each other. The objective of Fisher LDA is to determine discriminative features while preserving as much of the class discrimination as possible. The solution is to compute the optimal transformation of the feature space based on a criterion that minimizes the within-class scatter (of the data set) and maximizes the between-class scatter simultaneously. This criterion can also be used as a separability measure for feature selection. We use the trace criteria, $J = tr(S_w^{-1} S_b)$ where S_w denotes the within-class scatter and S_b denotes the between-class scatter matrix. This criterion computes the ratio of between-class variance to the within-class variance in terms of the trace of the product (the trace is just the sum of eigenvalues of $S_w^{-1} S_b$). We select the top k eigenvalues to determine the key dimensions of the w_f vector.

3.2 Temporal regularity features

Temporal regularity captures *similarity* between content (content regularity) and consistency in *timing* of content creation (structural regularity).

Content regularity is given by the autocorrelation of the content, derived from computing a similarity measure on the baseline content feature vectors. We define a similarity measure based on the histogram intersection distance. *Structural regularity* is given by the entropy of the post time difference distribution. A splog will have low entropy, indicating algorithmically generated content.

3.2.1 Temporal Content Regularity (TCR)

We define a generalized autocorrelation on non-numeric data (blog post content) to estimate the TCR value. Intuitively, the autocorrelation function (conventionally depicted as $R(\tau)$) of a time series (on numeric data) is an estimate of how a future sample is dependent on a current sample. A noise like signal will have a sharp auto-correlation function, while a highly coherent signal's autocorrelation function will fall off gradually. Since splogs are usually financially motivated, we conjecture that their content will be highly similar over time. However human bloggers will tend to post over a diverse set of topics, leading to a low auto-correlation value.

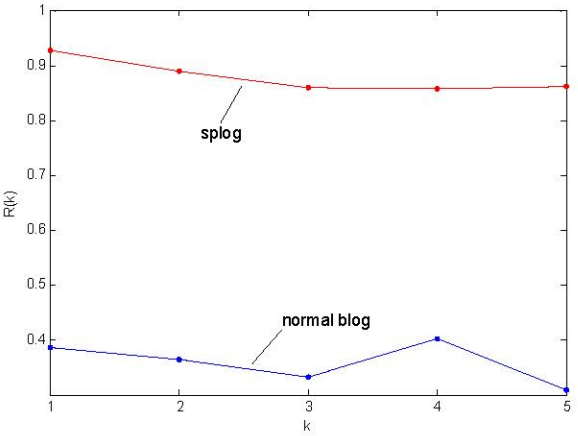

Figure 1: The figure shows the difference in the autocorrelation function between a splog and a normal blog. Notice that auto-correlation function for a splog is very high and nearly constant, while the values for a normal blog are relatively low and fluctuate. These graphs have been derived from a splog and a normal blog from the TREC dataset.

We now present the calculation of the generalized discrete time autocorrelation function $R(k)$ for blog posts. The posts are time difference normalized – i.e. we are only interested in the similarity between the current post and a future post in terms of the number of posts in between (e.g. will the post after next be related to the current post), ignoring time. This is a simplifying assumption, but is useful because many posts do not have the time meta data associated with them. The autocorrelation function is then defined as follows:

$$R(k) = 1 - d(p(l), p(l+k)),$$

$$d(p(l), p(l+k)) \triangleq 1 - E\left(\frac{\sum_{i}^{N} \min\left(w^{l}{}_{f}(i), w^{l+k}{}_{f}(i) \right)}{\sum_{i}^{N} \max\left(w^{l}{}_{f}(i), w^{l+k}{}_{f}(i) \right)} \right), \quad <1>$$

where E is the expectation operator, $R(k)$ is the expected value of the autocorrelation between the current l^{th} post and the $(l+k)^{th}$ post; d is the dissimilarity measure, $w^{l}{}_{f}(i)$ refers to the i^{th} dimension of the tf-idf vector w_f of the l^{th} post. The estimation of TCR for a blog b_i , denoted as $TCR(b_i)$, is obtained by the m-dimensional vector, i.e. $TCR(b_i)=R(k)$ for $k=1,2,\ldots,m$.

3.2.2 Temporal Structural Regularity (TSR)

We estimate TSR of a blog by computing the entropy of the post time difference distribution. In order to estimate the distribution,

we use hierarchical clustering with single link merge criteria on the post interval difference values

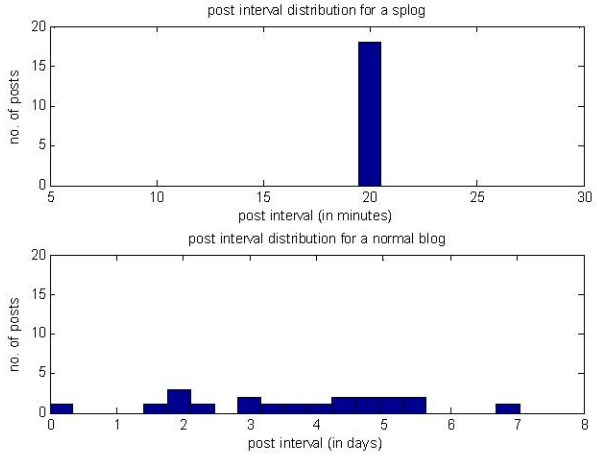

Figure 2: The figure shows the differences in the temporal regularity between a splog and a normal blog. The splog posts with a post time of every 20 minutes, while the normal blog has post time interval that is highly varied and ranges to several hours to nearly a week.

The original dataset is initialized into N clusters for N data points. Two clusters are merged into one if the distance (linkage) between the two is the smallest amongst all pair wise cluster distances. We use the average variance stopping criteria for clustering. Once the clusters are determined, we compute cluster entropy as a measure of TSR for each blog b_i,:

$$B_e = -\sum_{i=1}^{M} p_i \log_M p_i,\ p_i = \frac{n_i}{N}, \qquad <2>$$

and $TSR(b_i)=1-B_e/B_{max}$, where B_e is the blog entropy for blog b_i, B_{max} is the maximum observed entropy over all blogs, N is the total number of posts in the blog, n_i and p_i are the number of posts and the probability of the i^{th} cluster respectively, and M is the number of clusters. Note that for some blogs including normal or splogs, post time is not available as part of the post metadata. We treat such cases as missing data—if a blog does not have post time information, its TSR feature is ignored.

3.3 Link regularity estimation

Link regularity (LR) measures website linking consistency for a blogger. We expect that a splog will exhibit more consistent linking behavior since the main intent of such splogs is to drive traffic to affiliate websites. Secondly, we conjecture that there will be a significant portion of links that will be targeted to affiliated websites rather than normal blogs / websites. Importantly these affiliate websites will *not* be authoritative and we do not expect normal bloggers to link to such websites.

We analyze the splog linking behavior using the well known HITS algorithm [5]. The intuition is that splogs target focused set of websites, while normal blogs usually have more diverse targeting websites. We use HITS with out-link normalization to compute hub scores. The normalized hub score for a blog is a useful indicator of the blog being a splog.

In the HITS algorithm, each website w_i has both a hub score h_i and an authority score a_i. Given a web graph derived from M websites, good hubs and good authorities are identified by the following mutually reinforcing relationship: $a=A^T h$, $h=Aa$, where a is the vector containing authority scores, h is the vector containing hub scores, and A is the adjacency matrix with each element $A_{ij}=1$ indicating a hyperlink from website w_i to w_j.

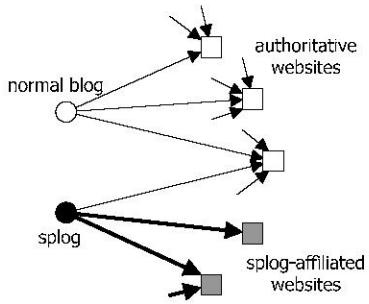

Figure 3: Normal blogs tend to link to authoritative websites while splogs frequently link to affiliate websites that are not authorities. The thickness of the arrow from the splog indicates the frequency with which the splog links to a non-authoritative, affiliate website.

A blog graph is a subgraph of the web graph. We construct an $N\times M$ adjacency matrix B from a directed bi-partite graph derived from N blogs and M websites, where $B_{ij}=1$ indicates there is a hyperlink from blog b_i to website w_j. Each blog is assigned a hub score and each website is assigned an authority score. The vector $h^{(b)}=(h_1,\ldots,h_N)^T$ contains the blog hub scores and the vector $a^{(w)}=(a_1,\ldots,a_M)^T$ contains the website authority scores. We compute the normalized blog hub score as a measure of LR:

$$h^{(b)} = B^{(out)} a^{(w)}, a^{(w)} = (B^{(out)})^T h^{(b)} \qquad <3>$$

and $LR(b_i)= h_i^{(b)}$, where matrix $B^{(out)}$ is the out-degree normalized adjacency matrix obtained by $B^{(out)}=D^{-1}B$, where $D=\text{diag}((o_i,\ldots,o_N)^T)$, and $o_i=\sum_j B_{ij}$ is the out-degree of blog b_i.

The out-degree normalization plays an important role in the LR estimation. Without the normalization, a splog that has considerable links pointing to the affiliated websites, can easily boost its hub score by creating just a few hyperlinks pointing to the authoritative websites. To suppress this effect, we normalize the blog out-degrees so that each linked websites of a blog is equally important, in terms of their contribution of the blog hub score. It means every selection of the linked websites is critical. In order to obtain higher hub score, a blog needs to have links to many different authoritative websites. Hence the normalized hub score of a blog can serve as a useful indicator differentiating splogs form normal blogs.

Classification. Our splog detector combines these new features (TCR, TSR, LR) with traditional content features into a large feature vector. We then use standard machine learning techniques (SVM classifier with a polynomial kernel) to classify each blog into two classes: splog or normal blog.

4 EXPERIMENTS

In this section, we present preliminary experimental results on the splog detection.

Data set description. We use the TREC (the Text Retrieval Conference) Blog-Track 2006 dataset for analysis. This dataset is a crawl of 100,649 feeds collected over 11 weeks, from Dec. 6, 2005 to Feb. 21, 2006, totaling 77 days. After removing duplicate feeds and feeds without homepage or permalinks (entries), we have about 43.6K unique blogs. We focus our analysis on this subset of blogs having homepage and at least one entry. In the TREC dataset, we manually labeled 8K normal blogs and 800+ splogs. We decided to create a symmetric set for evaluation containing 800 splogs and 800 normal blogs.

4.1 Detection performance

We use standard machine learning techniques (SVM classifier implemented using libsvm package [2], with a polynomial kernel) to classify each blog into two classes: splog or normal blog. A five fold cross-validation technique is used to evaluate the performance of the splog detector. Our performance measures include the following: AUC (area under the ROC curve), accuracy, precision and recall.

The results in Table 1 are interesting. They reveal that by *using the regularity-based features* we get significant improvement. The regularity-based features, designated as R in Table 1, are constructed as a 7-dimensional feature vector using TCR, TSR and LR measures. The baseline content features, designed as base-*n*, are *n*-dimensional feature vectors constructed using the content-based analysis alone. The baseline and non-baseline features jointly work very well – in each case when the baseline content features are merged with the regularity-based features, designed as R+base-*n*, the performance improves over using content alone. The performance gain by using the regularity-based features is promising—the size of regularity-based features is relatively small, compared to the large size content features; however, the improvement is significant, especially in low-dimensional feature vector cases.

Table 1: The table shows a comparison of the baseline content scheme (base-*n*, where *n* is the dimension of the baseline feature) against the regularity-based features (designated as R), and the combination of baseline with the regularity features (designated as R+base-*n*). The table indicates a significant improvement due to the regularity-based features is smaller with increase in the number of dimensions to the baseline features.

Feature	AUC	accuracy	precision	recall
base-256	0.951	0.873	0.861	0.889
R+base-256	0.971	0.918	0.915	0.923
base-128	0.924	0.859	0.864	0.853
R+base-128	0.954	0.896	0.889	0.905
base-64	0.896	0.820	0.817	0.825
R+base-64	0.950	0.883	0.870	0.901
base-32	0.863	0.795	0.810	0.771
R+base-32	0.937	0.871	0.852	0.898
base-16	0.837	0.711	0.662	0.861
R+base-16	0.922	0.856	0.832	0.893
R	0.807	0.753	0.722	0.821

The promising results indicate that the temporal and link structures of a blog, is a key distinguishing characteristic and in this specific application, allows us to distinguish between splogs and normal blogs.

5 CONCLUSIONS

In this paper, we propose new framework to detect splogs in the blogosphere. In our approach, splogs are recognized by their content, temporal and link structures. A key rationale for exploiting the structural property in splog detection is in the robustness of these features. While blog content is highly dynamic and vary over time, the temporal and link structures captured by the regularity based features reveal a stable blog character.

The unique structural properties of splogs are captured by three types of regularity-based features, including (a) TCR: temporal content regularity (self-similarity of content), (b) TSR: temporal structural regularity (regular post times), and (c) LR: regularity in the linking structure (frequent links to non-authoritative websites). We have evaluated our approach using standard classification performance metrics. The experimental results are promising, indicating the regularity-based features work well in the splog detection application. As part of future work, we plan to develop more sophisticated mathematical representations of blog structural properties.

6 REFERENCES

[1] *Wikipedia, Spam blog* http://en.wikipedia.org/wiki/Splog.

[2] C.-C. CHANG and C.-J. LIN (2001). *LIBSVM: a library for support vector machines.*

[3] Z. GYÖNGYI, H. GARCIA-MOLINA and J. PEDERSEN (2004). *Combating web spam with TrustRank*, Proceedings of the 30th International Conference on Very Large Data Bases (VLDB) 2004, Toronto, Canada,

[4] Z. GYÖNGYI, P. BERKHIN, HECTOR GARCIA-MOLINA and J. PEDERSEN (2006). *Link Spam Detection Based on Mass Estimation*, 32nd International Conference on Very Large Data Bases (VLDB), Seoul, Korea,

[5] J. M. KLEINBERG (1999). *Authoritative sources in a hyperlinked environment.* J. ACM **46**(5): 604-632.

[6] P. KOLARI (2005) *Welcome to the Splogosphere: 75% of new pings are spings (splogs)* permalink: http://ebiquity.umbc.edu/blogger/2005/12/15/welcome-to-the-splogosphere-75-of-new-blog-posts-are-spam/.

[7] P. KOLARI, A. JAVA, T. FININ, T. OATES and A. JOSHI (2006). *Detecting Spam Blogs: A Machine Learning Approach*, Proceedings of the 21st National Conference on Artificial Intelligence (AAAI 2006), Boston, MA, July 2006.

[8] A. NTOULAS, M. NAJORK, M. MANASSE and D. FETTERLY (2006). *Detecting spam web pages through content analysis*, Proceedings of the 15th international conference on World Wide Web, Edinburgh, Scotland, May 2006.

[9] G. SHEN, B. GAO, T.-Y. LIU, G. FENG, S. SONG and H. LI (2006). *Detecting Link Spam using Temporal Information*, Proc. of ICDM-2006, to appear, 2006,

[10] UMBRIA (2006) *SPAM in the blogosphere* http://www.umbrialistens.com/files/uploads/umbria_splog.pdf.