Splog Detection Using Self-similarity Analysis on Blog Temporal Dynamics

Yu-Ru Lin Hari Sundaram Yun Chi Junichi Tatemura Belle L. Tseng

Arts Media and Engineering Program
Arizona State University

NEC Laboratories America
Cupertino, CA 95014

e-mail: {yu-ru.lin, hari.sundaram}@asu.edu, {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. The presence of splogs degrades blog search results as well as wastes network resources. In our approach we exploit unique blog temporal dynamics to detect splogs.

There are three key ideas in our splog detection framework. We first represent the blog temporal dynamics using self-similarity matrices defined on the histogram intersection similarity measure of the time, content, and link attributes of posts. Second, we show via a novel visualization that the blog temporal characteristics reveal attribute correlation, depending on type of the blog (normal blogs and splogs). Third, we propose the use of temporal structural properties computed from self-similarity matrices across different attributes. In a splog detector, these novel features are combined with content based features. We extract a content based feature vector from different parts of the blog – URLs, post content, etc. The dimensionality of the feature vector is reduced by Fisher linear discriminant analysis. We have tested an SVM based splog detector using proposed features on real world datasets, with excellent results (90% accuracy).

Categories and Subject Descriptors

H.3.3 [**Information Systems**]: *Information Search and Retrieval*; H.3.5 [**Information Systems**]: *Online Information Services*; H.5.4 [**Information Systems**]: *Hypertext/Hypermedia*

General Terms

Experimentation, Measurement, Algorithms, Human Factors.

Keywords

Blogs, temporal dynamics, regularity, spam, splog detection, topology, self-similarity

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 new media for communication, and the presence of splogs degrades blog search results as well as wastes network resources.

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 characteristics and content. There are three key ideas in our splog detection framework.

1. We represent the blog temporal dynamics using self-similarity matrices defined on the histogram intersection similarity measure of the time, content, and link attributes of posts. The self-similarity matrices function as a generalized spectral analysis tool. It allows investigation of the temporal changes within the post sequence.
2. We study the blog temporal characteristics based on a visual transformation derived from the self-similarity measures. We show that the blog temporal characteristics reveal correlation between attributes, depending on type of the blog (normal blogs and splogs).
3. We propose two types of novel temporal features to capture the splog temporal characteristics: (1) Regularity features are computed along the off-diagonals and from the coherent blocks of the self-similarity matrices on single attribute. (2) Joint features are computed from self-similarity matrices across different attributes.

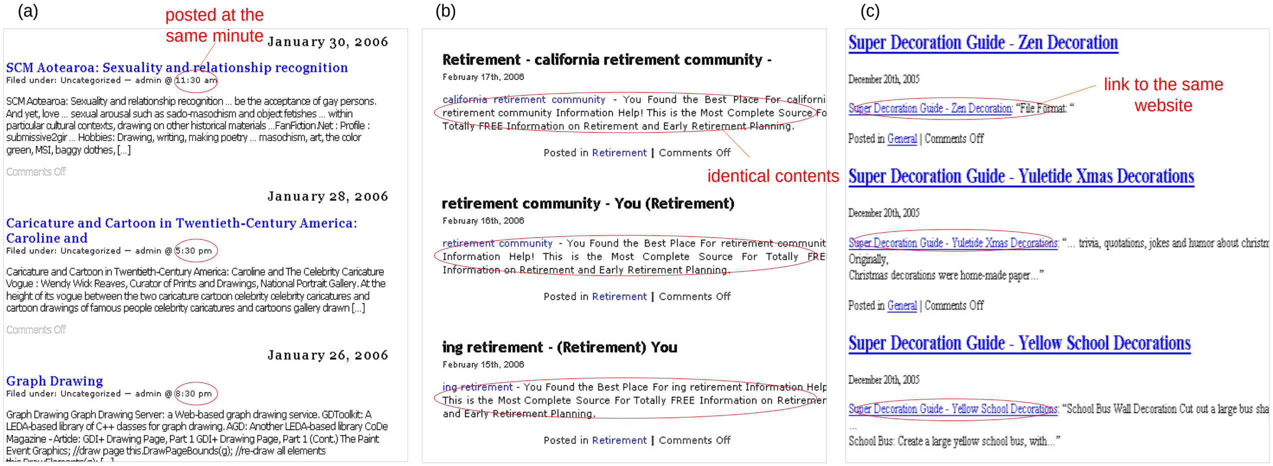
In a splog detector, these novel features are combined with content based features. We extract a content based feature vector from different parts of the blog – URLs, post content, etc. The dimensionality of the feature vector is reduced by Fisher linear discriminant analysis. We develop an SVM based splog detector using proposed features and have tested our approach on real world datasets with excellent results.

The rest of the paper is organized as follows. In next two sections we discuss related work and examine the splog phenomena. In section 4, we examine the self-similar aspects. In section 5 and 6, we present our approach of characterizing blog temporal dynamics based on self-similarity analysis. We present experimental results in section 7. Finally we present our conclusion and discuss future work in section 8.

2. RELATED WORK

Splogs are relatively new phenomena; however there has been work on web spam detection. While there are critical differences between the two, a review of web-spam research provides useful insight. Prior work to detect web spams can be categorized into content and link analysis. In [11] the authors distinguish web spams from normal web pages based on statistical properties in

AIRWeb 2007, May 8, 2007. Banff, Alberta, Canada.
Copyright 2007 ACM 978-1-59593-732-2 ...$5.00.



January 30, 2006

SCM Aotearoa: Sexuality and relationship recognition
Filed under: Uncategorized — admin @ 11:30 am

SCM Aotearoa: Sexuality and relationship recognition ... be the acceptance of gay persons. And yet, love ... sexual arousal such as sado-masochism and object fetishes ... within particular cultural contexts, drawing on other historical materials ...FanFiction.Net : Profile : submissive2gir ... Hobbies: Drawing, writing, making poetry ... masochism, art, the color green, MSI, baggy clothes, [...]

Comments Off

January 28, 2006

Caricature and Cartoon in Twentieth-Century America: Caroline and
Filed under: Uncategorized — admin @ 5:30 pm

Caricature and Cartoon in Twentieth-Century America: Caroline and The Celebrity Caricature Vogue : Wendy Wick Reaves, Curator of Prints and Drawings, National Portrait Gallery. At the height of its vogue between the two caricature cartoon celebrity celebrity caricatures and cartoon drawings of famous people celebrity caricatures and cartoons gallery drawn [...]

Comments Off

January 26, 2006

Graph Drawing
Filed under: Uncategorized — admin @ 8:30 pm

Graph Drawing Graph Drawing Server: a Web-based graph drawing service. GDToolkit: A LEDA-based library of C++ classes for graph drawing. AGD: Another LEDA-based library CoDe Magazine - Article: GDI+ Drawing Page, Part 1 GDI+ Drawing Page, Part 1 (Cont.) The Paint Event Graphics; //draw page this.DrawPageBounds(g); //re-draw all elements

Retirement - california retirement community -
February 17th, 2006

california retirement community - You Found the Best Place For californi retirement community Information Help! This is the Most Complete Source Fo Totally FREE Information on Retirement and Early Retirement Planning.

Posted in Retirement | Comments Off

retirement community - You (Retirement)
February 16th, 2006

retirement community - You Found the Best Place For retirement communit Information Help! This is the Most Complete Source For Totally FRE Information on Retirement and Early Retirement Planning.

Posted in Retirement | Comments Off

ing retirement - (Retirement) You
February 15th, 2006

ing retirement - You Found the Best Place For ing retirement Information Help This is the Most Complete Source For Totally FREE Information on Retiremer and Early Retirement Planning.

Posted in Retirement | Comments Off

Super Decoration Guide - Zen Decoration

December 20th, 2005

Super Decoration Guide - Zen Decoration: "File Format: "

Posted in General | Comments Off

Super Decoration Guide - Yuletide Xmas Decorations

December 20th, 2005

Super Decoration Guide - Yuletide Xmas Decorations: "... trivia, quotations, jokes and humor about christm Originally,
Christmas decorations were home-made paper..."

Posted in General | Comments Off

Super Decoration Guide - Yellow School Decorations

December 20th, 2005

Super Decoration Guide - Yellow School Decorations: "School Bus Wall Decoration Cut out a large bus sha
...
School Bus: Create a large yellow school bus, with..."

Figure 1: Examples of repetitive patterns in posting times, post contents, and affiliated links in splogs.

the content, such as number of words, average word length, or term frequencies in title, anchor text, tokenized URLs, etc. As examples of web spam detection using link analysis, [6,5] 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. They classify a webpage as spam by estimating the *spam mass* – the amount of PageRank score contributes by other spam pages. In [12] they detect the web (link) spams using temporal link information extracted from two snapshots of link graphs.

Splog has been considered as a special case of web spam pages [8,10]. The characteristics of splogs is investigated in [9]. In order to combat splogs, in [8,10] the authors suggest using a set of content and link features and compare features in terms of classification performance by SVM classifier. In their work, each blog is treated as a single and static web page.

Blogs have unique features. Unlike web spam where the content is usually static, a splog needs to have fresh content in order to continuously drive traffic and often the content is generated by an automated framework. Therefore, extracting and using temporal dynamics is critical to detecting splogs. Relying only on content features is not sufficient because spammers copy content from normal blogs. 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.

Data analysis based on "recurrence plots" (RP) was introduced by Eckmann et al. [3] to visualize how a dynamic system comes back to a state similar to a former state after some time. The idea of RP can be generalized into a self-similarity matrix. In multimedia, there has been prior work on temporal analysis of audio and/or visual signals. Foote et al. [4] propose self-similarities as a way to visualize musical structures. Multimedia objects are usually uniformly sampled time-series data. In this work we generalize this approach to arbitrary non-numeric time-series data (blog posts) as opposed to analyzing continuous auditory data.

3. WHAT ARE SPLOGS?

The motive for creating a splog is solely to drive 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 keyword stuffing or content duplication. As blogs have become increasingly mainstream, the presence of splogs has a detrimental effect on the blogosphere.

There are some alarming statistics about splogs. The authors in [13] reported that for the week of Oct. 24, 2005, 2.7 million blogs out of 20.3 million (10-20%) were splogs, and that an average of 44 of the top 100 blogs search results in the three popular blog search engines came from splogs. It has been estimated that 75% of new pings came from splogs; more than 50% of claimed blogs pinging the website weblogs.com are splogs [7]. The statistics reveal that splogs can cause problems including (1) the degradation of information retrieval quality and (2) the significant waste of network and storage resources.

Typical characteristics. We refer to a blog created by an author who has the intention of spamming to be a splog. Note that a splog is evaluated at the blog level, not individual pages (single homepage or permalink pages). Additionally, a blog that contains spam in the form of comment spam or trackback spam is not considered to be a splog. We now list some typical characteristics:

1. *Machine-generated content*: splog posts are generated algorithmically. Some might contain nonsense, gibberish, and repetitive text while others might copy or weave with text from other blogs or websites.
2. *No value-addition*: splogs provide useless or no unique information. Note that there are blogs using automatic content aggregation techniques to provide useful service, e.g. daily tips, product reviews, etc. – although the content is gathered from other blogs or news sources, we consider these blogs as legitimate blogs because of their value addition.
3. *Hidden agenda, usually an economic goal*: splogs have commercial intention – they display affiliate ads or out-going links to affiliate sites.

These characteristics can also be found in other types of spams like web spam. However, splogs have additional unique properties:

Uniqueness of splogs. Splogs are different from generic web spams in the following aspects.

1. *Dynamic content*: blog readers are mostly interested in recent posts. Unlike web spams where the content is static, a splog continuously generates fresh content to attract traffic.

2. *Non-endorsement link*: In web pages, a hyperlink is often interpreted as an endorsement of other pages. However, since most blogs have editable area open for readers to contribute, spammers are able to create hyperlinks (comment links or trackbacks) in normal blogs, links in blogs cannot be simply treated as endorsements.

Temporal and link structures in splogs. Because of algorithmically generated content, splogs tend to have repetitive patterns in the post sequence. Figure 1 shows three example splogs that have identical posting times (during the day), post content, or links to affiliated websites appearing in their posts.

In comparison, normal (human) bloggers show variation in blog content and post time, and have a diverse set of outgoing links. We shall develop an approach that captures the different structural properties between normal blogs and splogs. In the following sections we show how self-similarity analysis of a blog is useful for distinguishing splogs from normal blogs.

4. SELF-SIMILARITY

We now analyze the temporal dynamics of blog by examining the temporal self-similarity of its prime attributes, such as the post content, citation links, tags, etc. The intuition behind using self-similarity lies in its ability to reveal temporal structures.

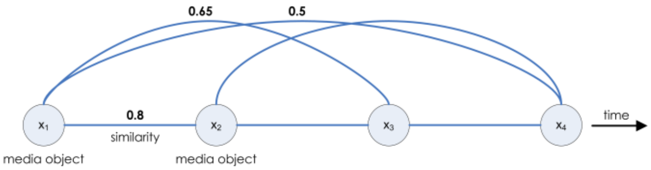


Figure 2: A topological graph induced on time sequence of media objects (e.g. blog posts), due to a specific similarity measure $s(i,j;\,\alpha)$ on the attribute α.

Let us assume that we have a sequence of N media objects (e.g. blog posts) x_i, from the same blog that are ordered in time – i.e. $t(x_i) \leq t(x_{i+m})$ for all $m \geq 0$. Assume that the media objects can be described using a set of attributes and for each attribute α, we define an attribute-dependent similarity measure $s(i,i+m;\,\alpha)$. This measures the similarity between any pair of objects x_i and x_{i+m}, using attribute α. We further create an attribute-dependent topological matrix S_α using the topology where the elements of the matrix $S_\alpha(i,j)$ are defined as follows: $S_\alpha(i,j) = s(i,j;\,\alpha)$, for $i,j \in \{1,\ldots,\ N\}$. Since the media objects are ordered in time, the topological matrix reveals the temporal self-similarity of the sequence x_i with respect to attribute α. In the following discussion, we shall refer to S_α as a self-similarity matrix on a particular attribute α. Figure 2 shows an example topological graph.

The self-similarity matrix functions as a generalized autocorrelation over any time series of media objects – if we take the average over all the values along each diagonal in the upper triangular matrix, this would be equivalent to computing the autocorrelation of the non-numeric sequence. Note that the nodes in Figure 2 refer to *posts from the same blog* and the edges refer to the similarity between posts on a specific attribute. We now examine the self-similarity matrices computed on the temporal sequences of blog posts. We focus on three prime attributes: post time, content and links.

4.1 Post time

A blog post usually contains a time stamp indicating when the post is created. We first examine the post timing using self-similarity on the post time attribute. We use two different time scales – (a) at the micro-time scale (in daily time) and (b) at the macro-time scale (in absolute time). The similarity measures are defined as follows:

$$\begin{aligned} S_{micro}(i,j) &= \mathrm{mod}\left(|\,t_i - t_j\,|, \delta_{day}\right), \\ S_{macro}(i,j) &= |\,t_i - t_j\,|, \end{aligned} \qquad <1>$$

where t_i and t_j are the post time of post p_i and p_j, respectively, and δ_{day} is time of a day (e.g. if the post time is expressed in seconds, δ_{day}=24×60×60=86400). The micro time similarity reveals the relationships between posts that may be days apart, but were posted at a similar time. It indicates regularity in posting time – perhaps some bloggers only post at specific times during the day. The macro time analysis is able to reveal post discontinues at large time scales, which might due to vacations, etc. Ref. Figure 3 for an example that demonstrates micro and macro time structures in blog post times. Note that along any row i, starting from the main diagonal, we have the similarity between post p_i and future posts. The white lines along the off-diagonals in the micro-time matrix suggest that the post creation time is similar in the daily time scale at different posts, and the white blocks in the macro-time matrix suggest the creation time of successive posts is close in absolute time.

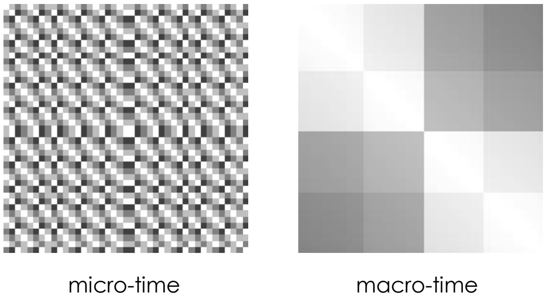


Figure 3: The plots show the micro- (in daily time) / macro-time structure in the posts times of the blogs. The brightness value is proportional to similarity in both images – the highest similarity is scaled to have the largest brightness value.

4.2 Post content

The similarity measure on post content is defined using histogram intersection on the tf-idf vectors. Let h_i and h_j be the tf-idf vectors (after stemming and stop-word removal) for posts p_j and p_j. The similarity between any two posts is defined as:

$$S_c(i,j) = \frac{\sum_{k=1}^{M} \min\left(h_i(k), h_j(k)\right)}{\sum_{k=1}^{M} \max\left(h_i(k), h_j(k)\right)} \qquad <2>$$

where, c refers to the similarity based on the content attribute and M is the size of the vector. Note that the posts are ordered in time. The corresponding element in self-similarity matrix is then the content similarity between the two posts.

Figure 4 (a) shows an example plot of the content-based temporal self-similarity matrix. It reveals that there is significant temporal correlation between posts. It also suggests that the blogger mostly tends to stay on a topic (large white blocks in the figure), and may occasionally post on different topics (causing black horizontal lines) before returning to the topics.

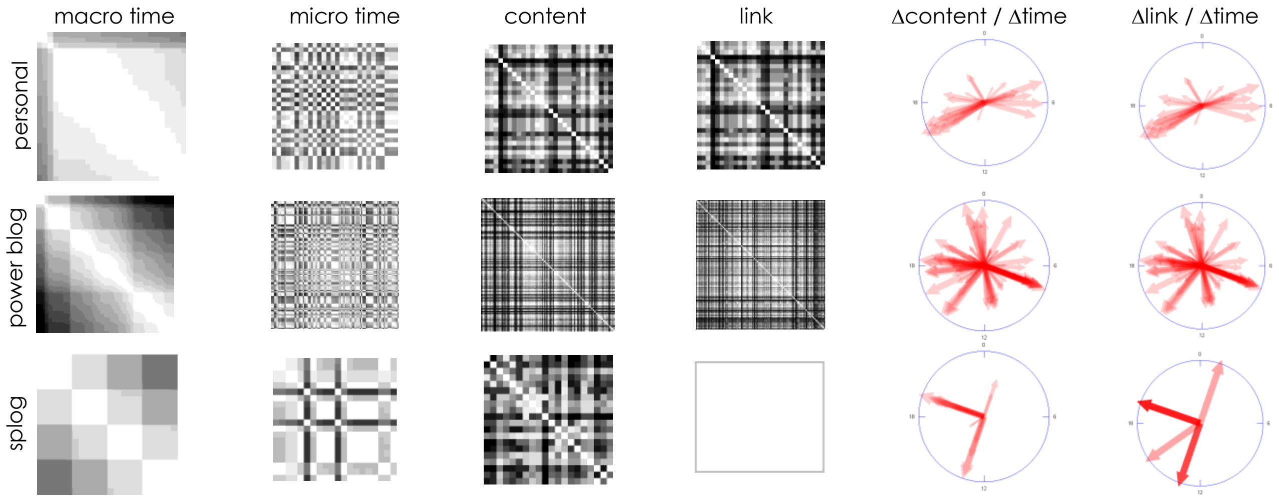


Figure 5: The figure shows the self similarity of the content, link and post-times, as well as self-similar clocks on content and link with respect to post time. Three blogs – a normal personal journal, a power blogger and a splog are shown as examples to demonstrate the proposed visualization framework allows distinguishing the temporal dynamics of different types of blogs.

4.3 Post Links

The similarity measure on the links is defined in a similar manner to eq. <2> except that the tf-idf vectors are now calculated on the target links (stemmed by the root domain), rather than on the words. Hence:

$$S_l(i,j)=\frac{\sum_{k=1}^{M}\min\left(h_i(k),h_j(k)\right)}{\sum_{k=1}^{M}\max\left(h_i(k),h_j(k)\right)} \qquad \text{<3>}$$

where, l refers to the similarity on the link attribute, h_i and h_j are the link based tf-idf vectors and M is the size of the vector.

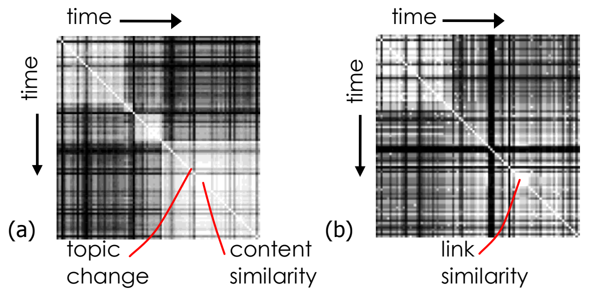


Figure 4: (a) Self-similarity matrix on the content attribute. (b) Self-similarity matrix on the link attribute.

Figure 4 (b) shows the link based self-similarity matrix. It reveals a similar block-based structure and changes as in Figure 4 (a), and we can see that changes to the content and link patterns are usually coincident.

The self-similarity matrices exhibit several advantages: First, the media objects are represented by their relationship with the other objects. The size of the matrix used to represent a dataset of N objects is always $N\times N$, regardless of the content complexity of the objects. Second, because the self-similarity matrices are constructed according to the time order of objects, they allow investigation of the temporal relationship among objects.

5. BLOG TEMPORAL SIGNATURES

In this section we present a visualization framework to demonstrate that splogs have distinct temporal dynamics compared to normal blogs. We use self-similarity analysis in the visualization to distinguish amongst different blog types.

In order to examine the temporal dynamics of blogs, we represent the similarity relationship in blog temporal sequence using a “clock” metaphor. The idea is to show how the non-time attributes (content or links) change with respect to the time attribute. Here we show how the attribute values of content and link change over the micro-time attribute. This approach can be applied to macro-time attribute to examine the long-term change of other attributes. As shown in Figure 6, we render each post as an arrow stemmed from the center of a circle. The length of an arrow indicates how the post is similar to its previous post in terms of the corresponding non-time attribute value, and the orientation of an arrow indicates when the post is created in a day.

Let $\rho_{i,\alpha}$ be the length of the i^{th} post-arrow corresponding to the attribute α, and θ_i be the orientation of the i^{th} post-arrow. We then compute $\rho_{i\,\alpha}$ and θ_i, for $i \geq 2$, as follows:

$$\begin{aligned}\rho_{i,\alpha} &= \rho_{i-1,\alpha}+1-\log_2\left(1+S_\alpha\left(i,i-1\right)\right),\\ \theta_i &= \theta_{i-1}-2\pi S_{micro}\left(i,i-1\right)\end{aligned} \qquad \text{<4>}$$

Note that $\rho_{1,\alpha}\equiv 1$ since p_0 is not defined. The equation indicates the length of arrow ($\rho_{i,\alpha}$) is determined by the attribute similarity between the corresponding post and its previous one, and the angle of arrow (θ_i) represents the time of the post in the day.

These transforms reveal the blog temporal dynamics by showing the rate of change of specific attribute with respect to time. In Figure 6, the increase in arrow density suggests regularity in the post time. Spiral like growth in the length of the arrow suggests a slow change in the content.

We can see distinct temporal characteristics amongst normal blogs and splogs. A comparison of different blog types, including two normal blogs, personal and power blog, and one splog case, is shown in Figure 5. Personal blogs are used as an online personal journal. Power blogs are those focus on a specific niche. They are often used to constantly update readers on the status of an on-going project, to provide a review on a new service or product, or to promote businesses of a company.

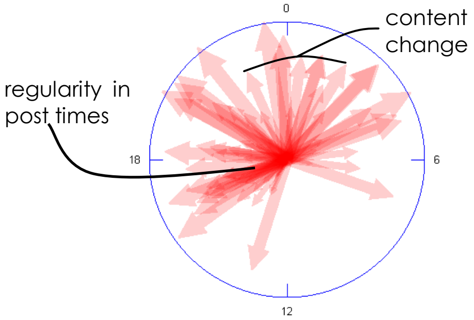


Figure 6: This plot shows the change of post content over its daily post time.

The figure shows very interesting differences and the following observations have been empirically observed to hold true across other samples. Normal blogs and splogs differ in their temporal characteristics in all three facets – post time, content and link.

Post time: Normal bloggers seem to prefer a regular posting time (e.g. morning / night) and the times show a spread, but are largely in a few time periods – they do not post at all times of the day. A splog will show machine like regularity – this can be either posting at fixed stable times, or a uniform distribution of times (throughout the day). A normal personal blogger as well as a power blog exhibit macro time breaks (e.g. due to vacation), this characteristic is absent in a splog. A power blogger can have *both* highly diverse posting time as well as machine like regularity as in Figure 5 – in this particular case, the blogger, actually has a script that regularly aggregates content from the web / other blogs and posts them on the blog at regular times.

Post content: A normal blogger typically stays on topic for a while, before showing a topic drift. A splog often copies content (to escape detection, as well as to appear as a search result related to a query) from different parts of the web / blogosphere and hence may show very high topic diversity. Some splogs on the other hand exhibit significant topic stability. We observe that a power blogger has a higher rate of change of content with time than a normal personal blogger. Interestingly, for both normal personal and power bloggers, changes to content appear to coincide with macro time breaks.

Post links: For normal, personal and power bloggers, changes to content affect the change to the linked sites in the same manner – this is to be expected as the links essentially function as supporting arguments in a blog post. However, a splog is driven by a strong commercial interest to drive traffic to affiliate sites. In the example in Figure 5, the splog always points to the same site (exhibiting strong link regularity), thus the temporal self-similarity in the figure does not change. Other splogs show a similar characteristic – i.e. they only show limited link temporal diversity.

6. TEMPORAL FEATURES

We now discuss our approach using the self-similarity analysis to derive features useful for splog detection. In section 6.1 we 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. Then in section 6.2 and 6.3, we propose two novel temporal features: the regularity features and the joint features.

6.1 Content based features

Content based features are regarded as useful features in detecting spam web pages [11]. 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.

6.1.1 *Fisher linear discriminant analysis (LDA)*

To avoid data over-fitting, we need to reduce the length of the vector w_f because 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”). Some particular terms might appear only in one or few blogs. 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 to be highly correlated with the class (in our case, normal blog vs. splog), 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 the sum of eigenvalues of $S_w^{-1}S_b$). We select the top k eigenvalues to determine the dimensions of the feature vector. Because the content of splogs is highly dynamic, i.e. the spam terms might change quickly, a large content-based feature vector tends to lose its generalizability. Here, we deliberately select a small k (k=32,…,256) to show how the following proposed temporal features can address this issue.

6.2 Regularity features

In the next two sub-sections we propose features that support our self-similarity characterization of the blog temporal dynamics, based on the self-similarity matrices introduced in section 4.

We shall use the following self-similarity matrices: (1) S_{macro}: macro-time, (2) S_{micro}: micro-time, (3) S_c: content and (4) S_l: link self-similarity matrix.

From the self-similarity matrices, we observe two types of patterns (ref. Figure 4 and the illustration in Figure 7): (1) high intensity off-diagonal lines appear when the attribute values are similar at different posts, and (2) high intensity blocks appear when the attribute values remain highly constant for some period. We conjecture that both patterns exhibit temporal regularity from different aspects, and we shall compute the regularity features from these patterns.

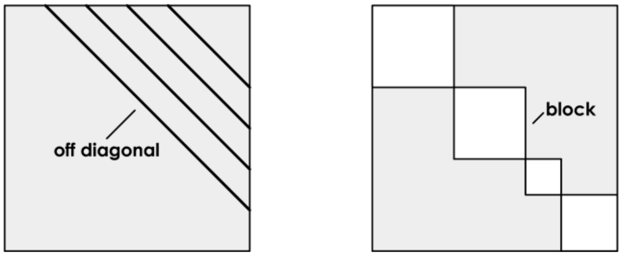


Figure 7: The features are computed using the off diagonals and blocks within the self-similarity matrix.

6.2.1 *Features along the Off-Diagonals*

We use three measures, mean, standard deviation and entropy to quantify the regularity patterns along the off-diagonals. The expectation along the diagonals of the topological matrix is equivalent to the generalized autocorrelation of the time series data under the specific similarity measure. Specifically the expectation along the k^{th} off-diagonal, is a measure of average similarity of a post to another post, with k-1 posts in between.

Intuitively, the autocorrelation function of a numeric time series data is an estimate of how a future sample is dependent on a current sample. A noise like signal will have a sharp autocorrelation function, while a highly coherent signal will fall off gradually. Here, we use the expectation of the off-diagonal values as a generalized autocorrelation on non-numeric blog post data. This captures how the attribute value of post sequence change, and we use the standard deviation to describe how the data deviates from the expectation. Additionally we compute the entropy of the off-diagonal elements to measure the amount of disorder among these elements.

Given a self-similarity matrix $M_\alpha \in \{S_{macro}, S_{micro}, S_c, S_l\}$, we compute the mean ($\mu_k$), standard deviation ($\sigma_k$) and entropy ($H_k$) along the k^{th} off-diagonal, $0 < k \le k_o$ for certain $k_o < N$, where N is the size of M_α. This is formally computed as follows:

$$\begin{aligned} &\mu_k(M_\alpha) = E[z], \\ &\sigma_k(M_\alpha) = \sqrt{\mathrm{var}[z]}, \qquad\qquad <5> \\ &H_k(M_\alpha) = -\sum_{i=1}^{D} p_i \log_D p_i, \end{aligned}$$

where, $z =$ diag(M_α,k) is the k^{th} off-diagonal, and the probabilities $p_i = d_i/D$ are computed after quantizing z into D bins, and d_i is the number of elements of z fall in the i^{th} bin. We typically use k_o=4 diagonals in our feature estimation. This is because as k_o is close to N, there are not enough samples in the data to estimate the variables accurately.

6.2.2 *Features from Coherent Blocks*

The block-wise features measure the similarity among posts within a group of consecutive posts. As blog posts usually exhibit short-term temporal coherence, coherent groups of posts can be easily observed as white blocks along the main diagonal on a self-similarity matrix. To extract such blocks, we only need to segment the main diagonal such that each segment associates with a block. We use an agglomerative hierarchical clustering method, with single link merge criteria on the pair-wise similarity values embedded in the self-similarity matrix. 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 simply use the average variance stopping criteria for clustering. (Other stopping criteria can also be applied here.) Once the clusters are determined, we further split a cluster if its data points (on the main diagonal) are not connected.

Let $b_k = \{\mathrm{M}_{\mathrm{u,v}}\}_{i \le u,v \le i+n-1}$ be a block that contains $n \times n$ connected elements (that induced from n data points) on the matrix. Similar to the diagonal-wise features, we now compute block-wise features – mean ($\mu_{b,k}$), standard deviation ($\sigma_{b,k}$), and entropy ($H_{b,k}$) for the k^{th} block, as follows:

$$\begin{aligned} &\mu_{b,k}(M_\alpha) = E[b_k], \\ &\sigma_{b,k}(M_\alpha) = \sqrt{\mathrm{var}[b_k]}, \qquad\qquad <6> \\ &H_{b,k}(M_\alpha) = -\sum_{i=1}^{D} p_i \log_D p_i, \end{aligned}$$

where p_j is the probability of values in the block that are quantized to D bins. Since the number of blocks on a matrix can be different, we simply use the expected value over all the blocks. That is, we use the overall mean $\mu_b = E[\mu_{b,k}]$, standard deviation $\sigma_b = E[\sigma_{b,k}]$ and entropy $H_b = E[H_{b,k}]$ as block-wise features. We expect that blogs with highly short-term coherency are likely to have high $\mu_b(M_\alpha)$, low $\sigma_b(M_\alpha)$ or low $H_b(M_\alpha)$ for some attribute α.

6.3 Joint features

We now discuss the joint features derived from a pair of self-similarity matrices on different attributes. From the matrices shown in previous section (ref. e.g. Figure 4), we observe that changes in different attributes are usually coincident. This effect is much stronger in splogs. For example, a splog might have a link to a "sports gear" website whenever posting about "river sport", or a link to a "sailboat" website whenever posting about "sail center". We conjecture that because the splogs are financially motivated, we expect correlations between attributes.

We compute the joint features to measure attribute correlation using joint entropy. The joint entropy measures how the distributions of two variables are related. It is computed as follows:

$$H(X,Y) = -\sum_{i=1}^{D_x}\sum_{j=1}^{D_y} p(x_i, y_j) \log p(x_i, y_j), \qquad <7>$$

Let $H_k(M_\alpha, M_\beta)$ be the joint entropy from the same k^{th} off-diagonals of a pair of matrices M_α and M_β, where α and β are two different attributes. Let $H_b(M_\alpha, M_\beta)$ be the joint entropy over the blocks of M_α and M_β.

$H_k(M_\alpha, M_\beta)$ is computed from the joint probability $p_k^{(d)}(x_i, y_i)$, where $p_k^{(d)}(x_i, y_i)$ indicates, after quantizing the k^{th} off-diagonal of M_α and the k^{th} off-diagonal of M_β into D_x and D_y bins respectively, the probability of an element on the k^{th} off-diagonal being contained in the bin x_i of M_α and in the bin y_i of M_β.

The joint entropy $H_b(M_\alpha, M_\beta)$ is computed from the joint probability $p^{(b)}(x_i,y_i)$, where M_α and M_β are segmented into D_x and D_y blocks respectively, and $p^{(b)}(x_i,y_i)$ indicates the probability of an element on the matrix being contained in the block x_i of M_α and in the block y_i of M_β. This analysis captures the block-structural similarity across the two attributes.

7. EXPERIMENTS

We now present preliminary experimental results on the splog detection. We show the discriminability of proposed temporal features in section 7.1, and in section 7.2, we compare the detection performance of the proposed features with the baseline content-based features.

Dataset description. In this work, 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, we have about 43.6K unique blogs. We focus our analysis on this subset of blogs having homepage and at least one entry.

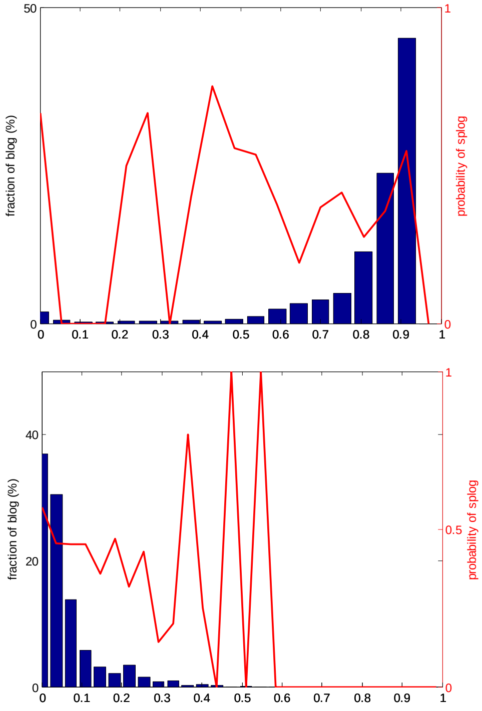


Figure 8: Top: $\mu_4(S_{macro})$. Bottom: $\sigma_3(S_{macro})$.

Annotation. We have developed an annotation tool for annotators (researchers and graduate students working at NEC Lab) to label the TREC-Blog dataset. For each blog, the annotators examine its content, out-going links, appearance of affiliate ads, etc. and assign one of the following five labels to the blog: (N) *Normal*, (S) *Splog*, (B) *Borderline*, (U) *Undecided*, and (F) *Foreign* Language. These labels are defined similar to the assessment task initiated in the web spam detection community[1]. A pilot study on the annotation results from a group of seven annotators shows the annotators have agreement on normal blogs but have varying opinions on splogs, which suggests that splog detection is not trivial even for humans.

Ground truth. We have labeled 9.2K blogs that are selected using random sampling. In order to disambiguate between splogs and non-splogs, and because most normal blogs are less ambiguous, we decided that those that are annotated as *Splog* need to be confirmed by a different annotator. Thus, we ended up with 8K normal blogs and 800+ splogs. We then randomly select 800 splogs and 800 normal blogs to create the evaluation set.

7.1 Feature Discriminability

In this section we discuss the impact of the proposed temporal features derived from the self-similarity analysis.

We examine the relationship between the feature value distribution and probability of finding a splog in the ground truth annotated data. In each of the figures in this sub-section, we see a histogram (shown using bars) and a curve. The histogram represents the distribution of non-splogs at each feature value. The curve shows the probability of splogs (i.e., the number of blogs divided by the total number of blogs) at each feature value. The splog probability curve in each figure is indicative of the utility of a temporal feature in splog detection, as the splog to non-splog ratio becomes either very high or very low for certain values of the features.

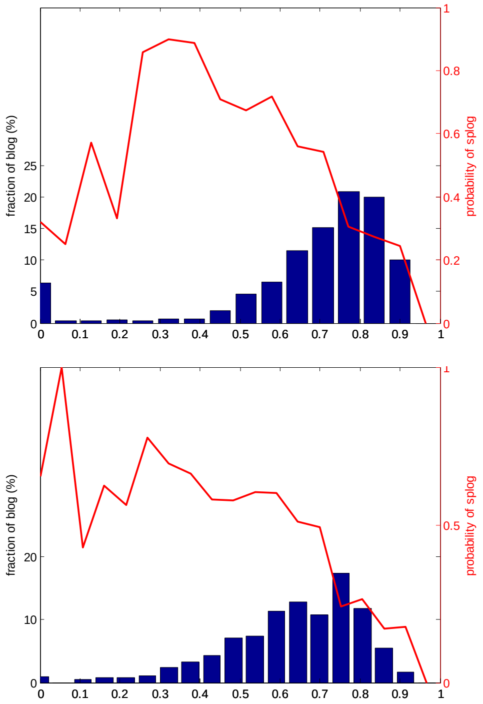


Figure 9: Top: $H_3(S_{macro}, S_{micro})$. Bottom: $H_2(S_{macro}, S_l)$.

In Figure 8 we see two examples of regularity features extracted along the diagonal of the self-similarity matrix. The top part of the figure shows the utility of the mean of the 4th off-diagonal on the macro-time self-similarity matrix, and the bottom figure shows the standard deviation of the 3rd off-diagonal elements on the macro-time matrix.

In Figure 9 we show examples of the utility of the joint features for splog detection. The top part of the figure shows the joint entropy between the macro- and micro-time matrices computed along the 3rd off-diagonal. The bottom figure shows the joint entropy along the 2nd off-diagonal for the macro-time and link self-similarity matrix.

7.2 Detection Performance

Our splog detector combines the new features (regularity and joint features) with traditional content features into a large feature vector. We then use standard machine learning techniques (SVM classifier implemented using libsvm package [2], with a radial basis function 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. We use well-known

[1] http://www.yr-bcn.es/webspam/datasets/uk2006-info/

performance metrics: AUC (area under the ROC curve), accuracy, precision and recall. In our experiment we regard splogs as positive cases.

The results in Table 1 are interesting. It shows that by *using the temporal characteristics* we get significant improvement. The temporal features, designated as *R* in Table 1, are constructed as a 32-dimensional feature vector. The baseline content features, designated as *base-n*, are *n*-dimensional feature vectors constructed using the content-based analysis alone. Features are selected separately using Fisher LDA. Table 1 shows the temporal structural features alone (*R*) out-performs the best 32 content features. It also suggests 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, designated as *R+base-n*, the performance improves over using content or temporal structures alone, indicating that they are complementary. The performance gain by using the temporal features is promising—the size of temporal features is relatively small, compared to the large size content features. In low-dimensional feature sets, the improvement by temporal features is significant. While the high-dimensional content features perform very well, there is a danger of over fitting.

The promising results suggest that the temporal features of a blog, is a key distinguishing characteristic and allows us to distinguish between splogs and normal blogs.

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 temporal features (*R*), and the combination of baseline with the temporal features (*R+base-n*). The table indicates a significant improvement by combining the temporal features. It also shows that the temporal features alone (*R*, with size 32) out-perform the best 32 content features (*base*-32).

Feature	AUC	accuracy	precision	recall
base-256	0.980	0.940	0.940	0.940
R+base-256	0.987	0.951	0.955	0.946
base-128	0.957	0.893	0.891	0.896
R+base-128	0.976	0.929	0.933	0.925
base-64	0.932	0.865	0.860	0.871
R+base-64	0.968	0.912	0.914	0.909
base-32	0.899	0.825	0.810	0.849
R+base-32	0.959	0.893	0.892	0.895
R	0.914	0.832	0.829	0.838

8. CONCLUSION

In this paper, we propose new framework based on blog temporal dynamics, to detect splogs in the blogosphere. In our approach, splogs are recognized by their temporal structures. While blog content is highly dynamic and time variant, the temporal structure captured by the regularity and joint features reveals a stable blog character. This stability makes time based structural features particularly attractive in splog detection.

We proposed the use of the topology of a time series of blog posts as an analysis framework. The topology is induced using a similarity measure on the posts. Topological analysis allows for a robust representation of a blog as it functions as a generalized spectral analysis tool. We showed how we can compute the self-similar characteristics of a specific attribute. We also showed how to extract statistical measures from the self-similar matrix representations.

The unique structural properties of splogs are captured by two types of temporal features. Regularity features are computed from the off-diagonals of the self-similar matrix as well as coherent blocks from self-similarity matrices. Joint features computed from self-similarity matrices across different attributes. We have evaluated our approach using standard classification performance metrics. The experimental results using the *topological analysis are excellent*, indicating that the temporal features work well in the splog detection application. The addition of traditional content features is shown to further improve performance.

There are several challenges that we propose to address as part of future research – (a) develop probabilistic representations of the topology and (b) short term topological analysis, including signature dynamics.

9. 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] J. Eckmann, S. O. Kamphorst and D. Ruelle (1987). *Recurrence plots of dynamical systems.* Europhysics Letters(4): 973-977.

[4] J. Foote, M. Cooper and U. Nam (2002). *Audio retrieval by rhythmic similarity*, Proceedings of the International Conference on Music Information Retrieval, 265-266,

[5] 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,

[6] 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,

[7] 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/.

[8] P. Kolari, T. Finin and A. Joshi (2006). *SVMs for the blogosphere: Blog identification and splog detection*, AAAI Spring Symposium on Computational Approaches to Analyzing Weblogs,

[9] P. Kolari, A. Java and T. Finin (2006). *Characterizing the Splogosphere*, Proceedings of the 3rd Annual Workshop on Weblogging Ecosystem: Aggregation, Analysis and Dynamics, 15th World Wid Web Conference,

[10] 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.

[11] 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.

[12] 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,

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