

The Splog Detection Task and A Solution Based on Temporal and Link Properties

Yu-Ru Lin, Wen-Yen Chen, Xiaolin Shi, Richard Sia, Xiaodan Song,
Yun Chi, Koji Hino, Hari Sundaram, Jun Tatemura and Belle Tseng

NEC Laboratories America

10080 N. Wolfe Road – Suite SW3-350, Cupertino, CA 95014

ABSTRACT

Spam blogs (splogs) have become a major problem in the increasingly popular blogosphere. Splogs are detrimental in that they corrupt the quality of information retrieved and they waste tremendous network and storage resources. We study several research issues in splog detection. First, in comparison to web spam and email spam, we identify some unique characteristics of splog. Second, we propose a new online task that captures the unique characteristics of splog, in addition to tasks based on the traditional IR evaluation framework. The new task introduces a novel time-sensitive detection evaluation to indicate how quickly a detector can identify splogs. Third, we propose a splog detection algorithm that combines traditional content features with temporal and link features that are unique to blogs. Finally, we develop an annotation tool to generate ground truth on a sampled subset of the TREC-Blog dataset. Initial experiments based on this ground truth set show promising results.

1. INTRODUCTION

The blogosphere is growing extremely fast and provides new business opportunities in areas such as advertisement, opinion extraction, and marketing. However, spam blogs (splogs) have become a major problem in the blogosphere—they reduce the quality of information retrieval results and waste network and storage resources [4,7]. Therefore, detecting splogs in the blogosphere has great importance.

In this paper, we propose our solution to detect splogs in the blogosphere. The main contributions of our work are as follows:

- Modeling the splog problem:** Unlike web or email spam, a splog is dynamic since it continuously generates fresh content to drive traffic. To solve the splog problem, we need to take advantage of the unique splog properties.
- Evaluation:** Splogs need to be identified as quickly as possible before they waste network and storage resources. We propose a time-sensitive evaluation framework to measure splog detection performance based on how fast the detection is made.
- Detection:** Our detection algorithm identifies unique features such as temporal and link properties useful for detecting splogs.

Most of previous work in spam detection comes from web spam detection. Prior work to detect web spams can be further categorized into content analysis [5,6] and link analysis [2,3]. Our work combines traditional features with temporal and link features that are unique to blogs.

2. WHAT ARE SPLOGS?

In this section, we provide a high-level definition of *splogs* and the splog problem we face today (Section 2.1), the typical splog characteristics (Section 2.2), and the differences between splog and other types of spam (Section 2.3).

2.1 Working definition of splogs

Spam blogs, which are called *splogs*, are undesirable weblogs that the creators use solely for promoting affiliated sites [1]. As blogs became increasingly mainstream, the presence of splogs has a detrimental effect in the blogosphere. According to multiple reports, the following are alarming statistics.

- 10-20% of blogs are splogs. For the week of Oct. 24, 2005, 2.7 million blogs out of 20.3 million are splogs [7].
- An average of 44 of the top 100 blogs search results in the three popular blog search engines came from splogs [7].
- 75% of new pings came from splogs; more than 50% of claimed blogs pinging *weblogs.com* are splogs [4].

The statistics exhibit serious problems caused by splogs, including (1) the degradation of information retrieval quality and (2) the tremendous waste of network and storage resources.

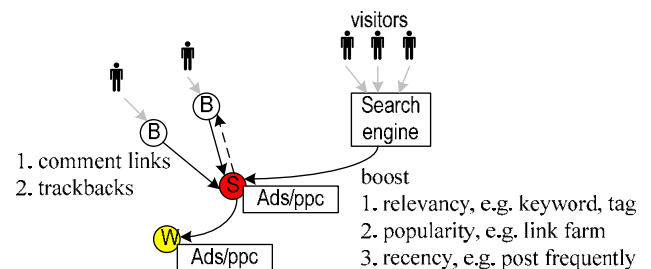


Figure 1: Splogs use different schemes to achieve spamming. “B” represents a blog, “S” represents a splog, and “W” refers to an affiliate site. There is usually a profitable mechanism (Ads/ppc) in the splog or affiliated site(s).

Figure 1 illustrates the overall scheme taken by splog creators. Their motive is to drive visitors to affiliated sites (including the splog itself) that have some *profitable mechanisms*. By profitable mechanism, we refer to web-based business methods, such as search engine advertising programs (e.g. *Google AdSense*) or pay-per-click (ppc) affiliate programs. There are several schemes used by spammers to increase the visibility of splogs by getting indexed with high ranks on popular search

engines. To deceive the search engine, the spammer may boost (1) relevancy (e.g. via keyword stuffing), (2) popularity (e.g. via link farm), or (3) recency (e.g. via frequent posts), based on some ranking criteria used by search engines. The increased visibility is unjustifiable since the content in splogs is often nonsense or stolen from other sites [1]. The spammer also attacks regular blogs through comments and trackbacks to boost the splog ranking.

2.2 Typical splog characteristics

In a typical splog, content is usually generated by machines in order to attract visitors through their appearance in either search engines or individual blogs. By splog, we refer to a blog created by an author who has the intention of spamming. Note that a blog that may contain spam in the form of comment spam or trackback spam is not considered a splog.

There are typical characteristics observed in splogs:

1. **Machine-generated content:** splog entries are generated automatically, usually nonsense, gibberish, repetitive or copied from other blogs or websites.
2. **No value-addition:** splogs provide useless or no unique information to their readers. There are blogs using automatic content aggregating techniques to provide useful service such as podcasting—these are legitimate blogs because of their value addition.
3. **Hidden agenda, usually an economic goal:** splogs have commercial intention that can be revealed if we observe any affiliate ads or out-going links to affiliate sites.

Some of these characteristics, such as no value-addition or hidden agenda, can also be found in other types of spams (e.g. web spam). However, splogs have unique properties that will be highlighted in the next section.

2.3 Uniqueness of splogs

Splogs are different from web spams in the following aspects.

1. **Dynamic content:** blog readers are mostly interested in recent entries. Unlike web spams where the content is static, a splog continuously generates fresh content to drive traffic.
2. **Non-endorsement link:** A hyperlink is often interpreted as an endorsement of other pages. It is less likely that a web spam gets endorsements from normal sites. However, since spammers can create hyperlinks (comment links or trackbacks) in normal blogs, links in blogs cannot be simply treated as endorsements.

Because of these two significant differences, the splog problem is different from that of traditional web spam as discussed next.

3. TASK DEFINITION

In this section we propose our evaluation methodology for comparing splog detection techniques on TREC blog dataset. We first describe the detection task framework in Section 3.1. Next, two detection tasks used in traditional information retrieval are given in Section 3.2. In Section 3.3 we propose an online detection task with novel assessment method.

3.1 Framework for detection task

The objective of a splog detector is to remove unwanted blogs. Blog search engines need splog detectors to improve the quality of their search results. Blog search engines differ from general web search engines in their growing contents – namely feeds. The detection decision is performed on a blog that consists of a growing list of entries. Because entries become available gradually, there can be time delay to gather enough evidences (i.e., entries) for detection. Since a splog will persist in the index until it is detected, earlier detection with few evidences is crucial for the overall search quality. We refer a detector that can make a decision with less evidence *fast*.

An illustration of how early splog detection is beneficial is shown in Figure 2. The grid represents how the amount of entries (x-axis) increases over time for each blog (y-axis). For a specific time, the gray area denoted by “downloaded in the storage” shows the number of blogs discovered with the corresponding amount of entries. As time passes, more blogs are indexed as well as growing amounts of entries, as shown by the dashed border and arrows.

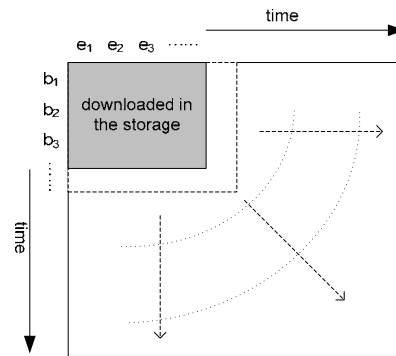


Figure 2: Blogs are discovered and downloaded over time. Similarly, the amount of entries downloaded grows over time. The gray area represents blog data that have been downloaded, and the dashed border and arrows show the downloading process continuing over time.

The target objective for blog search engines is to detect splogs as early as possible. As a result, we need to measure the speed of splog detection.

Traditional detectors are evaluated *offline*, where a batch of data is inputted into the detector and some performance metrics are calculated on the detection results. Because we want to also evaluate the speed of splog detection, we propose an *online* detection evaluation.

Another aspect of evaluation depends on the availability of ground truth information. Both offline and online detection can be evaluated with or without ground truth. Accordingly, there are four tasks as identified in Table 1.

Table 1: Four detection tasks are identified based on Offline and Online detections.

	Offline (Traditional)	Online (Time-Sensitive)
With Ground Truth	TASK 1	TASK 3
Without Ground Truth	TASK 2	TASK 4

3.2 Traditional IR-based detection

To compare different detection methods, there are two evaluation frameworks used in traditional information retrieval research and also widely applied in many TREC tracks.

3.2.1 Evaluation with ground truth

Evaluation is designed to compare detectors for an input set of blogs. Given a set of input blogs B with labels, the detectors can be evaluated by k -fold cross-validation, where the performance can be measured by metrics such as precision/recall, AUC, or ROC plot.

3.2.2 Evaluation without ground truth

To evaluate detector performances on a large dataset, there will be limited amount of labeled ground truth. Each detector makes its decision on the large dataset, and returns the detection results as a ranked list. The detector performance is evaluated by measuring the precision at top N (precision@ N) of the ranked list based on pooling of multiple detection lists.

Based on the availability of ground truth, splog detectors can be compared using one of the above offline evaluations. However to measure the speed of detection efficiency, we propose an online detection framework.

3.3 Online detection

As discussed above, the benefit of early splog detection is to quickly remove entries by splogs from the search index. Hence, we propose a new framework to evaluate time-sensitive detection performance.

We want to measure the detection performance on newly discovered blogs and observe how the decisions on these blogs can improve as more entries are available. First, blogs in the dataset are partitioned based on the time of discovery (i.e., the first appearance in the dataset). We assume the splog detector evaluates the blog contents at uniform frequency, i.e. $t_0 = t, t_1 = t + \Delta t, \dots, t_k = t + k \cdot \Delta t$. $B(t_i)$ is defined as a partition that consists of blogs discovered after time t_{i-1} and before t_i . $B(t_0)$ is the initial training set, usually given with labels (splog or non-splog).

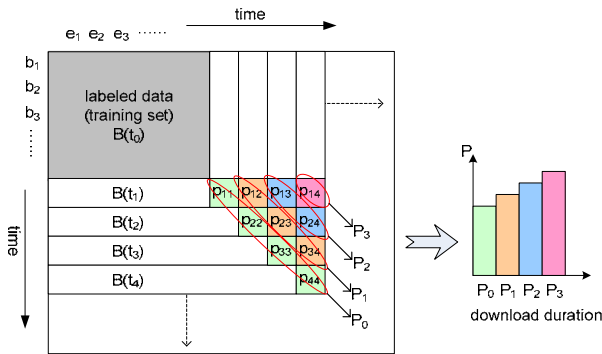


Figure 3: Online time-sensitive evaluation performance.

For each partition $B(t_k)$ ($k > 0$), the detector gives a decision at time t_j ($j \geq k$), for which the performance p_{jk} is measured. p_{jk} is the detection performance at time t_j on the partition at t_k ($B(t_k)$). In order to measure the speed of detection, we are interested in how the detector performance p_{jk} improves as j increases. Then, we introduce an overall performance measure of all decisions made with a specific delay. More specifically, for each delay $i =$

$j - k$, the overall performance P_i is given as an average, where $P_i = E[p_{jk} | i = j - k]$. The performance is plotted with i on the x-axis as shown in Figure 3, to demonstrate how quickly the detector can make a good decision. Note that each performance p_{jk} is measured based on the same evaluation metrics as the traditional offline evaluations. We expect the proposed online evaluation to provide significant insights on early detection of splogs.

4. OUR DETECTION METHOD

We have developed new techniques for splog detection based on temporal and linking patterns, which are unique features that distinguish blogs from regular web pages.

Due to the special characteristics of splogs, traditional content-based or link-based spam detection techniques are not sufficient. It is difficult to detect spams for individual pages (i.e., entries) by content-based techniques, since a splog can steal (copy) content from normal blogs. Link-based techniques based on propagation of trust from legitimate sites will work poorly for blogs since spammers can create links (comments and trackbacks) from normal blogs to splogs.

Our idea comes from the fact that a blog is a growing sequence of entries rather than individual pages. We expect that splogs can be recognized by their abnormal temporal and link patterns observed in entry sequences, since their motivation is different from normal, human-generated blogs. In a splog, the content and link structures are typically machine-generated (possibly copied from other blogs / websites). The link structure is focused on driving traffic to a specific set of affiliate websites. To capture such differences, we introduce new features, namely, temporal regularity and link regularity, which are described in the following subsection.

Our splog detector combines these new features with traditional content features and uses machine learning algorithm (SVM-based classifier) to classify each blog into two classes: splog or normal blog. We also plan on experimenting with other classifiers.

4.1 Temporal regularity estimation

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

Structural regularity is given by the probability distribution of time of content creation relative to the previous content. A high peak in the distribution indicates a high probability of a splog.

Content regularity is given by the autocorrelation of the content, derived from feature vectors such as term and citation frequencies as well as a similarity measure between vectors. We define a similarity measure based on the Kullback-Liebler divergence. The low decay rate of the auto-correlation function indicates a high probability of being a splog.

4.2 Link regularity estimation

Link regularity measures consistency in target sites pointed by a blog. The consistency exists because the links in a splog are used for driving traffics to specific affiliated sites. These affiliated sites are seldom pointed to by normal blogs. Therefore, if we cluster sites by using co-citation relationship, the link targets of normal blogs are likely to be diversified and the link targets of splogs are likely to be focused. As a result, we first

group the link targets of blogs through bi-partite clustering (we use SVD), and then use entropy to measure the diversification of the link targets.

5. DATA-PREPROCESSING AND GROUND TRUTH DEFINITION

We have made significant efforts to pre-process the TREC-Blog dataset and to establish ground truth for training and testing. Our major contributions are summarized as follows:

1. **Pre-processing:** The TREC-Blog 2006 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.
2. **Annotation tool:** We have developed a user-friendly interface (Figure 4) for annotators to label the TREC-Blog dataset. The detailed description of the tool is available at our webpage¹ and will be presented in the full version of this paper. Essentially, in the interface, the content of the blogs and their contents are fetched from the database and presented to the annotator.

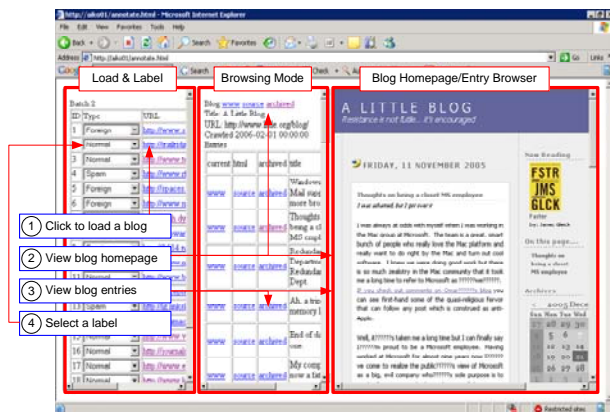


Figure 4: Splog Annotation Tool to view and label blogs.

Through the interface as shown in Figure 4, an annotator can browse the blog homepage and entries that have been downloaded in the TREC-Blog dataset, or visit the blog site directly online, in order to assign one of the following five labels: (N) Normal, (S) Splog, (B) Borderline, (U) Undecided, and (F) Foreign Language.

3. **Disagreement among annotators:** We performed a pilot study to investigate how different annotators identify splogs. We presented a set of 60 blogs to a group of 6 annotators, asking them to assign each blog to one of the five labels. One interesting result is that the annotators have agreement on normal blogs but have varying opinions on splogs (S/B/U), which suggests that splog detection is not trivial even for humans. We plan to conduct further intensive user studies.

4. **Ground truth:** As of August 24, 2006, we have labeled 9240 blogs by using our annotation tool. The 9240 blogs are selected using random sampling as well as stratified sampling methods. Among these 9240 blogs, 7905 are labeled as normal blogs, 525 are labeled as splogs, and the rest are borderline/undecided/foreign. The annotated splog percentage is lower than what has been reported because (1) some known splogs are pre-filtered from the TREC dataset, and (2) we have selected to examine the 43.6K subset of blogs that have both homepages and entries downloaded.

Using the annotation tool to generate a collection of ground truth, we built a baseline splog detector and our detector. Based on the offline and online detection tasks described in Section 3, preliminary results on our splog detector are promising. These experimental results will be described in detail in the full version of this paper.

6. CONCLUDING REMARKS

We study the splog detection problem as an important open task for TREC-Blog track. A splog is significantly different from web spam, and thus new detection tasks are identified. The new task measures how quickly a detector can identify splogs. We also provide a set of ground truth labeled through our annotation tool. We are currently building a splog detector that combines content features with temporal and link properties. Early experimental results on the annotated ground truth set are promising.

7. REFERENCES

- [1] Wikipedia, Spam blog <http://en.wikipedia.org/wiki/Splog>.
- [2] 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.
- [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] 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/>.
- [5] 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), July 2006, Boston, MA.
- [6] 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, May 2006, Edinburgh, Scotland.
- [7] UMBRIA (2006) *SPAM in the blogosphere* http://www.umbrialistens.com/files/uploads/umbria_splog.pdf.

¹ http://www.public.asu.edu/~ylin56/project/splog_detection/