

Splog Filtering based on Writing Consistency

Wei Liu, Songbo Tan, Hongbo Xu
Institute of Computing Technology, Chinese
Academy of Sciences, Beijing
{liuwei, tansongbo, hbxu}@software.ict.ac.cn

Lihong Wang
National Computer Network and Information
Security Management Center, Beijing
wlh@mail.nisac.gov.cn

Abstract

Splog is the key challenge in the access of blogosphere. Existing splog-filtering methods are restricted to the way for traditional web spam filtering, without considering the characteristics of blogs. Inspired by the observation that fake writers (writers of splogs) have striking higher consistent writing behavior than real writers (writers of legitimate blogs), we propose to detect splogs by distinguishing fake writers from real writers. To measure how consistent the writing behavior is, we propose the consistency-based features derived from writing interval, writing structure and writing topic. Then we designed a splog-filtering system which can use the consistency-based features effectively and flexibly. The experimental results on Blog06 data set show that, proposed measure can effectively detect splogs, reaching an accuracy of 90%. Compared with content-based methods, our approach can get a comparable accuracy with fewer features and smaller train set, indicating that writing consistency represents the essential difference between splogs and blogs.

1. Introduction

A traditional definition for blogs is that: Web pages containing entries displayed in reverse chronological order, with the function as online diaries [1]. Beyond publishing content, blogs also enable users to engage in conversations and generate tight communities, which make blogs become new influential social medium. With the astonishing expansion of the blogosphere [2], blogs receive a lot of attention in commercial concern as well as academic research. The blogosphere provides us with a rich and innovative data collection which makes it possible to look into people's interest, views, and feelings [3]. On top of this, many special information access tasks are inspired, such as blog classification, blog retrieval, temporal analysis, social network analysis [1].

How to guarantee the quality of blogosphere is the key issue for analyzing blog data successfully. The injection of splogs can seriously degrade the research results based

on blogosphere as well as waste network resources. Splogs are blogs with auto-generated content created for the purpose to attract search engine, raise the PageRank of target pages that are higher than they deserved, or distort the research results. A formal study taken by Kolari [4] shows some alarming statistics. Approximately 75% of pings are received from splogs, around 20% of blogs indexed by search engine are splogs. It is suggested that splogs now rival web spam and e-mail spam, becoming a major problem in the face of blog analysis.

Mishne [1] views the splog filtering process as the first step to get access into blogosphere, Finin [5] submitted splog detection to be a new task in Trec2007. As the splogs have extremely resourceful and innovative methods [5], many specific filtering methods are developed to combat splogs. Prior works to detect splogs can be categorized into two kinds: content analysis [6] [7] [8] and link analysis [9] [10]. They detect splogs by applying blog characteristics into web spam filtering techniques. In their work, each blog is treated as a special case of static web spam pages, without considering the unique features of blogs.

The key difference between blogs and other web pages is that blogs represent individuals. The series of posts in a blog can be used to analyze the blogger's writing behavior. It is observed that fake writers (writers of splogs) have striking higher consistent writing behavior than real writers (writers of legitimate blogs). Hence, we attempt to attack splog-filtering task from the perspective of "writing behavior", that is to detect splogs by distinguishing fake writers from real writers. The basic idea is that, the more consistent the blogger's writing behavior is, the more probable the blogger is a fake writer. In our approach, blog is treated as an individual rather than separate page. Our filtering method identifies splogs by analyzing the local content of a blog without additional information. There are three main contributions in this work:

Firstly, we proposed effective measurements to evaluate the consistency-based features from three perspectives, i.e., writing interval, writing structure and writing topic. Experiment on each feature shows that the

consistency distributions are significantly different between splogs and blogs.

Then we proposed writing consistency based splog-filtering system, which is flexible to add new consistency-based features or to change the combination algorithms.

Lastly, we conduct a series of experiments on the writing consistency, which show that it can be used as an effective and robust method for splogs-filtering task.

The remainder of this paper is organized as follows. Next section talks about related work. In section 3, we investigate three consistency features, and reveal the different distributions between splogs and blogs. Proposed consistency-based filtering system is described in section 4. In section 5, we present experimental results. Section 6 concludes this paper and discusses the future work.

2. Related work

Spam is not a new phenomenon, but the ease-to-create and fast-to-distribute nature makes it more pervasive in blogosphere. Existing splogs filtering methods can be categorized into content analysis and link analysis.

As examples of content-based splog filtering, [6] [7] introduce several new features such as bag-of-urls and bag-of-anchortext computed from different parts of blogs. After comparing the results of using different feature sets in terms of classification performance by SVM classifier, the bag-of-word from content and urls are the best features for splogs filtering. In their work, blogs are treated as isolated and static pages, ignoring their dynamic updated nature. Moreover, blogs are always written in an informal way and cover a variety of topics. As a result, large amount of training data is required for term-frequency based methods. Preparing labeled data is a time-consuming job, which makes it unfeasibility to online filter. Otherwise, the splog often has plagiarized content copied from legitimate blogs, hence it is not sufficient to rely only on content features.

Link spam in web pages has been investigated sufficiently in [11]. It is proved that link analysis can be used to detect link spam effectively. [9] introduces TrustRank to detect spam web pages by propagating the trust score from a reputable seed pages selected manually, following the citation link structure of the web. Kazunari [10] uses co-citation clusters derived from splogs seed to detect splog. Its basic premise is that blogs are unlikely to link to splogs. However, blog has a complicated link structure, including blogrolls (link to other blogs), inner-links (link to other posts in blog), and interaction links (links in comments and trackbacks). The citation links between blogs are no longer trustworthy because with the help of comments and trackbacks mechanisms, spammers can easily create links from good blogs to splogs. Therefore, the propagating of TrustRank and the clustering of blogs based on citation link structure are not

trustworthy. In addition, the global link structure of blogosphere is always unavailable.

Lin [8] first considers the temporal dynamics nature, and exploits blog temporal features to detect splogs. When combined with content-based features, it improves the classification accuracy by 10%. It is suggested that high self-similarity is a good indicator of splogs. The dynamics of blogs can be used to detect splogs.

Our work in this paper is essentially different from the methods mentioned above, since we are focusing on the writer's behavior of a blog. We propose a novel technique to detect splogs by distinguishing the fake writers from real writers. As blogs representing individuals, a blog can be mapping with a person's life, and reveal the dynamic behavior to the writer. From the post-time series, content structures and concerned topics, we can get a general picture of the writer, including writing frequency, writing habits and the writer's interests. It is observed that the consistency of writing behavior between splogs and blogs is different. The writing behavior of splogs is more inclined to higher consistency, which can be used to recognize fake writers, namely the owners of splogs.

3. Analysis of writing behavior

In this section, we investigate the writing consistency from three aspects, including writing interval, writing structure and writing topic. With several simple heuristic measurements, we reveal the prevalence of splogs relative to consistency of writing behaviors.

First we give some formal definitions. Assume that we have a sequence of N post ordered in times, from the same blog, such as $X = \{x_1, x_2, \dots, x_n\}$, where $t(x_i) \leq t(x_{i+k})$ for all $k \geq 0$. In this paper, we take $k = 1$, because we just consider the relationships between two immediate adjacent posts. Assume that the writing behavior can be described by a set of attributes and for each attribute α , we define a consistency function $C_\alpha(X)$. The function reveals how consistent the writer is on the attribute α .

3.1. Consistency on writing interval

Blog post contains a time stamp recording the time when it is created. In this subsection, the time stamps of posts are used to analyze the writer's writing consistency. As one important means for measuring data fluctuation, we employed variance analysis to concretize the consistency function on writing interval.

Assuming that we have a sequence time series of post in the same blog, and then we can obtain the time interval series $\{\Delta t_1, \Delta t_2, \dots, \Delta t_{n-1}\}$, where $\Delta t_i = t(x_{i+1}) - t(x_i)$, for $i \in \{1, 2, \dots, N-1\}$. The consistency function of writing interval is defined as follows:

$$C_i(\mathbf{X}) = \frac{1}{\sqrt{\frac{\sum_{i=1}^{n-1} (\Delta t_i - \bar{\Delta t})^2}{n-2}}} \quad (1)$$

where $\bar{\Delta t}$ is the average time intervals, the smaller the variance is, the higher the consistency function is.

In most cases, in order to keep the splogs being indexed, spammers update the splogs frequently with a fixed interval. In our first experiment, we investigate whether a high consistency of writing interval is a good indicator for fake writers. For this purpose, we plotted the distribution of the consistency function value of each blog in our data set (consists of 1,239 splogs and 3,788 blogs, more details about the data set are illuminated in section 5.1). The result is shown in Figure 1.

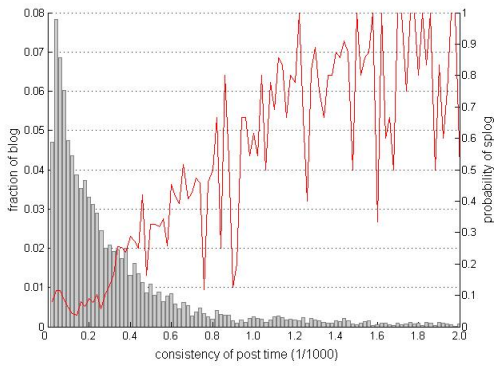


Figure 1. Consistency of writing interval.

This figure, like following figures in the rest of this section, consists of a bar graph and a line graph. The bar depicts the distribution of consistency value on a certain attribute (in this subsection, consistency of writing interval) of all blogs in our data set. The horizontal axis depicts a set of value ranges. The left scale of the vertical axis applies to the bar graph, and depicts the percentage of blogs fell into a particular range. The right scale of the vertical axis applies to the line graph, and depicts the percentage of splogs.

As can be observed in Figure 1, more than half of all blogs have low consistency values on writing interval. The prevalence of splogs is higher for blogs with high consistency, although the line graph gets noisier towards the right, due to the small number of sampled blogs with high writing consistency. The writing intervals of real writers are always stochastic. On the contrary, splogs always have fixed intervals predefined to keep blog updating frequently, resulting in an obviously higher consistency of writing interval. As can be seen in Figure 1. When the consistency value reaches 2.0, the variation of writing interval is $8(\text{min}) (\approx 1/(2.0/1000)(\text{s}))$, such a high consistency value of writing interval is abnormal for real writers. It is suggested that high consistency on writing

interval can imply a deviant writing behavior, which is a good indicator for fake writers.

3.2. Consistency on writing structure

The second attribute to describe a writer's writing behavior is writing style, including syntactical structures, diction, and figures of thought [12]. However, the NLP techniques required for these analyses are computationally expensive. As an alternative, we choose a more lightweight approach to measure the writing style, namely writing structure. One popular practice when creating splogs is keyword stuffing, with the hope that by mixing extraneous words with the legitimate content, the splog could match more queries. Hence, splogs always have excessive number of words.

Given a blog $X = \{x_1, x_2, \dots, x_n\}$, let s_i be the content length of x_i . Then we define the consistency function of writing structure like this:

$$C_s(\mathbf{X}) = \frac{\bar{s}}{\sqrt{\frac{\sum_{i=1}^n (s_i - \bar{s})^2}{n-1}}} \quad (2)$$

where the \bar{s} is the average length of content. A high structure consistency depends on two factors, i.e., high average and low variation. To estimate the role the consistency of writing structure plays in splogs, for every blog in our data set, we calculated the consistency function value. We define the content length to be the word number in main content (excluding makeup and comments).

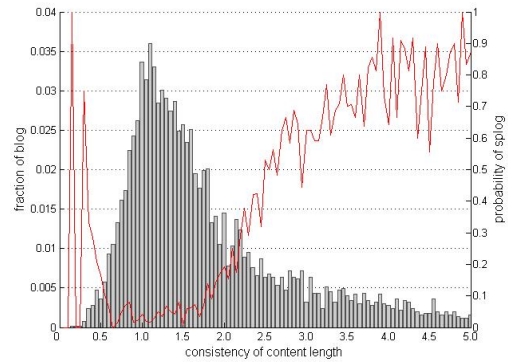


Figure 2. Consistency of writing structure.

Figure 2 illustrates the relationship between the distributions of the structure consistency and the probability that it is a splog. The noises in the left-most line are due to the small number of sampled blogs containing few words and high fluctuations. The splogs-probability line seems to have an upward trend, indicating that higher consistency on writing structure increase the possibility of being splogs.

It can be interpreted like this, for real writers, blog is a kind of online diary with a narrative nature. The length of content depends on the content and time spent to write, which varies with writer's environment, including interest, life and feelings. It is abnormal for a writer to write every post with considerable and similar length. As opposite to real writers, for fake writers, the sole purpose of holding a blog is to drive the search engine traffic by matching as many queries as possible with low cost. A popular practice is to use auto-generated or hijacked content from other resources, which make splogs tend to have long and fixed content length. After that, the consistency on writing structure is a good indicator for fake writers.

3.3. Consistency on writing topic

Writing topic represents the writer's interest, views and feelings, which vary with the writer's life and mood. The writing topic is a typical attribute for analyzing the writer's writing behavior. In this subsection, we examine whether the consistency on writing topic have different distributions between blogs and splogs.

Given a blog $X = \{x_1, x_2, \dots, x_n\}$, the consistency on writing topic is defined using cosine similarity on the tf-idf vectors. Let $c_i = \{w_{i1}, w_{i2}, \dots, w_{it}\}$ be the tf-idf vectors (after stemming and stop-word removal) for post x_i . The similarity between two posts is defined as:

$$sim(x_h, x_k) = \frac{\sum_{j=1}^t w_{hj} \times w_{kj}}{\sqrt{\sum_{j=1}^t w_{hj}^2} \times \sqrt{\sum_{j=1}^t w_{kj}^2}} \quad (3)$$

Then the consistency function of writing topic is defined as:

$$C_c(X) = \frac{1}{n-1} \sum_{i=1}^{n-1} sim(x_i, x_{i+1}) \quad (4)$$

For each blog in our data set, we plotted the consistency value on writing topic. The result is shown in Figure 3.

As can be observed in Figure 3, the prevalence of splogs is higher for blogs with high consistency value on writing topic. Most real writers have low consistency value on writing topic, which verifies the proposition mentioned before. Real writers tend to have various topics drawn from real lives. To the contrary, fake writers utilize duplicate content to keep a higher-than-deserved ranking. They mostly tend to stay on a certain topic, resulting in a high consistency on writing topic. Consequently, the consistency on writing topic is a useful indicator for fake writers.

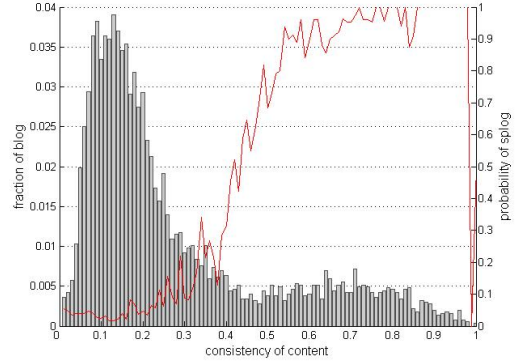


Figure 3. Consistency of writing topic.

4. Writing consistency-based filtering system

We now describe our writing consistency-based splog-filtering system, which is proposed to use consistency-based features efficiently. As Figure 4 shows, the process can be partitioned into two phases: consistency analysis on writing behavior and combination of consistency-based features.

4.1. Consistency analysis

During this phrase, various attributes that can be used to portray writing behavior are investigated. This is the first step to make use of writing consistency. As discussed in section 3, in this work, we describe writing behavior from three aspects: including writing interval, writing structure and writing topic. Then we propose using consistency function C_t , C_s and C_c to measure the consistency on each attribute, through which we obtain writing consistency-based features.

4.2. Combination of consistency-based features

In previous sections, we presented the consistency on three attributes of writer's writing behavior. It is proved that, high consistency on these attributes is an obvious symbol of fake writers. Nevertheless, when used individually, no technique detects most of the splogs without many mistakes. For example, consider the consistency of writing interval described in Section 3.1, the average probability of splogs for value of 1.0 and higher is 65%. But only about 13% of all blogs fall in this range. This number is far below the 25% splogs in our data set. In order to detect splogs more efficiently we combine the proposed consistency-based features. We first introduce the using of the Naïve Bayes probabilistic model to combine the consistency of writing behavior. Then in section 5, we show the performance of our features combined with popular classification algorithms.

Given a consistency vector $\mathfrak{X} = \{C_t, C_s, C_c\}$, where C_t is the consistency of writing interval, C_s is the consistency of writing structure, C_c is the consistency of writing topic.

The integral writing consistency is formally computed through the Naïve Bayes probabilistic model as follows:

$$\begin{aligned}
 & P(x \in B_s | (C_t, C_s, C_c)) \\
 &= \frac{P(x \in B_s) \times P((C_t, C_s, C_c) | x \in B_s)}{P(C_t, C_s, C_c)} \quad (5) \\
 &= P((C_t, C_s, C_c) | x \in B_s) \\
 &= P(C_t | x \in B_s) \times P(C_s | x \in B_s) \times P(C_c | x \in B_s)
 \end{aligned}$$

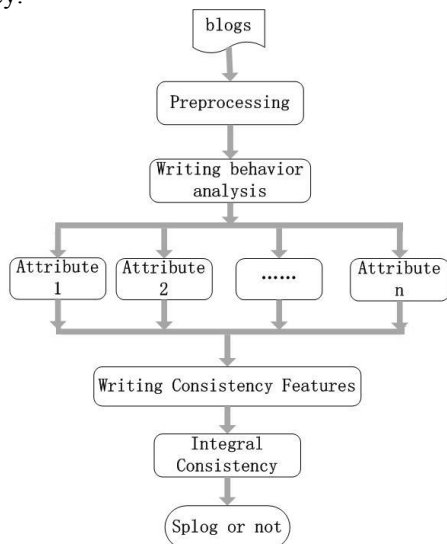
where B_s represents splogs. Because the fraction of splogs and the denominator are the same to every blog, with the conditional independence assumptions, the model can be expressed as the final concise form. For consistency on each attribute $C_\alpha \in \mathfrak{R}$, in this paper $\alpha \in \{t, s, c\}$, the independent probability distribution is estimated like this:

$$P(C_\alpha | x \in B_s) = \begin{cases} P(C_\alpha \geq \theta_\alpha | x \in B_s) & \text{if } C_\alpha \geq \theta_\alpha \\ 1 - P(C_\alpha \geq \theta_\alpha | x \in B_s) & \text{else} \end{cases}, \quad (6)$$

where θ_α is model parameter vector that can be approximated with relative frequencies from the training set. Consequently, we can compute the integral consistency of writing behavior for each writer of blog in our data set. The greater the integral consistency value is, the more probable the writer is a fake writer.

Our consistency based filtering-system is flexible and extensible. With the further investigation of writing behavior, new useful consistency-based features can be added in phrase one. And various strategies can be used to combine the consistency-based features.

In the following section, we apply several standard classification algorithms as our combination strategies, including SVM, Naïve Bayes model and decision-tree techniques, hoping to detect more splogs with greater accuracy.



Writing Consistency based Splog Filtering System Architecture
Figure 4. splog-filtering system.

5. Experiments

We now present the experimental results on our splog-filtering method. We first describe the data set in section 5.1. In section 5.2, we discuss our baseline method, present the classifiers and evaluation metrics we used. In the last subsection, we compare the classification performance between writing consistency-based features and content-based features.

5.1. Dataset description

In this work, we use the TREC (the Text Retrieval Conference) Blog Track 2006 dataset for analysis [13]. This dataset contains 11,649 feeds collected over 11 weeks, from Dec. 6, 2005 to Feb. 21, 2006. To simulate the real blogosphere, 17,969 known splogs are inserted into the collection (corresponding to 17.8% of the feeds).

We have labeled 6,000 blogs that are selected using random sampling. After removing duplicate feeds and feeds without any post, we ended up with 3,788 legitimate blogs and 1,239 splogs.

5.2. Experiment settings

We experimented with a variety of classification techniques: SVM classifier (implemented using libsvm package [14], with a polynomial kernel), Naïve Bayes model and decision-tree techniques (implemented using C4.5 decision tree [15]).

Link-based splog-filtering requires the entire link structure of the blogosphere, which is always unavailable. Content-based features are regarded as useful features in detecting splogs [6] [7] [13]. Therefore, we take the content-based method as the baseline schema. These features are used to discriminate the splogs from legitimate blogs, based on the term frequency divergence.

We first extract features from four different parts of a blog, including tokenized URLs, post titles, anchor text and post content, which are proved to be very effective for splog identification [6]. The total number of unique terms we get is greater than 20,000 (excluding words containing digits and no-alphabet characters). To reduce the length of the feature vector, we take the information gain (IG) as the feature selection method.

In order to evaluate the performance of our method, we randomly select 1,000 splogs and 1,000 legitimate blogs to create the evaluation set. We use well-known performance metrics: F1 measure, precision and recall. The definition of the F1 measure is shown as follows:

$$F1 = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \quad (7)$$

5.3. Detecting performance

We compared our writing consistency-based features with content-based features from three aspects: accuracy, feature number and size of training data.

5.3.1. Accuracy. We chose the top 10,000 content-based features as the baseline scheme. Using five-fold cross-validation technique, we evaluate our writing consistency-based features and content-based features on different classifiers. The results can be summarized in Table 1. From Table 1 we can see that, our writing consistency-based features (WRC) out-perform the content-based features (CF) on F measure in each case. The performance gain by using consistency-based features is encouraging: the size of our method is relatively small, compared to the large size of content-based features.

Table 1. Performance of our method.

Method	Precision	Recall	F1
CF + SVM	0.88	0.83	0.854
WRC + SVM	0.931	0.833	0.879
CF + Bayes	0.827	0.905	0.864
WRC + Bayes	0.843	0.891	0.866
CF + C4.5	0.787	0.856	0.82
WRC + C4.5	0.935	0.875	0.904

5.3.2. Feature number. In order to compare the performance in terms of feature number, we test the classification accuracy on different number of features. We set the feature number from 1,000 to 10,000 by increasing 1,000 at each time. The classification results after five-fold cross validation process are plotted in Figure 5. For each classifier, we use the same shape of lines to present the F1 value of consistency-based features and content-based features. Since the number of our writing consistency-based features is a constant, the lines for consistency-based features are straight lines parallel with the horizontal axis.

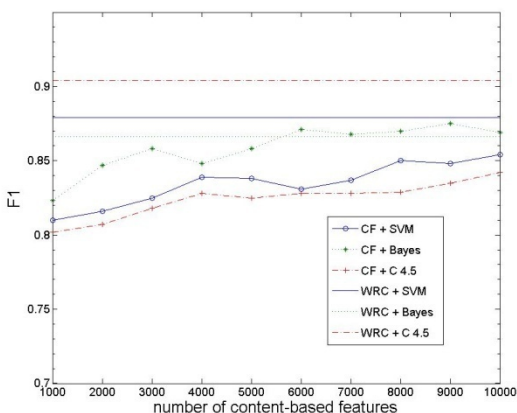


Figure 5. Compare consistency with content-based features in terms of feature number.

Take the results of naïve Bayes classifier represented by two dotted lines as examples. A clear upward trend can

be seen in the dotted curve with stars, suggesting that the performance of content-based features improves gradually with the increasing of feature number. However, when the feature number greater than 6,000, the curve of content-based method shows a tendency to decline. This makes intuitive sense: splogs usually copy content from legitimate sources, which degrade the discrimination of content-based features. So to improve classification accuracy by increasing feature number is impracticable.

5.3.3. Size of training data. In order to compare the training cost in terms of sample size, we employed a technique known as n-fold cross validation. N-fold cross-validation involves dividing the data set randomly into n partitions with equal sizes, and performing n training and testing steps, each step use n-1 partitions to train the classifier and the remaining partition to test its effectiveness. By means of this method, we can evaluate the performance in different sizes of training data.

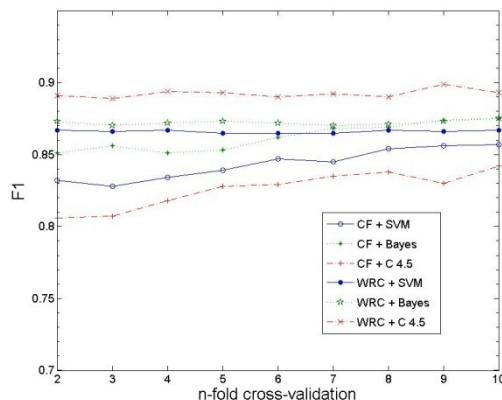


Figure 6. Compare consistency with content-based features in terms of sample size.

Figure 6 depicts the comparisons between consistency-based features and content-based features in terms of different sizes of training data using different classifiers. As can be observed in Figure 6, for each classifier, the classification performance of consistency-based features is stable with little fluctuation, while the classification performance of the content-based features shows upward trend with the increasing of training data size. This can be explained like this: writing consistency is a distinctive feature of writers. The difference of writing consistency between fake writers and real writers is an intrinsic indicator, and high consistency of writing behavior is a general phenomenon.

Consequently, the experimental results are encouraging for two reasons. First, using the writing consistency-based feature can substantially reduce the feature number without loss of accuracy. Second, writing consistency-based classifier need less training data to get

its stable classification accuracy, indicating a low training cost and better generalization ability.

6. Conclusions and future work

In this paper, we propose new approach based on writing behavior analysis, to detect splogs in the blogosphere. We first analyze writing consistency from three aspects: writing interval, writing structure and writing topic. We propose effective measurements to evaluate the consistency-based features. The experimental results demonstrate that the consistency distributions are significantly different between splogs and blogs.

To use the consistency-based features effectively, we propose a flexible and extensible filtering system based on writing consistency. In this system, more consistency-based features can be added and various classification algorithms can be used to combine the features.

We have evaluated our approach using three standard classification algorithms. The comparisons with content-based methods show that our consistency-based method can yield much better performance, even with fewer features as well as smaller training samples, indicating that writing consistency can be used to detect splogs effectively.

Filtering splogs from the perspective of writing behavior is a heuristic technique. All of our consistency-based features are independent on the language used in blogs, so our approach can be used to detect splogs written in English as well as in other languages.

Our method is not a perfect solution that can address the problem of splogs thoroughly, and some of the features can be manipulated by spam programs easily. In our future work, we will attempt to develop more complex features to characterize writing behavior. It is our hope that continued research in this area can make splogs more expensive than legitimate blogs.

7. Acknowledgments

This work is mainly supported by the projects (2007AA01Z441&2006AA010105&2007AA01Z416&0704021000&2006AA01Z452) and the 973 National Basic Research Program of China (2004CB318109 & 2007CB311100).

8. References

- [1] G. Mishne. "Information Access Challenges in the Blogspace". *International Workshop on Intelligent Information Access*, Finland, 2006.
- [2] <http://technorati.com/weblog/2006/11/161.html>
- [3] Chun-Yuan Teng, and Hsin-His Chen. "Detection of Bloggers' Interests: Using Textual, Temporal, and Interactive Features". *Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*, Washington D. C. USA, 2006, pp. 366-369
- [4] P. Kolari, A. Java, and T. Finin. "Characterizing the splogosphere". *Proceeding of the World Wide Web 2006 Workshop on the Weblogging Ecosystem*, Edinburgh, 2006.
- [5] P. Kolari, A. Java, T. Finin, J. Mayfield, A. Joshi, and J. Martineau. "Blog Track Open Task: Spam Blog Classification". *Technical report of TREC 2006 Blog Track*, September 2006.
- [6] P. Kolari, T. Finin, and A. Joshi. "SVMs for the blogosphere: Blog identification and splog detection". *Proceeding of the AAAI Spring Symp. On Computational Approaches to Analyzing Weblogs*. AAAI Press, California, 2006, pp. 92 - 99.
- [7] F. Salvetti, N. Nicolov. "Weblog Classification for Fast Splog Filtering: A URL Language Model Segmentation Approach". *Proceeding of the Human Language Technology Conference of the NAACL*, New York, 2006, Companion Volume: Short Papers, pp. 137 - 140.
- [8] Y.R. Lin, H. Sundaram, Y. Chi, J. Tatemura, and B.L. Tseng. "Splog Detection using self-similarity analysis on blog temporal dynamics". *Proceeding of the ACM Workshop on Adversarial information retrieval on the web*, Canada, 2007, pp. 1-8.
- [9] Z. Gyöngyi, H. Garcia-Molina and J. Pedersen. "Combating web spam with TrustRank". *Proceedings of the 30th International Conference on Very Large Data Bases (VLDB)*, 2004, Volume 30, pp. 576-587.
- [10] K. Ishida. "Extracting Spam Blogs with Co-citation Clusters". *Proceeding of the 17th international conference on World Wide Web*, ACM, China, 2008, pp. 1043-1044.
- [11] B. Davison. "Recognizing Nepotistic Links on the Web". *AAAI-20000 Workshop on Artificial Intelligence for Web Search*, Texas, July 2000.
- [12] http://en.wikipedia.org/wiki/Writing_style
- [13] C. Macdonald, I. Ounis. "The TREC Blog06 Collection: Creating and Analysing a Blog Test Collection". *DCS Technical Report TR-2006-224, Department of Computing Science*, University of Glasgow, 2006.
- [14] <http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/>
- [15] J.R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan-Kaufman, 1993.