# Extracting Spam Blogs with Co-citation Clusters

Kazunari Ishida
The University of Shimane
2433-2 Nobara-cho, Hamada-shi, Shimane 697-0016, JAPAN
+81-855-24-2275
k-ishida@u-shimane.ac.jp

## ABSTRACT

This paper reports the estimated number of spam blogs in order to assess their current state in the blogosphere. To extract spam blogs, I developed a traversal method among co-citation clusters of blogs from a spam seed. Spam seeds were collected in terms of high out-degree and spam keyword. According to the experiment, a mixed seed set composed of high out-degree and spam keyword seeds is more effective than individual seed sets in terms of F-Measure. In conclusion, mixed seeds from different methods are effective in improving the F-Measure results of spam extraction with co-citation clusters.

## Categories and Subject Descriptors

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

## General Terms

Algorithms, Measurement, Experimentation

## Keywords

Spam Blog Extraction, Co-citation Cluster, Advertisement Link

## 1. INTRODUCTION

Blogs are useful social media for researching consumer opinion on products and services, but there are many spam blogs that benefit from advertising affiliations. To get a sense of the overall number of spam blogs in the Japanese blogosphere, I observed updated blog data for the period of a week. Based on this observation period and other previous research [1], I found that spam blogs form co-citation clusters because they share advertisement links, even though other characteristics, e.g. term sequence, URL, anchor text, and tags were easily changed. Based on the nature of co-citation clusters in spam, I developed a traversal spam extraction method, employing a Shared Interest algorithm [2] [3] to extract co-citation clusters. The SI algorithm affords that a blog or a cited page can be a part of multiple clusters. Using the nature of overlap in co-citation clusters, the traversal method extracts spam blogs from a spam seed set, which is automatically extracted by simple schemes. This method does not require any vast amount of learning data from spam blogs, although other methods [4] employing SVM or other learning algorithms may require such data.

## 2. SPAM IN JAPANESE BLOGOSPHERE

I analyzed a data set of updated blogs collected from the Japanese blogosphere from August 26th to September 1st, 2007. I applied a heuristic to convert URLs of individual blog pages of a blogger into the top-level URL of the blogger. The resulting number of unique blogs was 691,674. Because of the heuristic, frequently updated blogs have a high out-degree. Figure 1 illustrates the out-degree distribution of blogs in log scale for both axes. To analyze relation between spam blogs and out-degree of blogs, I arranged blogs in terms of out-degree, divided them into subsets which accounted for approximately 10% of all data, picked up 20 samples from each set, and counted spam in each sample.

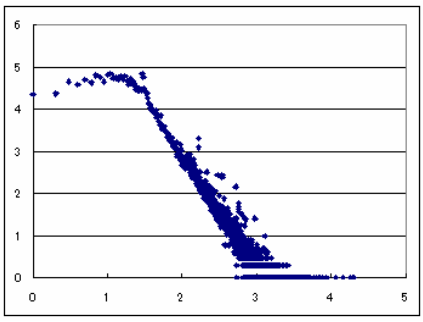


**Figure 1. Out-degree Distribution**

I evaluated each blog in each sample manually. A general criterion for evaluating spam blogs is how much value a blogger adds to his/ her blogs. Spammers automatically or semi-automatically copy parts of the other blogs and paste them on their blogs with many commercial affiliate links. As a result, they do not add any value to their blogs. Valuable blogs contain bloggers' opinions, experiences, and other creative content. When these valuable blogs contain advertisement links, the links are also valuable because readers are interested in their contents and tend to find useful, related products or services.

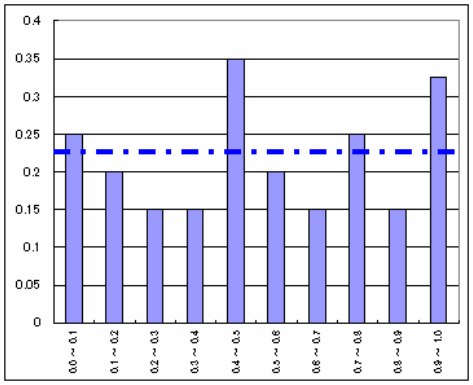


**Figure 2. Spam Distribution**

Figure 2 summarizes the percentage of spam in each subset, which accounts for approximately 23% of all data. In addition, I investigated high out-degree blogs, which are 1% (7038) of all blogs in the data set, and found that the high out-degree blogs are usually spam blogs (19 spam blogs out of 20 samples, or 95%).


*WWW2008*, April 21–25, 2008, Beijing, China.
ACM 978-1-60558-085-2/08/04.

## 3. SPAM EXTRACTION METHOD

I developed a traversal spam extraction method with co-citation clusters and a spam seed set:

1. Prepare a spam seed set.
2. Prepare a set of co-citation clusters.
3. Pick a blog from the spam seed set and put it into a spam pool.
4. Pick co-citation clusters containing the blog from in step 3 and remove those clusters from the set of co-citation clusters.
5. Extract all blogs contained in the clusters from step 4 and put them into a spam seed set. This does not include blogs already picked from the spam seed in step 3.
6. Repeat steps 3–5. When the spam seed set is empty, the procedure is finished.

To prepare the set of co-citation clusters, I used two criteria to select clusters from all clusters extracted from the blogosphere with the SI algorithm. The first criterion is density, which is defined by dividing the number of edges that a cluster has by all possible edges of the cluster. When the density is high, blogs in a cluster share interest because of many shared co-citation links. The density values 0.1, 0.5, and 0.9 will be used for the experiment in section 4. The second criterion is co-citation ratio, which is defined by dividing the number of blogs by the number of cited pages within a cluster. When the ratio is low, blogs in a cluster share interest because they have many co-cited pages. The co-citation ratio values 1, 10, and 100 will be used for the experiment in section 4.

To prepare spam seed sets, I used simple two schemes: high out-degree and spam keyword list. To prepare high out-degree seeds, I used the top 1% of blogs in terms of out-degree because of the high spam rate described in section 2. To prepare the spam keyword list, I selected 606 keywords for commercial affiliation and 341 keywords for adult content semi-automatically using representative keyword extraction as developed in [3]. Employing the keyword list, keyword seeds were collected from the one week data set with a certain threshold, i.e. 1, 3, 5 in this paper. The thresholds indicated that a spam blog contained spam keywords more than 1, 3, or 5 times. To investigate the effect of mixing spam seeds, I mixed high out-degree seeds and keyword seeds. The statistics of seven seeds, i.e. high out-degree seeds, three keyword seeds (1, 3, 5), and three mix seeds (1, 3, 5) are be summarized in section 4.

## 4. EXPERIMENT

I applied the traversal spam extraction method to the one week data set to find an effective spam seed set and parameter settings for density and co-citation ratio. I prepared co-citation clusters for the data set with same parameters as in [2]. Table 2 summarizes statistics of the data set and seven spam seeds sets I prepared. Precision was calculated by counting spam blogs within a set of 20 samples for each seed set. Recall was derived from the estimated number of spam blogs. Figure 3 illustrates the results of the experiments in terms of F-measure defined as [(Precision * Recall * 2) / (Precision + Recall)]. The blue dotted line represents the baseline F-measure of the data set (0.374). The table shows mean, min, and max values for each seed set with the two parameters, density and co-citation ratio. Figure 3 shows that the mixed seeds are more effective than the high out-degree seeds or the keyword seeds. There are few overlaps between high out-degree seeds and keyword seeds mixed into the mixed seed set, a result of the different schemes used to collect blogs. The wide variety of spam in the mixed seed is effective for applying the spam extraction method with co-citation clusters.

**Table 2. Data Set and Spam Seeds**

| Blog data | #blog | Data ratio | Spam ratio (Precision) | Estimated #spam | Recall |
|---|---|---|---|---|---|
| All data | 691674 | 1 | 0.23 | 158889 | 1.00 |
| High out degree seed | 7038 | 0.010 | 0.95 | 6686 | 0.04 |
| Keyword 1 seed | 60688 | 0.088 | 0.75 | 45516 | 0.29 |
| Keyword 3 seed | 14749 | 0.021 | 0.95 | 14012 | 0.09 |
| Keyword 5 seed | 7597 | 0.011 | 0.95 | 7217 | 0.05 |
| Mix 1 seed | 64527 | 0.093 | 0.65 | 41943 | 0.26 |
| Mix 3 seed | 20349 | 0.029 | 0.95 | 18314 | 0.12 |
| Mix 5 seed | 13613 | 0.020 | 0.95 | 12932 | 0.08 |

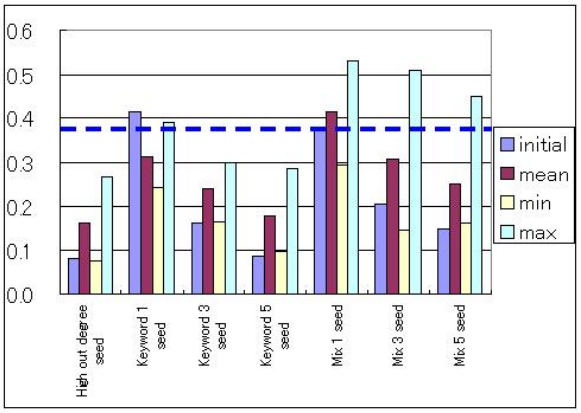


**Figure 3. Results of Spam Extraction**

## 5. CONCLUSION

This paper reported the percentage of spam blogs in the Japanese blogosphere. It also provided the traversal spam extraction method with co-citation clusters. According to the experiment conducted by this method, mixed seed is more effective in extractracting spam than high out-degree seeds or keyword seeds in terms of F-measure.

## 6. REFERENCES

[1] K. Ishida, "Extracting Latent Weblog Communities: A Partitioning Algorithm for Bipartite Graphs," Proceedings of the 2nd Annual Workshop on the Weblogging Ecosystem-Aggregation, Analysis and Dynamics in WWW2005, May 10–14, 2005.

[2] K. Ishida, "A Multi Clustering Algorithm based on Line Graph and Local Similarity," IPSJ-FI06083009, Vol.2006, No.59, pp. 69–76, 2006 (in Japanese).

[3] K. Ishida, "Mining Collective Opinions from the Blogosphere," Proceedings of CITSA2007, pp. 154–161, 2007.

[4] P. Kolari, T. Finin, and A. Joshi, "SVMs for the Blogosphere: Blog Identification and Splog Detection," AAAI Symposium on Computational Approaches to Analyzing Weblogs, pp. 92–99, 2006.