Combining Textual Content and Hyperlinks in Web Spam Detection

F. Javier Ortega[1], Craig Macdonald[2], José A. Troyano[1], Fermín L. Cruz[1], and Fernando Enríquez[1]

[1] Departamento de Lenguajes y Sistemas Informáticos
Universidad de Sevilla
Av. Reina Mercedes s/n 41012, Sevilla, Spain
{javierortega,troyano,fcruz,fenros}@us.es
[2] Department of Computing Science
University of Glasgow
Glasgow, G12 8QQ, UK
{craigm}@dcs.gla.ac.uk

Abstract. In this work[1], we tackle the problem of spam detection on the Web. Spam web pages have become a problem for Web search engines, due to the negative effects that this phenomenon can cause in their retrieval results. Our approach is based on a random-walk algorithm that obtains a ranking of pages according to their relevance and their spam likelihood. We introduce the novelty of taking into account the content of the web pages to characterize the web graph and to obtain an a priori estimation of the spam likelihood of the web pages. Our graph-based algorithm computes two scores for each node in the graph. Intuitively, these values represent how bad or good (spam-like or not) a web page is, according to its textual content and the relations in the graph. Our experiments show that our proposed technique outperforms other *link-based* techniques for spam detection.

Keywords: Information retrieval, Web spam detection, Graph algorithms, PageRank, web search.

1 Introduction

Web spam is a phenomenon where web pages are created for the purpose of making a search engine deliver undesirable results for a given query, ranking these web pages higher than they would otherwise [7]. Basically, there are two forms of spam intended to cause undesirable effects: Self promotion and Mutual promotion [4]. Self promotion tries to create a web page that gains high relevance for a search engine, mainly based on its content. This can be achieved through many techniques, such as word stuffing, in which visible or invisible keywords are inserted in the page, in order to improve the retrieved rank of the page for the

[1] Partially founded by Spanish Ministry of Education and Science (HUM2007-66607-C04-04).

most common queries. Mutual promotion is based on the cooperation of various sites, or the creation of a wide number of pages that form a *link-farm*, that is a large number of pages pointing one to another, in order to improve their scores by increasing the number of in-links to them.

There are different approaches to deal with the problem of web spam. They are usually classified into two groups, depending on the spam mechanism that they attempt to identify. *Content-based* techniques use the textual content of the web pages to classify them. These methods usually examine the distribution of statistics about the contents in spam and not-spam web pages, such as the number of words in a page, the HTML invisible content, the most common words in a page in relation with the ones in the entire corpus, etc. [4,5].

On the other hand, *link-based* techniques focus on the structure of the graph made up of the web pages and the hyperlinks among them. TrustRank [6] belongs to this group. It consists in a graph-based ranking algorithm that gives more weight in the computation to a set of hand-picked, trusted web pages. Another interesting work on this field is presented in [1].

Our contribution consists in a method that integrates concepts from both techniques for spam detection, combining the information from the content and the links of the web pages in order to build a ranking where spam web pages are demoted.

The organization of the rest of the paper is as follows. In the next section, we introduce the intuition behind our approach. The experimental design and results are shown in Section 3. Finally, we provide our conclusions concerning the present work, and discuss some ideas for future work.

2 Combining Links and Contents

Our approach consist of two parts: a graph-based ranking algorithm and a set of content-based heuristics. The aim of these metrics is to obtain some a piori information about the spam-likelihood of a web page, in accordance to its content. The values of the heuristics are included in the algorithm in order to introduce a bias in the computation of the ranking of web pages, in such way that the system reduces the final rank of every web page related to some (a priori) spam page, and vice versa. In our system we have implemented two content-based metrics: the average word length of each web page, and the number of repeated words. The ranking algorithm is an extension of PageRank [8], but it computes two scores for each node in the web graph: $PR^+(v_i)$, representing the relevance of node v_i in the graph, and $PR^-(v_i)$, that is the spam likelihood of web page v_i. The ranking of web pages is built regarding the difference between PR^+, and PR^+. These scores are computed following Equations (1) y (2).

$$PR^+(v_i) = (1-d)e_i^+ + d \sum_{j \in In(v_i)} \frac{PR^+(v_j)}{|Out(v_j)|} \tag{1}$$

$$PR^-(v_i) = (1-d)e_i^- + d \sum_{j \in In(v_i)} \frac{PR^-(v_j)}{|Out(v_j)|} \tag{2}$$

where v_i is a node in the web graph (a web page), $In(v_j)$ and $Out(v_j)$ are the set of inlinks and outlinks of v_j, respectively. The algorithm iterates over the graph applying the equations (1) y (2), until the highest difference between the scores of a node in two consecutive iterations is less than a certain threshold t. Vectors e_i^+ and e_i^- contain the information about the content-based metrics. In this way, we can give more relevance to some web pages over the rest, depending on the values of e_i^+ and e_i^-. These webs are the *seeds* of our algorithm. We have implemented three methods for the selection of these seeds:

- Most Positive and Negative web pages (MPN): the N web pages with highest and lowest content-based metrics are chosen as seeds. Each seed is initialised with a value of $e_i = 1/N$.
- Most Positive and Negative web pages with Metrics (MPN-M): similar to the previous method, but the values of the seeds are computed as follows: $e_i = Metrics_i/N$.
- All the web Pages as Seeds (APS): every web page in the collection are used as seeds, applying the same formula to obtain their initial weights.

3 Evaluation

Since our approach does not classify the web pages between spam or non-spam, it does not make sense to perform an evaluation in terms of classification accuracy. In our experiments, we use the PR-buckets evaluation method, also followed in other works on the application of graph-based algorithms to the spam detection task, such as [2,6]. Their intuition is that it is more important to correctly detect the spam in high PageRank valued sites, because they will often appear in the top positions in the search results for many queries. The aim of this evaluation method is to easily determine the number of spam web pages detected mainly in the highest positions of the ranking. We have used TrustRank as the baseline for our experiments. As mentioned above, this technique is based on a graph-based ranking algorithm that chooses by hand a set of a priori relevant web pages.

Table 1. Cumulated errors for each method: TrustRank (TR), Most Positive and Negative web pages (MPN), Most Positive and Negative web pages with Metrics (MPN-M) and All the web Pages as Seeds (APS). The best result for each case is highlighted.

#	Web pages	TR	MPN	MPN-M	APS
1	14	**0**	2	**0**	**0**
2	68	2	5	**1**	3
3	212	17	16	**4**	8
4	649	40	48	**21**	32
5	1719	73	104	**66**	101
6	3849	155	244	**124**	199
7	6513	254	392	**180**	297
8	9291	371	557	**255**	416
9	12102	448	742	**350**	537
10	14914	511	937	**440**	650

The experiments have been performed using the WEBSPAM-UK2006 Dataset [3], specifically compiled for the research on web spam detection. The corpus has about 98 million web pages and 120 million links. A subset of 11 thousand web hosts have been labelled as spam or not-spam, obtaining 10 million spam web pages in total. We have used the Terrier [2] system in order to efficiently index the corpus.

In Table 1 we show the results of TrustRank and our approach, with its three methods for the selection of seeds. The number of spam web pages in the 10 first buckets (the first positions of the ranking) are shown. MPN-M is the approach that obtains the best results for all the buckets.

4 Conclusions and Future Work

In this work we present a web spam detection system that takes into account both content and link-based information. The system has been evaluated with a standard corpus, achieving very good results and even outperforming a state-of-art technique, TrustRank.

We plan to further our work by implementing new content-based heuristics and studying the impact of these metrics in the overall performance of the system. It is also very interesting to prove the influence of the seed set in the final results of the system, and to experiment with alternative approaches for the selection of seeds.

References

1. Becchetti, L., Castillo, C., Donato, D., Baeza-Yates, R., Leonardi, S.: Link analysis for web spam detection. ACM Transactions on the Web 2(1), 1–42 (2008)
2. Benczur, A.A., Csalogany, K., Sarlos, T., Uher, M., Uher, M.: Spamrank - fully automatic link spam detection. In: Proceedings of the First International Workshop on Adversarial Information Retrieval on the Web (AIRWeb), pp. 25–38 (2005)
3. Castillo, C., Donato, D., Becchetti, L., Boldi, P., Leonardi, S., Santini, M., Vigna, S.: A reference collection for web spam. SIGIR Forum 40(2), 11–24 (2006)
4. Cormack, G.V., Smucker, M., Clarke, C.L.A.: Efficient and Effective Spam Filtering and Re-ranking for Large Web Datasets. Computing Research Repository, abs/1004.5 (2010)
5. Fetterly, D., Manasse, M., Najork, M.: Spam, damn spam, and statistics: using statistical analysis to locate spam web pages. In: Proceedings of the 7th International Workshop on the Web and Databases, pp. 1–6. ACM, New York (2004)
6. Gyöngyi, Z., Garcia-Molina, H., Pedersen, J.: Combating web spam with trustrank. In: Proceedings of the Thirtieth International Conference on Very Large Data Bases, vol. 30, pp. 576–587, VLDB Endowment, Toronto (2004)
7. Najork, M.: Web spam detection. In: Encyclopedia of Database Systems, pp. 3520–3523. Springer, US (2009)
8. Page, L., Brin, S., Motwani, R., Winograd, T.: The pagerank citation ranking: Bringing order to the web. World Wide Web Internet And Web Information Systems (66), 1–17 (1999)

[2] `http://terrier.org/`