Larger is Better: Seed Selection in Link-based Anti-spamming Algorithms

Qiancheng Jiang, Lei Zhang, Yizhen Zhu, Yan Zhang [*]
Department of Machine Intelligence, Peking University
Beijing 100871, China
{jiangqc, zhangl, zhuyz, zhy}@cis.pku.edu.cn

ABSTRACT

Seed selection is of significant importance for the biased PageRank algorithms such as TrustRank to combat link spamming. Previous work usually uses a small seed set, which has a big problem that the top ranking results have a strong bias towards seeds. In this paper, we analyze the relationship between the result bias and the number of seeds. Furthermore, we experimentally show that an automatically selected large seed set can work better than a carefully selected small seed set.

Categories and Subject Descriptors

H.3.3 [**Information Systems**]: Information Search and Retrieval

General Terms

Algorithms, Experimentation, Measurement

Keywords

Biased PageRank, Link Spamming, Seed Selection

1. INTRODUCTION

Web spam is one of the most intractable mischievousness to the search engines. They exploit many illegal means to benefit from high ranking positions. Many link-based anti-spamming techniques have been proposed so far[1, 3, 2, 4] for combating them. In general these approaches are all biased PageRank algorithms. As mentioned in previous work [1, 4], the seed selection plays an important role in differing good pages from bad ones. Traditional approaches such as TrustRank[1] and ParentPenalty[3] usually use a manual process to carefully select a small seed set. However, this process is always time consuming. It is lumbersome and awkward for periodical refreshing of the seed sets, especially when the spamming tricks are adaptive and the web environment is rapidly evolutive. Besides, when the number of seeds is small, the top ranking results are almost all occupied by seeds or their neighbors due to the refilled value of each seed per iteration in these algorithms. As far as we know, until recently no previous work has taken these issues into consideration. In this paper, we demonstrate our preliminary results on these research points.

2. RESULT BIAS ANALYSIS

Among all of the biased PageRank algorithms, propagating rank values via links from a small seed set is a general option. However, using a small seed set has a big problem: the ranking results have a strong bias towards seeds. That is, the top ranking results are always occupied by the seeds.

The bias is mainly due to the damping factor. During the computation, each seed will be refilled with $(1-\alpha_d)\cdot 1/N_s$ after each iteration, where α_d is damping factor and N_s is the number of seeds. Therefore, the side effect is large when the seed set is small. So seeds can occupy most of the top positions in the final result. To reduce this result bias, a feasible way is increasing the number of seeds for reducing the refilled value. Since the number of seeds cannot be guaranteed always enough, we should decide the minimal number. The point is how many top results are users concerned. If a user only concerns top 100 results and wishes they are less affected by result bias, the number of seeds can be small. Whereas a user concerns a large range of top ranking results, the number of seeds should be large.

In actual fact we can estimate the number of seeds using the assumption as follows: when we concern top N results, we assume that if $(1-\alpha_d)\cdot 1/N_s$ is less than the $(\gamma N)^{\mathrm{th}}$ page's score (γ is an expansion coefficient), it is less effected by result bias. So we can first get the $(\gamma N)^{\mathrm{th}}$ page's score then calculate the number of seeds. For example, when $N = 100$, $\gamma = 10$ and $\alpha_d = 0.85$, if the 1000th page has a score of 4×10^{-5}, we can get $N_s = 3750$.

3. EXPERIMENT

We perform experiments on a partial set of pages crawled by Tianwang search engine (developed by network lab, Peking University) in Nov. 2005. It contains 13.3 M pages with about 232 M links on 358,245 sites, most of which belong to $.cn$ domain.

3.1 Result Bias and Number of Seeds

With TrustRank[1], we start from 50 seeds and double the number each time. At each point, we randomly select

[*] Corresponding author

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

different seed set 4 times and calculate the average number of seeds that top 100 and top 1000 results contain. The result is shown in Table 1. It indicates that the top results are nearly all occupied by seeds when the seed set is small. The number of seeds in top 100 results reaches the nadir at the case of 3200.

Table 1: Result Bias for TrustRank

number of seeds	number of seeds in top 100 results	number of seeds in top 1000 results
50	50	50
100	95.5	100
200	92	200
400	78.75	400
800	49.25	800
1600	21.75	812
3200	15	587.25
6400	18.25	419.25
12800	33.5	428.5

To explore this trend more preciously, we start from 1600 seeds and enlarge the number by 100 each time. We perform this experiment four times at each point and get the average. The result is shown in Figure 1. The x-axis shows the number of seeds while the y-axis represents the corresponding ratio. We see this ratio runs to stable when the number of seeds is about 4000. By checking the scores, we find the 1000th site's TrustRank value is about 3.98×10^{-5}, which is perfectly matched with our estimation in Section 2.

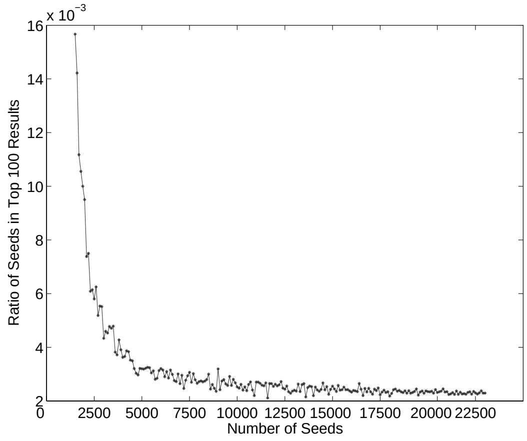

Figure 1: Ratio of average number of seeds in top 100 results to total number of seeds

3.2 Combating Link Spamming

In order to find out the impact of the number of seeds on the ability of combating link spamming, we use a method similar to that in TrustRank[1]. We generate a list of sites in descending order of their PageRank scores and segment these sites into 20 buckets. Each of the buckets contains a different number of sites with scores summing up to 5% of the total PageRank scores. We construct a sample set of 1000 sites by selecting 50 sites at random from each bucket. Then we perform a manual evaluation to determine their categories. Each site is classified into one of the following categories: reputable, spam, pure directory, and personal blog. Any site uses any spamming techniques will be put into spam category. We throw away the non-existent sites and reselect another one.

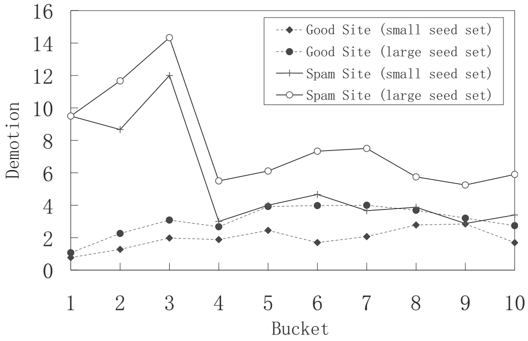

Figure 2: The bucket-level demotion of TrustRank scores with different seed sets

To compare the anti-spamming abilities of different seed sets, we select a small seed set $\mathfrak{X}$ using a method similar to TrustRank [1]. At the same time, we select all the sites in the *.gov* domain and *.edu* domain as a large seed set $\mathfrak{L}$. Figure 2 shows the bucket-level demotion of TrustRank scores when using $\mathfrak{X}$ and $\mathfrak{L}$. Good sites (reputable and directory) with high rankings have little demotion, i.e. retain high ranking values. There is no obvious difference when using these two seed sets. The average demotion of the good sites is almost less than 4. However, spam sites have more demotion and using $\mathfrak{L}$ is much better than using $\mathfrak{X}$. The demotions are always larger than 5.8 with $\mathfrak{L}$.

4. CONCLUSION AND FUTURE WORK

In this paper, we reveal that a large seed set can achieve a better performance than a small seed set on detecting web spam for biased PageRank algorithms. What is more, instead of carefully selecting a small seed set, we can select a large number of seeds automatically. For example, we can just select sites in the *.gov* and *.edu* domains as seeds. No doubt that this process is time saving. So when using a large seed set, we can obtain good result as well as simplification of selecting process.

Our future work will explore some unanswered question about seeds selection. For example, how to exploit large seed sets more effectively and can we get "useful" bad seeds from good ones? We will focus on these problems in the future.

5. ACKNOWLEDGEMENT

This work is supported by NSFC under Grant No.60673129, 60773162 and 60573166, Beijing Natural Science Foundation under Grant No.4073034

6. REFERENCES

[1] Z. Gyongyi, H. Garcia-Molina, and J. Pedersen. Combating web spam with trustrank. In *VLDB '04*.

[2] V. Krishnan and R. Raj. Web spam detection with anti-trust rank. In *AIRWeb'06*, August 2006.

[3] B. Wu and B. D. Davison. Identifying link farm spam pages. In *WWW '05*, May 2005.

[4] L. Zhang, Y. Zhang, Y. Zhang, and X. Li. Exploring both content and link quality for anti-spamming. In *CIT '06*, page 37, Washington, DC, USA, 2006.