From Good to Bad Ones: Making Spam Detection Easier *

Li Zhao  Qiancheng Jiang  Yan Zhang†
Key Laboratory of Machine Perception ( Ministry of Education )
Peking University, Beijing 100871, China
lizhao@pku.edu.cn, futuremac@gmail.com, zhy@cis.pku.edu.cn

Abstract

Previous researches of the anti-spamming have proved that results can be greatly improved when bad seeds are used together with good ones. However, how to select bad seeds efficiently is a big challenge. In this paper we discuss how to select bad seeds based on good seeds selection. The experiments running on over 13M web pages show that our method is practical and time-saving. Moreover, the selected bad seeds can enhance the performance of a good seed set on effectively filtering out spam from normal pages.

1. Introduction

With the search engines’ increasing importance in the people’s life, there are more and more attempts to mischievously influence the page rankings. This kind of action called web spamming is always illegal, since it misleads both search engines and users seriously. Furthermore, it has a negative economic and social impact on the whole web community. Spammers are playing tricks on the search engines by all means, for example, term spamming, link spamming, cloaking and redirection [1]. Among these tricks, link spamming is the most insidious and devastating one.

Many anti-spamming techniques have been proposed so far [2, 7, 10, 6, 4, 8]. TrustRank [2] improves the PageRank by using good seeds. It can effectively demote the pages that adopt link spamming tricks. Baoning Wu and Brian D. Davison propose algorithms for detecting link farms automatically by first generating a desirable seed set and then expanding it [7]. In actual fact, almost all of these biased ranking algorithms employ a seed set and this set plays an important role in identifying web spam.

It is mentioned in previous work that a mixed seed set which combines good and bad seeds together can lead to a better result. Nevertheless, previous research either ignores the discussion about bad seed selection or they do not use bad seeds at all. For example, the bad seeds are not involved in TrustRank algorithm. Though Baoning et al. present an algorithm for selecting bad seeds in their work [7], however, the propagation ability of the selected seed set cannot be guaranteed. Usually a good “bad seed” should be both “famous” (actually notorious) and powerful in penalizing its parents. In fact, choosing a powerful bad seed set is arduous and expensive.

Our research goal is finding a method to select bad seeds efficiently and effectively. In this paper, we intend to choose bad seeds based on the ranking demotion that is caused by good seeds’ trust propagation. Good seed selection method is well described in TrustRank algorithm. However, manually evaluation the selected seeds is a time-consuming process. In this paper we will present an easier way to choose a large good seed set and therefore we can obtain a group of powerful bad seeds more efficiently.

The rest of the paper is organized as follows. The related work is given in Section 2. In Section 3, we present some preliminary knowledge of seed selection. Section 4 describes the bad seed selection method in detail. Section 5 demonstrates the experimental results. Finally we draw the conclusion in Section 6.

2. Related Work

Previous work on the spam detection is mainly based on a ranking system. They have paid more attention to demote the spam sites’ ranking position. PageRank, HITS [5, 3] are the earliest ranking algorithm of the web pages. A page which has more in-links or out-links can gain a higher score. However, this ranking mechanism is exploited by many link spamming tricks such as link farm.

With corresponding to these link spamming tricks, many novel ranking algorithms which are called “biased PageRank” are proposed. Based on the assumption that non-spam sites seldom point to spam sites and vice versa, TrustRank [2] starts with a set of good seeds and propagate the trust via out-links. Similarly Krishnan et al. [4] propose that

---

*Support by NSFC under Grant No.60673129, 60773162 and 60573166, Beijing Natural Science Foundation under Grant No.4073034
†Corresponding author
disturbance can either be propagated via in-links. Wu et al. [6] start a biased random walk from a small spam seed set and then extract link farm and link exchange centroids.

There are also some methods propagating values in both directions. Wu et al. [8] compute trust and distrust separately and then combine them by simply linear summing. In our previous work [9], we propose a novel QoC-QoL algorithm which exploits bidirectional links to propagate ranking values.

All above methods always have a seed set at the beginning, good seed set or spam seed set or both. So a seed set is a crucial factor of these ranking methods. Baoning et al. [7] present a method of selecting bad seed set, but they neglect the propagation ability of the seed set. Krishnan et al. [4] also give a heuristic way to look for spam seeds, however, it is vague and hard to operate.

3. Preliminaries

3.1. Seed Selection in TrustRank

TrustRank is an effective algorithm which uses good seeds to give good pages higher scores. In order to find a suitable set which is not only useful in identifying good pages but also reasonably small to limit the number of oracle invocations, TrustRank discusses two strategies of seed selection. One is the high inverse PageRank. This can guarantee a good propagation coverage. The other one is the high PageRank. Good trust scores of sites with high PageRank is probably propagated to pages that are likely to be at the top of result sets. Thus, the good seeds can be selected by the following steps. First, the inverse PageRank of the pages are computed. Then the pages are ranked in decreasing order of their inverse PageRank and pages at the top L (limit of oracle invocations) compose a candidate seed set. Finally, the computer experts manually evaluate this set to select appropriate pages as good seeds.

3.2. QoC-QoL Algorithm

The QoC-QoL algorithm is introduced in our previous work [4]. QoC-QoL evaluates a page with its content and link quality separately. For a page \( p \), \( QoC(p) \) measures its content quality while \( QoL(p) \) computes its link quality. A page’s \( QoC \) is determined by its parent pages’ content and link quality, while the \( QoL \) value is determined by both its child pages’ content and link quality. The formulas for these two metrics are given as below:

\[
QoC(p) = \sum_{q_i \text{: points to } p} \left( \alpha \cdot \frac{QoC(q_i)}{ON(q_i)} + (1 - \alpha) \cdot \frac{QoL(q_i)}{ON(q_i)} \right)
\]

\[
QoL(p) = \sum_{r_i \text{: pointed by } p} \left( \beta \cdot \frac{QoC(r_i)}{IN(r_i)} + (1 - \beta) \cdot \frac{QoL(r_i)}{IN(r_i)} \right),
\]

Where \( ON(q_i) \) is the number of outgoing links of page \( q_i \), and \( IN(r_i) \) is the number of incoming links of the page \( r_i \). The \( \alpha, \beta \) are two decay factors. The equivalent matrix equation form is:

\[
QoC = \alpha M^T \cdot QoC + (1 - \alpha) M^T \cdot QoL
\]

\[
QoL = \beta N^T \cdot QoC + (1 - \beta) N^T \cdot QoL,
\]

In this form \( M \) is the line-normalized Web graph matrix, \( N \) is the line-normalized transpose matrix. These two metrics are both convergent [9]. In practical work, we use the square average of \( QoC \) and \( QoL \) (denoted as SAV) as an improvement metric of TrustRank. Since the QoC-QoL algorithm can make use of bidirectional links, therefore the propagation ability is more powerful. It is more effective than TrustRank. Moreover, it can exploit the web structure more deeply such as selecting much representative seeds according to web link graph. So the propagation coverage of a seed set can be larger as well as the normal and spam pages are distinguished more accurately.

4. Efficiently Selecting Powerful Bad Seeds

4.1. Bad Seed Selection Algorithm

Both TrustRank and QoC-QoL algorithm can significantly demote spam sites (or pages). Thus, we compute the SAV values (or TrustRank scores) for each web site in the experimental data set twice with different seed sets. For each computation, we list the sites in decreasing order according to the SAV difference. Then we select the first 1000 sites. The sites appear in the intersection of both computations, which have strong possibility to be spam sites, constitute a candidate seed set.

The function for bad seeds is to penalize its parents via its incoming links propagation. To maximize the influence of this penalty, we should select the sites with much inlinks. Thus we choose the sites with high SiteRank values from the candidate set as the final bad seeds. The algorithm is shown in Figure 1.

In order to deep into the algorithm, we use a simple example to illustrate the process of BADSEEDSELECTION. In Figure 2, we give a small web graph where good pages are drawn in white and bad pages in black. It is assumed that \( G_1 = \{1, 2\} \) and \( G_2 = \{3, 4\} \). Then going through the step 2 and 3, \( B_1 = \{5, 6, 7, 8\} \) is selected based on \( G_1 \) while
Input:
M  transition matrix of the web graph
N  transpose transition matrix of the web graph
G_i  good seed set, where \( i = \{1, 2\} \)
Output:
B  the bad seed set

Begin
1. Compute SiteRank values for each site \( s \),
2. Compute SA V values by using \( G_1 \) and \( G_2 \) for each site \( s \),
3. Sort the value (SiteRank - SA V) of each site \( s \) in descending order and select the top 1,000 sites separately as \( B_1 \) and \( B_2 \),
4. Get the intersection of \( B_1 \) and \( B_2 \) and put the top 100 sites with high SiteRank scores together as candidate seed set \( C \),
5. Manually assess the sites in \( C \) and then gain the final bad seed set \( B \).
End

Figure 1. BADSeedSelection Algorithm

Figure 2. A small web graph with good (white) and bad (black) nodes

\[ B_2 = \{6, 7, 8, 9\} \] is selected based on \( G_2 \). Finally, we can get \( B = B_1 \cap B_2 = \{6, 7, 8\} \) as the bad seed set, if we do not go further to evaluate it.

4.2. Making the Selecting Process Efficient

In our algorithm, computing SiteRank scores and SA V (square average of QoC and QoL) values are automatically and quite efficient. The time-consuming steps are selecting good seeds and step 5. To guarantee the quality of the seed set, step 5 must be committed manually. However, since we have limited the candidate set to a very small scale, this step is acceptable. To make the whole process of our algorithm more efficient, we should save time for selecting good sets. Usually there are two methods to select good seeds. One is the way described in the TrustRank. The other is selecting a large number of the sites which belong to .gov and .edu domain. The latter method is described in our previous work [9]. It is faster and more efficient than the method used in the TrustRank. At the same time, a large number of .gov and .edu sites works better than a small carefully selected seed set in most cases. The BADSeedSelection algorithm can achieve a quite good performance by using this method.

4.3. Why Using Two Computations

In the BADSeedSelection algorithm, we choose two different good seed sets and compute the SA V values twice. Is this necessary for obtaining bad seeds? We know both TrustRank and QoC-QoL are biased ranking algorithms. When a good seed set is used to compute the ranking scores, a site with no incoming links from any seed will have the minimum score, no matter it is actually a spam or a normal site. Therefore, the top 1000 sites with high SA V demotion will contain many non-spam sites, which will make our selection and evaluation more expensive and less effective. To alleviate this effect, we employ multiple seed sets. Different seed sets can decrease the false positives. A candidate in set \( C \) has a strong possibility to be a spam site in our BADSeedSelection algorithm. Thus the manual evaluation in step 5 is more efficient.

5. Experiment

5.1. Data Set

In order to evaluate the performance of our BADSeedSelection algorithm in selecting powerful bad seeds, we perform a number of experiments on a real web graph which is a partial set of pages crawled by Tianwang search engine (developed by network lab, Peking University) in Nov. 2005. The data set is consist of 13.3M pages and 232M links among these pages. We divide these pages into 358,245 hosts according to their URLs, most of which belong to .cn domain.

5.2. Seed Set Selection

As described in Section 4.2, we have two methods to select good seeds: carefully selecting a small seed set just like TrustRank or simply select a large number of .gov or .edu sites that mentioned in our work [9]. In the following experiments, we enumerate three combinations of the two methods to compare which method is the most effective in filtering spam sites. The different mixture of good seeds and the number of bad seeds selected are listed in Table 1. As shown in Table 1, we have three groups of seeds combinations. In the first group, we manually select two small good seed sets which contain 38 and 45 elements respectively. The bad seeds are generated by running the BADSeedSelection algorithm step by step. At first there
Table 1. Difference mixture of good seeds

<table>
<thead>
<tr>
<th>Group</th>
<th>Good Seeds (1)</th>
<th>Good Seeds (2)</th>
<th>Bad Seeds</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group1</td>
<td>38</td>
<td>40</td>
<td>45 (716)</td>
</tr>
<tr>
<td>Group2</td>
<td>38</td>
<td>24,943</td>
<td>38 (466)</td>
</tr>
<tr>
<td>Group3</td>
<td>10,952</td>
<td>13,911</td>
<td>43 (920)</td>
</tr>
</tbody>
</table>

are 716 hosts appearing in both $B_1$ and $B_2$. After evaluating the top 100 hosts with high SiteRank scores, we get a bad seed set including 45 hosts finally. Differing from the first group, we exploit a small good seed set and a large seed set which consists of 24,943 .gov or .edu hosts in Group2. Accordingly, we gain a bad seed set with 38 hosts. The combination of two large good seed sets in the third group produces a bad seed set that contains 43 bad seeds. We have three bad seed sets through our BADSEEDSELECTION algorithm so far. In the next part, we exploit both these bad seed sets and good seed sets to combat spam sites.

5.3. Performance for Filtering Spam

In order to validate whether these seed sets are really powerful in detecting spam sites, we settle on a sample of 1000 sites which can give enough information and still be manageable. First we sort the sites in descending order according to their SiteRank scores. Then the sites are divided into 20 buckets and the summation of the SiteRank values in each bucket is 5% of the total SiteRank score. We construct a sample set of 1000 sites by selecting 50 sites randomly from each bucket and then manually labeled these sites with four tags: reputable, spam, pure directory and personal blog. Once a site does not exist, we throw it away from the sample and reselect another one to supplement. The spam sites in the sample is 16% and this ratio is reasonable. The distribution of the sample set in total is presented in Figure 3. We also give the composition of each bucket in our sample set. The distribution is shown in Figure 4. There are spam sites in all of the buckets. In the top 10 buckets, the number of spam sites are unacceptable.

Before evaluating our algorithm, we regard the TrustRank as a baseline. The same as the process of sampling, we sort the sites in descending order by their TrustRank scores. Then the sites are divided into 20 buckets according to the sum of the TrustRank. At last we make a statistics on the sample sites in each bucket. The results are illustrated in the Figure 5. In this figure, the horizontal axe represents the bucket number from 1 to 20. The vertical axe corresponds to the proportion of each category in the bucket. The reputable and directory sites are treated as normal (non-spam) sites. Their contributions are shown in white and light gray segments respectively. The dark gray segment stands for the personal blog whose quality is difficult to distinguish. The empty part is the portion of spam sites. In the following comparison, we will pay more attention to the first 10 buckets because they always contain most of the top results and more significant for the users.

As shown in Figure 5, the results of the TrustRank algorithm with small good seed set is acceptable. The reputable sites are over 90% in the top 5 buckets. It is a big pro-
motion to the raw distribution in the sample set. However, there is much room to be improved in bucket 4 to 10. A few of buckets only contain 80% reputable sites, which is not enough for top results.

As an improvement, we use both good seeds and bad seeds to compute the SAV values of each sites. The initial values of good seeds sum up to 1 while the bad seeds sum up to -1. The value of each site is divided equally and the QoC and QoL values are equivalent at the beginning. The decay factor $\alpha$ and $\beta$ in the equation are both set to 0.5.

Since we already have three groups of bad seeds, now we mix the bad seeds and different good seeds to detect spam sites. At first, we mix 45 spam sites we selected in group1 and 38 good seeds. Figure 6 shows the performance on detecting spam of this mixture. Compared with the results of TrustRank shown in Figure 5, there is much improvement in the top buckets. The percentage of reputable sites in the first 3 buckets improves and the average percentage of spam sites in the top 10 buckets decreases notably.

In the following experiment, we mix the second group of 38 bad seeds with the same good seed set. Figure 7 exhibits the results of this combination. This experimental result is approximately the same as the result in Figure 6. It surpasses the effect of the TrustRank and can optimize the ranking of the results. These two experiments indicate that a mixture of good seeds and bad seeds can reinforce the ability of spam detection with QoC and QoL algorithm.

Besides using small good seeds, we also combine the large good seed set with the bad seed set. We use 10,952 good seeds which are selected automatically from .gov and .edu and the 43 bad seeds as a seed set next. The result of this experiment is illustrated in Figure 8.

In the top 3 buckets the effect is perfect and there is no spam sites in these buckets. The combination with large good seeds can achieve a little bit better effect than the above two. We also notice that the percentage of personal blogs whose quality are hard to detect is lower than the other two. This phenomenon may mainly lie on the fact that some bloggers post their blog urls to some famous sites, but this action takes no effect when we use large good seed sets.

Through these experiments, we can find that the mixture of good seeds and bad seeds is a good improvement of these ranking algorithms. This method is effective to demote the spam sites and maintain the rank position of the reputable sites at the same time. The results also suggest that the bad seeds combined with the large seed set is more effective to filter spam sites.

6. Conclusions

In this paper, we mainly proposed a method of selecting bad seeds which are used for combating web spam. We not only indicate some ways to select bad seeds, but also give the way to make use of them. In our algorithm, the bad seed set is mixed with the good seed set to compute the SAV scores of each page. We display the process of our selecting bad seeds in our experiment which is time-saving and easy to operate. At the same time, the experimental results show that the mixed seed set is effective in identifying web spam sites regardless of their way of combination. We also find that the mixture of large good seeds with the bad seeds is the best choice among these combinations. Since a large good seed set can be selected automatically, this kind of approach is the most efficient one in practice.

It is no doubt that there are still some interesting unsolved problems in our research. For instance, we have not find the most appropriate proportion of the good seeds and the bad seeds in the mixed set. Such issues are challenges in our future work.

References
