DATA SET SELECTION IN ANTI-SPAMMING ALGORITHM - LARGE OR SMALL

1 Ms. Jyoti Pruthi  
Research Scholar, Singhania University, Jhunjhunu (Raj.), India  
jyoti.pruthi756@gmail.com

2 Dr. Ela Kumar  
Associate Professor, Gautam Buddha University, Greater Noida, India  
Ela_kumar@gmail.com

ABSTRACT

Seed selection is of significant importance for the algorithms that are used 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 but the current study says that it is difficult to check the performance of the algorithm on the large seed set. In this paper, we analyze the relationship between the result bias and the number of seeds. Furthermore, we experimentally show the comparison between the large and the small seed set.

Keywords: Web Spam, Biased PageRank, Link Spamming, Seed Selection, Trust Rank

1. INTRODUCTION

Web spam is often defined as the practice of manipulating web pages in order to cause search engines to rank some web pages higher than they would without any manipulation. Spam is harmful not only because it can impact search engine results, but also because it can contribute more spam.

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 ant 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 Parent Penalty [3] usually use a manual process to carefully select a small seed
set. However, this process is always time consuming. It is cumbersome and awkward
for periodical refreshing of the seed sets, especially when the spamming tricks are
adaptive and the web environment is rapidly changing. Moreover, none of these
methods has taken the propagation coverage and result bias of the seed set into
consideration, which are critical to the final results. 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.

The web is enormous and continuously growing. This makes it impossible for
people to fight spam by manually checking each individual web page. All the
fighting methods should be automatic. In addition, the algorithms against spam
should be tested on a large-scale data set because some algorithms work well for a
small sample of the web, but fail to fight spam for data from the whole web. The
fact that some algorithms work for today’s web, but might not give good results in
the future is another reason for testing on a large-scale data set.

In our work, we discover that the propagation ability of a seed set has a great
relationship with web graph structure [5, 6] and the propagating direction. Web
graph is an extremely sparse graph in nature which can be divided into several parts.
Seeds in different parts have different propagation characteristics. In previous work
such as TruskRank [1], the seed set must be carefully selected from different parts in
order to maximize the propagation coverage.

2. PRELIMINARIES
2.1 Web Graph Model:
The web can be modeled as a directed graph \( G = \{V, E\} \) whose nodes correspond to
static pages (\( V \)) on the web, and whose edges correspond to hyperlinks (\( E \)) between
these pages. The web graph (\( G \)) is massive containing billions of nodes and edges.
In addition, \( G \) is dynamic or evolving, with nodes and edges appearing and
disappearing over time. One study [10] models the structure of the web as a Bowtie
structure. In this model, the majority of the web pages are a strongly connected
graph. Mathematically, the graph structure can be encoded as a matrix Eq (i) where
\[
G[i, j] = \begin{cases} 
1 & \text{if } i \text{ connects to } j \\
0 & \text{otherwise} 
\end{cases}
\]  

In addition, transition matrix (\( T \)) Eq (ii) and inverse transition matrix (\( I \)) Eq (iii)
captures the outdegree and indegree of the web graph and they can be defined as:
Transition Matrix.
\[
T[i, j] = 1/\text{outdegree (}j\text{)} \quad \text{if } j \text{ connects to } I 
\]
Inverse Transition Matrix:
\[
I[i, j] = \begin{cases} 
1/\text{indegree}(j) & \text{if } i \text{ connects to } j \\
0 & \text{if } i \text{ do not connect } j
\end{cases}
\] (iii)

2.2 Biased Page Rank:
Page Rank [11] is one of the most popular link based methods to determine a page’s global relevance or importance. Page rank assigns an importance score (page rank) proportional to the importance of other web pages which point to it. Page rank \( r \) is defined as the first eigenvector of the matrix \( A \) where \( A \) is defined as follow:
\[
A_{ij} = \beta T_{ij} + (1 - \beta)/N
\] (iv)

where \( T \) is the transition matrix, \( N \) is the total number of web pages and \( \beta \) is a decay factor and \( 0 < \beta < 1 \). While page rank assigns a score proportional to generic popularity of a page, biased page rank or topic-specific page rank [10] measures the popularity within a topic or domain. Here the equivalent random surfer model is as follows. When the random surfer teleports, he picks a page from a set \( S \) of web pages which is called the teleport set. The set \( S \) only contains pages that are relevant to the topic. Corresponding to each teleport set \( S \), we get a different rank vector. In matrix Eq (v) representation:
\[
A_{ij} = \begin{cases} 
\beta T_{ij} + (1 - \beta)/|S| & \text{if } i \text{ to } S \\
\beta T_{ij} & \text{otherwise}
\end{cases}
\] (v)

where \( A \) is a stochastic matrix as before. Here, we have weight all pages in the teleport set \( S \) equally, but we could weight them differently if we wish.

2.3 Trust Rank:
The TrustRank algorithm [12] is a semi-automatic algorithm designed for spam detection and PageRank demotion. It starts with a seed of hand-picked trusted pages and then launches the PageRank algorithm with a teleportation vector with nonzero entries only on this seed of trusted pages. Thus the initial trust score of trusted pages will propagate through hyperlinks. The fundamental idea is that trusted pages link to trusted pages and spam pages are linked to by spam pages [7]. Each coordinate of the vector obtained this way gives the TrustRank score of the associated page. Pages with a high TrustRank score are trusted and considered to be non spam, pages with a small Trustrank score are untrusted and considered to be spam. The boundary between trusted and untrusted pages is an arbitrary threshold in TrustRank.
3. EXPERIMENT

3.1 DataSet:
Seed set selection is the most important component of the TrustRank algorithm. Different seed selections lead to different results. The seed selection process employed by the authors of TrustRank may not guarantee a broad coverage of the Web. We perform experiments on a WEBSPAM-UK2007 database that contains partial set of pages crawled by search engine. The base data is a set of 105,896,555 pages in 114,529 hosts in the .UK domain.

3.2 Result - Number of Seeds:
We start from 50 seeds and double the number each time. At each point, we randomly select 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 level at the case of 3200.

<table>
<thead>
<tr>
<th>Number of seeds</th>
<th>Number of seeds in top 100 results</th>
<th>Number of seeds in top 1000 results</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>50</td>
<td>50</td>
</tr>
<tr>
<td>100</td>
<td>95.5</td>
<td>100</td>
</tr>
<tr>
<td>200</td>
<td>92</td>
<td>200</td>
</tr>
<tr>
<td>400</td>
<td>78.75</td>
<td>400</td>
</tr>
<tr>
<td>800</td>
<td>49.25</td>
<td>800</td>
</tr>
<tr>
<td>1600</td>
<td>21.75</td>
<td>812</td>
</tr>
<tr>
<td>3200</td>
<td>15.58</td>
<td>7.25</td>
</tr>
<tr>
<td>6400</td>
<td>18.25</td>
<td>419.25</td>
</tr>
<tr>
<td>12800</td>
<td>33.5</td>
<td>428.5</td>
</tr>
<tr>
<td>25600</td>
<td>34.35</td>
<td>501.01</td>
</tr>
</tbody>
</table>

Table1: Result for seed selection

To explore this trend more precisely, 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.
3.3 Combat 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.
To compare the anti-spamming abilities of different seed sets, we select a small seed set $X$ using a method similar to TrustRank [1]. Good sites (reputable and directory) with high rankings have little demotion, i.e. retain high ranking values. There is no 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 $L$ is much better than using $X$. The demotions are always larger than 5.8 with $L$.

4. CONCLUSION

As the web grows in size and value, search engines play an increasingly critical role, allowing users to find information of interest. However, today’s search engines are seriously threatened by malicious web spam that attempts to subvert the unbiased searching and ranking services provided by the engines. In this paper, we reveal that a large seed set can achieve a better performance than a small seed set on detecting web spam. 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.

REFERENCES


ABOUT AUTHORS

1. **Jyoti Pruthi** is pursuing PhD (computer science) from Singhania University, Rajasthan. She received her MPhil (computer science) degree in 2008. She holds an MCA from MDU, Rohtak. She is currently working as Assistant Professor, MCA Department with Manav Rachna College of Engineering, Faridabad. Earlier, she worked with CSSL, Gurgaon as a Software Engineer. Her area of specialization includes Search Engine, Spamming & Anti-Spamming Techniques.

2. **Dr Ela Kumar** PhD (Artificial Intelligence) from University of Delhi. She holds an M.Tech (CSE) and B.E (ECE) from IIT, Roorkee. She is currently working as Associate Professor with Gautam Buddha University, Gr Noida. Her area of specialization includes Search Engine, AI, Ad-Hoc Networks. She holds to her credit around 19 Years of experience of teaching and research. Besides publishing 30 research papers in journals of repute, she has written 3 books.