Web Page Ranking using Link Attributes [*]

Ricardo Baeza-Yates Emilio Davis
Center for Web Research, CS Dept.
Universidad de Chile
Blanco Encalada 2120, Santiago, Chile
{rbaeza,edavis}@dcc.uchile.cl

ABSTRACT

We present a variant of PageRank, WLRank, that considers different Web page attributes to give more weight to some links, improving the precision of the answers.

Categories and Subject Descriptors

H.3.5 [**Information Systems**]: On-line Infor. Systems

General Terms

Algoritmhs, experimentation.

Keywords

Web link ranking, PageRank.

1. INTRODUCTION

Nowadays all search engines use some kind of Web page link-based ranking in their ranking algorithms. Without doubt, this has been the result of the success of Google, and its PageRank link algorithm [1]. A taxonomy of different link ranking algorithms is presented in [2].

In all published link ranking algorithms, all links have the same importance. However, web page developers give more importance to some links using different HTML tags, because some Web resources are more important than others. Hence, a link ranking technique that gives different weights to links may improve over uniform weight links.

In this work we present a variant of PageRank that gives weights to link based on three attributes: relative position in the page, tag where the link is contained, and length of the anchor text. Our results show that our algorithm, WLRank, improves over PageRank.

2. PAGERANK

The idea behind PageRank is that good pages reference good pages. Hence, pages that are referenced by good pages have higher PageRank. Although there are several formulations of PageRank, we use the random surf metaphor. Suppose that you are a user surfing the Web in a random fashion, such that, if you are in a page, with certain probability you get bored and leave the page, or you choose uniformly at random to follow one of the links on the page where you are (removing self links). Hence, the probability of being in page p is

$$PR(p) = \frac{q}{T} + (1-q)\sum_i \frac{PR(r_i)}{L(r_i)}$$

where T is the total number of pages, q is the probability of leaving page p (in the original work $q = 0.15$ is suggested), r_i are the pages that point to page p, and $L(r_i)$ is the number of links in page r_i. These values can then be used as page ranking, and can be computed by an iterative algorithm converging quite fast, as we are interested in the ranking order rather than the actual ranking values. The term q is called *damping factor* as decreases exponentially link spamming based in sequences of links that return to a page.

3. OUR VARIANT

WLRank (Weighted Links Rank) assigns the ranking value $R(i)$ to page i using the following equations:

$$R(i) = \frac{q}{T} + (1-q)\sum_j \frac{W(j,i)R(j)}{\sum_k W(j,k)} \ ,$$

$$W(j,i) = L(j,i)(c + T(j,i) + AL(j,i) + RP(j,i)) \ ,$$

where given a link from page j to page i we have:

- $L(j,i)$ is 1 if the link exists, or 0 otherwise, and c is a constant that gives a base weight to every link,
- $T(j,i)$ is a value that depends on the *tag* where the link is inserted,
- $AL(j,i)$ is the length of the anchor text of the link divided by a constant d that depends that estimates the average anchor text length in characters, and
- $RP(j,i)$ is the relative position of the link in the page weighted by a constant b.

As in PageRank, $R(i)$ corresponds to the probability to reach page i while surfing the Web. If $W(j,i) = L(j,i)$ we have the original PageRank. The changes are explained below.

The term $T(j,i)$ is a sequence of constants depending on the tag where the link is. For example, if the link is inside a <h1> tag, will have a high $T(j,i)$ value, a little less for <h2>, etc. The same for others emphasis tags like <strong> or <b>.

[*] This work was partially supported by Fondecyt Grant 1020803.

WWW2004, May 17–22, 2004, New York, New York, USA.
ACM 1-58113-912-8/04/0005.

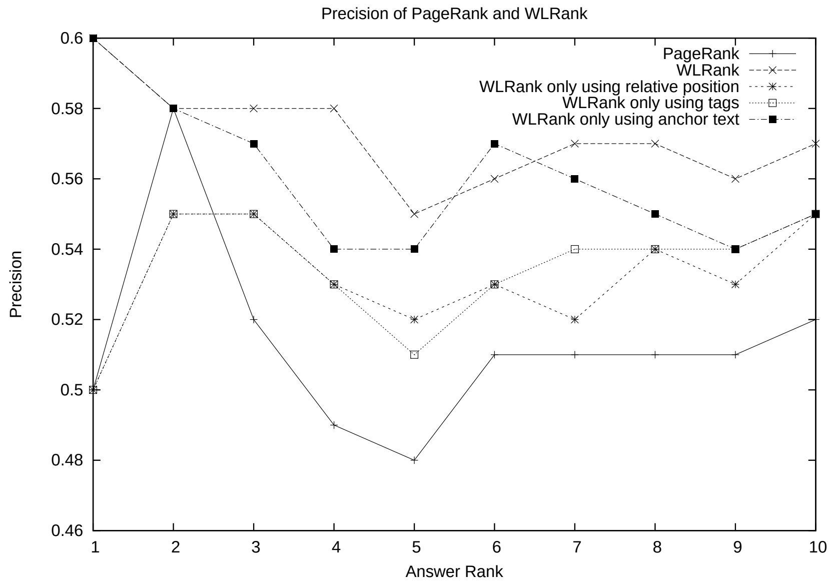

Figure 1: Precision of PageRank and WLRank (and all the individual factors).

The term $AL(j,i)$ gives more value to links where the creator explained in more detail what Web resource is being linked. For example, this gives less weight to links described with *home* or *here*.

Finally, the term $RP(j,i)$ gives more weight to links that are at the beginning of the page rather that at the end of the page (physically in the HTML code, not necessarily in the browser view).

4. EVALUATION

To test WLRank we used a crawling over the .CL domain of 460 thousand pages, and several users that provided binary relevance judgments on the first 10 answers for several queries. Our users came from different backgrounds to emulate what happens in practice in search engines.

For the test we used WLRank with $c = 1$, $b = 1$, and $d = 100$. We also considered unit weights for only two tags: <b> y <h1>. In addition to WLRank, we computed the effect of each term alone as well as PageRank, all of them over the same Web collection to have a valid comparison.

Each user was requested to pose two or three query to the system for all the cases (that is at most 150 relevance judgments per user) without knowing which ranking was being used in each case. The assumption was that each user would be an expert on the selected query. Due to lack of space we cannot include all the queries and judgments, but the queries ranged from generic terms such as *education* or *computing* to specific ones like *aspirin* or *mozilla*.

Using the judgments of a total of 20 queries, we computed the precision on the first k answers. That is, precision is the number of relevant answers over the number of answers considered, obtaining the results shown in figure 1. From the graph we can see that the most effective attribute is anchor text length, and that all of them improve upon PageRank which uses uniform link weights.

One way to compare how better is WLRank with respect to PageRank is using a *perfect* ranking, which only gives relevant results. Table 1 shows the total error with respect to a perfect ranking for the first k answers up to 10. We can see that WLRank improves PageRank precision a 13% on average per answer for the first 10 answers.

Table 1: Comparison of PageRank and WLRank against a perfect ranking.

Answer	Perfect - PageRank	Perfect - WLRank
1	0.5	0.4
2	0.43	0.43
3	0.48	0.42
4	0.51	0.43
5	0.52	0.45
6	0.49	0.44
7	0.49	0.43
8	0.49	0.43
9	0.49	0.44
10	0.49	0.44
Total error	**4.89**	**4.29**

5. CONCLUSIONS

Our results show that using weighted links can improve the precision of search engines. The best attribute seems to be anchor text length, but others can be better. On the other hand, relative position was not so effective, indicating that the logical position not always matches the physical position. Future work includes tuning the weight factors for each term and further user evaluation.

6. REFERENCES

[1] S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine. In *7th WWW Conference*, Brisbane, Australia, April 1998.

[2] C. Ding, X. He, P. Husbands, H. Zha, and H. Simon. Pagerank, hits and a unified framework for link analysis. *LBNL Tech Report 49372*, 2001-2002.