# The PowerRank Web Link Analysis Algorithm

Yizhou Lu[1], Benyu Zhang[2], Wensi Xi[3], Zheng Chen[2], Yi Liu[4], Michael R. Lyu[4], Wei-Ying Ma[2]

[1]School of Mathematical Sciences, Peking University, Beijing 100871, P.R. China
luyizhou@pku.edu.cn

[2]Microsoft Research Asia, 49 Zhichun Road, Beijing 100080, P.R. China
{byzhang, zhengc, wyma}@microsoft.com

[3] Virginia Polytechnic Institute and State University, Blacksburg, VA. 24060 U.S.A.
xwensi@vt.edu

[4]Computer Science and Engineering Department, the Chinese University of Hong Kong.
{yliu,lyu}@cse.cuhk.edu.hk

## ABSTRACT

The web graph follows the power law distribution and has a hierarchy structure. But neither the PageRank algorithm nor any of its improvements leverage these attributes. In this paper, we propose a novel link analysis algorithm "the PowerRank algorithm", which makes use of the power law distribution attribute and the hierarchy structure of the web graph. The algorithm consists two parts. In the first part, special treatment is applied to the web pages with low "importance" score. In the second part, the global "importance" score for each web page is obtained by combining those scores together. Our experimental results show that: 1) The PowerRank algorithm computes 10%~30% faster than PageRank algorithm. 2) Top web pages in PowerRank algorithm remain similar to that of the PageRank algorithm.

## Categories and Subject Descriptors

I.m [**Computing Methodologies**]: MISCELLANEOUS; G.2.2 [**Discrete Mathematics**]: Graph theory

## General Terms

Algorithms, Performance, Experimentation

## Keywords

Page Rank Algorithm, Power Distribution, Hierarchy Structure

## 1. Introduction

The PageRank algorithm [2] was developed to determine the importance of a web page by calculating the principle eigenvector of the web page's adjacency matrix. Due to the huge amount of pages in the Web, PageRank computing becomes a very time consuming job. Researchers have developed many technologies to speed up the page rank computation from different prospectives. Kamyar et, al, [4] exploits block structures of web to enhance PageRank computations. They computed local rank of pages of different domains separately and merged them together as a new starting value of PageRank computation. Arasu [1] suggested applying methods such as Multi-grid, Gauss-Seidel, successive overrelaxation method (SOR) on the web adjacency matrix to obtain the rank score vector. However, none of the tasks discussed above explored the hierarchy structure and the power law distribution of the web graph to help improve the calculation of PageRank. In this paper, we propose a novel link analysis algorithm which combines both above attributes to improve the page rank computation by

- Reducing the computational complexity;
- Keeping the order of top ranked pages similar to that of the PageRank algorithm.
- Reducing the probability of mult pages having same score.

## 2. Web graphs and its power law distribution

The web graph is proved to follow the power law distribution [3]. Using web page in-degree as an example, the power law distribution can be explained as "the number of web pages with in-degree k is proportional to $k^{-\beta}$". In the web graph that follows the power law distribution, the "importance" value of a webpage flows more easily from the low in-degree pages to the high in-degree ones. This leads to a larger conditional expectation value of the "importance" score, if the page has high in-degree. Mathematically, we write it as the following theorems.

**Theorem:** 1) For the $i_{\text{th}}$ node of a graph, $r_i$ is its "popularity" score form PageRank, and *InDi* is its in-degree. n is the number of nodes in the graph, and *M*, $N_i$ are constants (see 2) ). The conditional expectation value of its "popularity" score satisfies:

$$E(r_i \mid InD_i > M) > \frac{MN_i}{20n}.$$

2) Suppose P denotes the markov matrix of the web, $x^k$ denotes the score vector of *k*th iteration in PageRank computation, $k_{threshold}$ denotes the maximum of iteration, *j* denotes the *j*th node. Then,

$$N_i = \Pr(P_{ji} > \frac{1}{2}) \bullet \min_k \{\Pr(x_j^k > \frac{1}{10n})\},\ k{=}1,\ldots,k_{threshold}.$$

"Pr($P_{ji}$>1/2)" means that how many pages linked to $i_{\text{th}}$ page have low out-degree. "min$_k${Pr( $x_j^k > \frac{1}{10n}$ )}" means that how many pages have average "popularity" score in the final score vector. Due to the power-law distribution, $N_i$ is large. The value of $\frac{MN_i}{20n}$ can be larger than average "popularity" score (1/n).

## 3. The PowerRank Algorithm

As the previous theorem revealed, high in-degree pages have higher expectation for the "importance" score. From the result, we can deduce that low "importance" web pages are expected to have low in-degree. If we take a special treatment on these pages, for example, cutting them off from the web graph, the web graph link structure will remain similar as before. Such treatment would then reduce the computing time and preserve the similar rank result.

To identify the low-ranked pages, we first rank the hosts or domain nodes of the web by their in-degree. Then we cut off the low in-degree hosts or domains. Pages located in such nodes (hosts or domains) are also cut off. The remaining nodes are continued to the next level of calculation for "importance". Finally, those "popularity" scores for pages remained in the calculation, or the pages cut off from the calculation are combined

WWW 2004, May 17-22, 2004, New York, NY USA.
ACM 1-58113-844-X/04/0005.

to form a global “importance” score. We name this method “PowerRank” algorithm. It is described in detail in the following:

Suppose there are only three levels of web hierarchies: domain, host and Webpage. Suppose the Page URL is http://www.acm.org/index.html, its host URL is www.acm.org, and its Domain URL is acm.org. The PowerRank algorithm contains four steps:

- *First, PageRank algorithm is applied on domains. After several iterations, the low-ranked domains are cut off.*
- *Second, PageRank is applied on hosts. Similar to the first step, after several iterations, the low-ranked hosts are cut off.*
- *Third, a similar calculation is applied on web pages, and low-ranked pages are cut off. By our theorem, the structure of the remaining graph should be similar to that of the original web graph. Applying a ranking algorithm here will obtain a similarity rank order and save computing time.*
- *Finally, the global “importance” scores of the pages in the cut-off hosts (domains) are calculated by multiplying their local PageRank scores with the scores of their nested hosts.*

The advantages of PowerRank algorithm are:

**Advantage 1** The PowerRank algorithm introduces a new framework of computing the “importance” of the web pages. In this algorithm, the time complexity is significantly reduced compared to the traditional PageRank algorithm.

**Advantage 2** Our algorithm can be applied to any improvements on the PageRank algorithm that we introduced in the first section.

# 4. Experiments

## 4.1 Experiment Setup

We use a Web Page sub-graph of the web (denoted as AE) to conduct our experiments. AE contains about 88 Million pages, 4.4 Million hosts, and 3.3 Million domains.

To study the effects of different cut-off criteria, we conduct six experiments. These experiments are represented with three percentage numbers: Different domains and host cutting-off ratios are indicated in the first and second percentage numbers, while the third number indicates the percent of web pages left. These six experiments are 2%-2%-96%, 4%-6%-90%, 8%-12%-80%, 10%-20%-70%, 15%-25%-60%, 20%-30%-50%, which are labeled as AEI, AEII, AEIII, AEIV, AEV and AEVI, respectively.

## 4.2 Results

We compared the six experiments’ score vectors and the PageRank vector.

**Rank Vectors Comparison** We record the L1-norm[⊥] for the minus vector of original PageRank score vector and our score vectors. The comparison results are shown in Table 2. The L1-norm ranges from 0.11 to 0.63.

**Table 2: L1-norm of minus vectors under different threshold**

| Experiment Label | L1-norm of minus vector |
|---|---|
| AEI | 0.112598434090614 |
| AEII | 0.139604702591896 |
| AEIII | 0.246725514531136 |
| AEIV | 0.420685678720474 |
| AEV | 0.526982128620148 |
| AEVI | 0.634842157363892 |

**Time Comparison** We compare the computing time of AEI, AEIII and AEIV with the original PageRank algorithm. Results are shown in Figure 1. The dotted bar is the ratio of remaining pages’ number to the total number of pages’. The stripped bar is the ratio of remaining links’ number to the total number of links. The squared bar is the ratio of computing time to the original PageRank computing time.

From Figure 1 we can find that the computing time of the PowerRank algorithm is less than that of the PageRank algorithm. In AEIII, the computation time reduced about 20%.

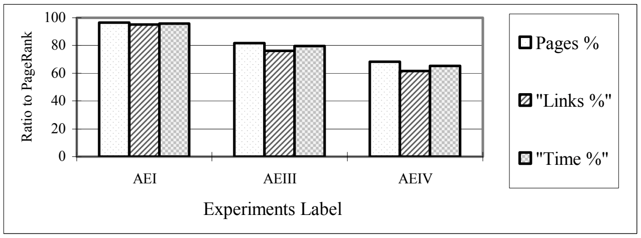

**Figure 1: Nodes, links and time ratio in different experiments**

**High Rank Score Comparison** We also compare the top 5% pages’ PowerRank score and PageRank score. In their log-log plots, we found that they fit the line $y = x$ well. That means the PowerRank provides a similar ranking result to PageRank algorithm for top-ranked pages.

**Low Rank Score Comparison** Two pages with the same “importance” value in the final calculation results are called a *tie*. We compare the tie counts of PowerRank result and PageRank result in the lower 80% pages. The tie counts in AEIV, AEV, and AEVI are less than that of PageRank. The tie counts of lower 15% pages in AEII, AEIII are less than that of PageRank as well.

# 5. Conclusion

In this paper, we introduce the “PowerRank” algorithm, which takes the advantage of the power law distribution attribute and the hierarchy structure of the web graph. Our experimental results on this algorithm show that the PowerRank algorithm computes 10%~30% faster than the PageRank algorithm. Moreover, the top ranked web pages in PowerRank algorithm remain very similar to those of the PageRank algorithm. Finally, it reduces the score “ties” in the final calculation results.

# 6. Acknowledgement

The work was done in Microsoft Research Asia; and the authors from CUHK were partially supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. CUHK4182/03E)

# 7. REFERENCES

[1] A. Arasu. PageRank Computation and the Structure of the Web: Experiments and Algorithms. 11th International WWW Conference, May 2002.

[2] S. Brin, L. Page, R. Motwami, and T. Winograd. The PageRank citation ranking: bringing order to the web. Stanford University Technical Report, 1998.

[3] M. Faloutsos, P. Faloutsos, C. Faloutsos. On Power-Law Relationships of the Internet Topology. Proceedings of ACM SIGCOMM, Aug. 1999.

[4] S. D. Kamvar, T. H. Haveliwala, C. D. Manning and G. H. Golub. Exploiting the Block Structure of the Web for Computing. Stanford University Technical Report, 2003.

[⊥] L1-norm $\| \vec{x} \|_1$ of vector $\vec{x} = (x_1, x_2, \ldots, x_n)$ is defined as: $\left\|\vec{x}\right\|_1 = \sum_{i=1}^{n} \left|x_i\right|$