PageRank, HITS and a Unified Framework for Link Analysis*

Chris Ding[a], Xiaofeng He[a], Parry Husbands[a], Hongyuan Zha[b], Horst D. Simon[a]
[a]Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA
[b]Pennsylvania State University, University Park, PA 16802, USA
{chqding,xhe,pjrhusbands,hdsimon}@lbl.gov, zha@cse.psu.edu

ABSTRACT

Two popular link-based webpage ranking algorithms are (i) PageRank[1] and (ii) HITS (Hypertext Induced Topic Selection)[3]. HITS makes the crucial distinction of *hubs* and *authorities* and computes them in a mutually reinforcing way. PageRank considers the hyperlink *weight normalization* and the equilibrium distribution of *random surfers* as the citation score. We generalize and combine these key concepts into a unified framework, in which we prove that rankings produced by PageRank and HITS are both highly correlated with the ranking by in-degree and out-degree.

Categories and Subject Descriptors

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

General Terms

Algorithms, Experimentation

1. HITS ALGORITHM

In the HITS algorithm[3], each webpage i has both a hub score y_i and an authority score x_i. The intuition is that a good *authority* is pointed to by many good *hubs* (this defines the $\mathcal{I}^{op}$ operation) and a good *hub* points to many good *authorities* (this defines the $\mathcal{O}^{op}$ operation). This mutually reinforcing relationship can be represented as the following general operations,

$$\mathbf{x} = \mathcal{I}^{op}(\mathbf{y}), \quad \mathbf{y} = \mathcal{O}^{op}(\mathbf{x}). \tag{1}$$

Here vectors $\mathbf{x} = (x_1, \cdots, x_n)^T$ and $\mathbf{y} = (y_1, \cdots, y_n)^T$ contain the authority score and hub score of each webpage, respectively. $\mathcal{I}^{op}$ and $\mathcal{O}^{op}$ can be written as

$$\mathcal{I}^{op}(\cdot) = L^T, \quad \mathcal{O}^{op}(\cdot) = L. \tag{2}$$

L is the adjacency matrix of the web graph. Final scores are obtained at convergence through the iterations,

$$\begin{aligned} \mathbf{x}^{(t+1)} &= \mathcal{I}^{op}(\mathcal{O}^{op}(\mathbf{x}^{(t)}) = L^T L \mathbf{x}^{(t)} \\ \mathbf{y}^{(t+1)} &= \mathcal{O}^{op}(\mathcal{I}^{op}(\mathbf{y}^{(t)}) = L L^T \mathbf{y}^{(t)} \end{aligned} \tag{3}$$

*LBNL-50007. Nov. 2001. The full paper is available online: `www.nersc.gov/~cding/papers/#web`.

SIGIR'02, August 11-15, 2002, Tampere, Finland.
ACM 1-58113-561-0/02/0008.

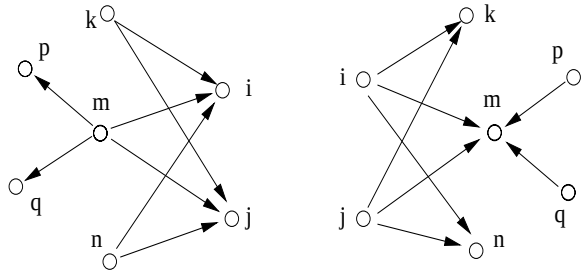

Figure 1: Importance of hyperlink weight normalization. Left: webpages i, j are co-cited by webpages k, m, n. However, since webpage m also cites other webpages p, q, the co-citation of i, j by m is not as significant as the co-citation by either k or n. This fact can be compensated by normalizing the weights on the out-bound links of a webpage; now the co-citation by m is only 2/4=50% as important as the co-citation by either k or n.
Right: webpages i, j co-reference webpages k, n, m. However, since webpage m is also referenced by other webpages p, q, the co-reference of i, j to m is not as significant as the co-reference to either k or n. This fact can be compensated by normalizing the weights on the in-bound links of a webpage.

where $\mathbf{x}^{(t)}, \mathbf{y}^{(t)}$ denote scores at the t-th iteration.

Co-citation and co-reference

We emphasize the role of co-citation and co-reference(Fig.1). Since $L^T L$ determines the authority ranking, we call $L^T L$ the *authority matrix*. We prove that $L^T L = D_{in} + C$, where D_{in} is the diagonal matrix containing in-degrees of all nodes, and C is the co-citation matrix. This shows the close relationship between authority and co-citation.

Since $L L^T$ determines the hub scores, we call $L L^T$ the *hub matrix*. We prove that $L L^T = D_{out} + R$, where D_{out} is the diagonal matrix containing out-degrees of all nodes, and R is the co-reference matrix. This shows the close relationship between hubs and co-references.

Assuming the web graph is a fixed degree sequence random graph, HITS results in average case can be solved in closed form [2], which proves that authority ranking by HITS is *identical* to the ranking by in-degrees. Similarly, hub ranking in HITS is identical to the ranking by out-degrees.

Scheme	$\mathcal{I}^{op}$	$\mathcal{O}^{op}$
HITS	L^T	L
PageRank	$L^T D_{out}^{-1}$	$L D_{in}^{-1}$
Auth-Rank	$L^T D_{out}^{-1}$	L
Hub-Rank	L^T	$L D_{in}^{-1}$
Sym-Rank	$D_{in}^{-1/2} L^T D_{out}^{-1/2}$	$D_{out}^{-1/2} L D_{in}^{-1/2}$

Table 1: $\mathcal{I}^{op}$ and $\mathcal{O}^{op}$ operations for HITS, PageRank, Auth-Rank, Hub-Rank, and Sym-Rank.

2. PAGERANK

The key feature of PageRank is the hyperlink weight normalization, as shown in Fig.1 from the perspective of co-citation and co-reference. We may state this as *Internet Democracy*: each website (webpage) has a total of one vote. Another key feature is that PageRank adopts a web surfing model based on a Markov process in determining the scores:

$$\mathbf{x} = \mathcal{I}^{op}(\mathbf{x}),\ \mathcal{I}^{op}(\cdot) = L^T D_{out}^{-1}.$$

The equilibrium distribution of random surfers on webpages is a measure of a webpage's "importance", is the authority score in PageRank.

Hubs in PageRank

We generalize the weight normalization idea to in-bound hyperlinks. One reason is illustrated in Fig.1. Co-reference to a webpage with a large in-degree is not as significant as co-reference to a webpage with a small in-degree. For example, the fact that we all make reference to a highly referenced site such as *New York Times* says little about whether we are similar. But if two persons make reference to Knuth's *The Art of Computer Programming*, it is likely that both persons are interested in computer algorithms.

We propose to define hub in PageRank using the same random surfer model as in definition of authority. The hub scores are obtained through

$$\mathbf{y} = \mathcal{O}^{op}(\mathbf{y}),\ \mathcal{O}^{op}(\cdot) = L D_{in}^{-1} \quad (4)$$

3. A UNIFIED FRAMEWORK

The most important feature of HITS is the mutual reinforcement between hubs and authorities, while the most important feature of PageRank is the hyperlink weight normalization. These features are summarized in Table 1. They can be generalized and combined. We also clarify and formalize *weight propagation* and *random surfing* as two different but related method to compute ranking scores. All these form a unified framework for link analysis.

In this framework, one can easily design new ranking algorithms. We study three new ranking algorithms: the Auth-Rank, the Hub-Rank and the Sym-Rank. Their $\mathcal{I}^{op}, \mathcal{O}^{op}$ operations are defined in Table 1. All these three rankings combine both features of HITS and PageRank, thus they are expected to be somewhere between the rankings produced by HITS and PageRank. The most important results are: all three rankings can be solved in closed-form. The authority rankings of Auth-Rank and Sym-Rank are identical to the ranking by indegrees. The hub rankings of Hub-Rank and Sym-Rank are identical to the ranking by outdegrees. We therefore conclude that even though PageRank and HITS use different methods to compute the link-based ranking, their final rankings will correlate highly with rankings by indegree or outdegree. HITS ranking and PageRank ranking are very similar, too.

Experiment 1. This dateset was supplied by the Internet Archive and was extracted from a crawl performed over 1998-1999. It has 4,906,214 websites and represents a site-level graph of the Web. Rankings are shown below.

```
Authority Ranking
Hits InDgr Page  URL
  1    4    6    www.yahoo.com
  2    3    3    www.geocities.com
  3    1    1    www.microsoft.com
  4    6    5    members.aol.com
  5    2    2    home.netscape.com
  6   10   12    www.excite.com
  7   11   15    www.lycos.com
  8    9    9    members.tripod.com
  9   15   11    ourworld.compuserve.com
 10    5    7    www.netscape.com
 11   20   25    www.cnn.com
 12   28   22    www.webcom.com
 13   33   20    sunsite.unc.edu
 14    7    4    www.adobe.com
 15   35   24    www.teleport.com
 16   17   26    www.altavista.digital.com
 17   25   16    www.w3.org
 18   19   28    www.infoseek.com
 19   18   19    www.angelfire.com
 20   21   34    www.hotbot.com
```

Experiment 2. This dataset is about the topic *Running* which contains a total of 13152 webpages. This is a subcategory of a larger category *Fitness* from the Open Directory Project (*www.dmoz.org*). Rankings are shown below.

```
Authority Ranking
Hits InDgr Page   URL
 1    2    1    www.runnersworld.com/
 2    5    5    sunsite.unc.edu/drears/running/
 3    4    2    www.usatf.org/
 4    1    3    www.coolrunning.com/
 5    6    6    www.clark.net/pub/pribut/spsport
 6    8    9    www.runningnetwork.com/
 7    9    8    www.iaaf.org/
 8   14   20    www.sirius.ca/running.html
 9   12   12    www.wimsey.com/~dblaikie/
10   15   17    www.kicksports.com/
11    7    7    www.nyrrc.org/
12   18   14    www.usaldr.org/
13   20   32    www.halhigdon.com/
14   25   30    www.ontherun.com/
15   10   89    www.runningroom.com/
16   23   31    www.webrunner.com/webrun/running
17   22   23    www.doitsports.com/
18   21   79    www.arfa.org/
19   19   16    www.adidas.com/
20   11   13    www.uta.fi/~csmipe/sport/
```

In both datasets, HITS ranking and PageRank ranking are highly correlated with indegree ranking.

4. REFERENCES

[1] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. *Proc. of 7th WWW Conferece*, 1998.

[2] C. Ding, H. Zha, X. He, P. Husbands, and H. Simon. Analysis of hubs and authorities on the web. *Lawrence Berkeley Nat'l Lab Tech Report* 47847, May 2001.

[3] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *J. ACM*, 48:604–632, 1999.