PageRank, HITS and a Unified Framework for Link Analysis *

Chris Ding†, Xiaofeng He†, Parry Husbands†, Hongyuan Zha‡, Horst Simon†

Abstract

Two popular webpage ranking algorithms are HITS and PageRank. HITS emphasizes *mutual reinforcement* between *authority* and *hub* webpages, while PageRank emphasizes hyperlink *weight normalization* and *web surfing* based on random walk models. We systematically generalize/combine these concepts into a unified framework. The ranking framework contains a large algorithm space; HITS and PageRank are two extreme ends in this space. We study several normalized ranking algorithms which are intermediate between HITS and PageRank, and obtain closed-form solutions. We show that, to first order approximation, all ranking algorithms in this framework, including PageRank and HITS, lead to same ranking which is highly correlated with ranking by indegree.

1 Introduction

Two most popular ranking algorithms are (i) the PageRank algorithm developed by Brin and Page [3] and used in the search engine Google, and (ii) the HITS (Hypertext Induced Topic Selection) algorithm developed by Kleinberg[5]. HITS makes the distinction between *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. There are a number of further extensions and developments [1, 6, 2].

We generalize the key concepts of mutual reinforcement and hyperlink weight normalization into a unified framework. We clarify and formalize the notion of similarity mediated score propagation and random surfing score propagation schemes. In this unified framework, new extensions of HITS or PageRank can be easily designed and analyzed (Table 1 captures the main results). We analyze three new extensions, the out-link normalized ranking (OnormRank), the in-link normalized ranking (InormRank), and symmetric normalized ranking (SnormRank).

All three new ranking algorithms have closed-form solutions. The authorities in OnormRank using random surfing score propagation (see §5.2) are precisely given by node indegrees (see Eq.5.13), and similar results for hub ranking in InormRank, Using similarity mediated score propagation (see §5.1), authorities scores are precisely given by square root of indegrees (Eq.7.14) and hub scores are given by square root of outdegrees (see Eq.7.15). By construction, all three new ranking algorithms combine mutual re-inforcement with hyperlink weight normalization; therefore their rankings are close to the rankings produced by HITS and PageRank. From these, we conclude that both HITS and PageRank authority rankings have high correlation with the ranking by indegree. The difference between rankings produced by different algorithms reflects slightly different but useful emphasis. These results provide theoretical basis for the general intuition that in web ranking, indegree and outdegree are of fundamental importance.

2 HITS Algorithm

In the HITS algorithm[5], each webpage p_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,

$$\mathbf{x} = \mathcal{I}^{op}(\mathbf{y}), \quad \mathbf{y} = \mathcal{O}^{op}(\mathbf{x}). \tag{2.1}$$

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. The mutual reinforcement operations $\mathcal{I}^{op}$ and $\mathcal{O}^{op}$ in HITS can be written in the following matrix representations

$$\mathcal{I}^{op}(\cdot) = L^T, \quad \mathcal{O}^{op}(\cdot) = L. \tag{2.2}$$

where L is the adjacency matrix of the directed web graph. The final authority and hub scores of every webpage can be obtained through an iteratively updating

*The full length version of this paper is available from http://www.nersc.gov/~cding/papers.

†NERSC Division, Lawrence Berkeley National Laboratory, University of California, Berkeley, CA 94720, {chqding,xhe,pjrhusbands,hdsimon}@lbl.gov. Supported by U.S. Department of Energy under contract DE-AC03-76SF00098.

‡Department of Computer Science and Engineering, Pennsylvania State University, University Park, PA 16802. zha@cse.psu.edu.

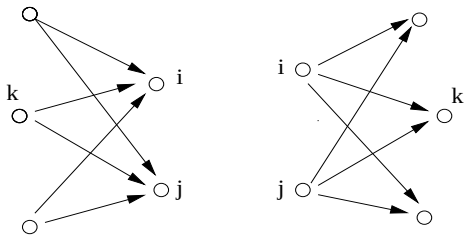

Figure 1: Left: webpages p_i, p_j are co-cited by webpage p_k. Right: webpages p_i, p_j co-reference webpage p_k.

process. If we use $\mathbf{x}^{(t)}, \mathbf{y}^{(t)}$ to denote authority and hub scores at the t^{th} iteration, the iterative processes to reach the final solutions are

$$\begin{aligned} \mathbf{x}^{(t+1)} &= \mathcal{I}^{op}(\mathcal{O}^{op}(\mathbf{x}^{(t)}) = L^T L \mathbf{x}^{(t)} \\ \mathbf{y}^{(t+1)} &= \mathcal{O}^{op}(\mathcal{I}^{op}(\mathbf{y}^{(t)}) = L L^T \mathbf{y}^{(t)} \end{aligned} \tag{2.3}$$

Since $L^T L$ determines the authority ranking, we call $L^T L$ the *authority matrix*. Similarly, we call LL^T the *hub matrix*. The final solutions $\mathbf{x}^*, \mathbf{y}^*$ are the principal eigenvectors of $L^T L$ and LL^T, which are the singular value decomposition of L. In practical applications, a modification of HITS [1] by suppressing the contribution from different webpages from same host (site or root in URL) is often adopted.

2.1 Co-citation and co-reference. The authority and hub matrices have interesting connections [5] to co-citation and co-reference in the fields of citation analysis.

If two distinct webpages p_i, p_j are co-cited by many other webpages p_k as in Fig.1, p_i, p_j are likely to be related in some sense. Thus co-citation C_{ij} is a similarity measure, defined as the number of webpages that co-cite p_i, p_j. One can show the authority matrix $L^T L$ is

$$L^T L = D_{in} + C,$$

where $D_{in} = \text{diag}(\mathbf{d}_{\text{in}})$. and $\mathbf{d}_{\text{in}} = (b_1, \cdots, b_n)^T$ is a vector of in-degrees. Thus $L^T L$ is the sum of co-citation and indegree [4]. This shows the close relationship between authority and co-citation.

The fact that two distinct webpages p_i, p_j co-reference several other webpages p_k (right panel in Fig. 1) indicates that p_i, p_j have certain commonality. Co-reference (also called bibliographic coupling) measures the similarity between webpages. We use $R = (R_{ij})$ to denote the co-reference with R_{ij} defined to be the number of webpages co-referenced by p_i, p_j. One can show that the hub matrix LL^T can be expressed as

$$LL^T = D_{out} + R,$$

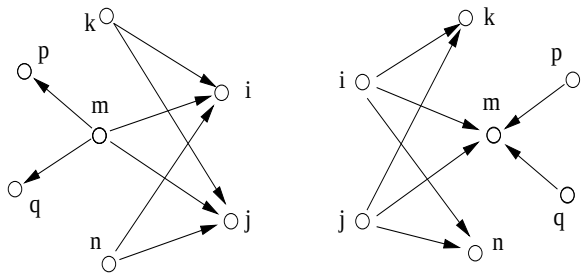

Figure 2: Importance of hyperlink weight normalization. **Left**: webpages p_i, p_j are co-cited by webpages p_k, p_m, p_n. However, since webpage p_m also cites webpages p_p, p_q, the co-citation of p_i, p_j by p_m is not as significant as the co-citation by either p_k or p_n. This fact can be compensated by normalizing the weights on the out-bound links of a webpage; the co-citation of p_i, p_j by p_m is then 2/4=50% as important as the co-citation by either p_k or p_n.
Right: webpages p_i, p_j co-reference webpages p_k, p_n, p_m. However, since webpage p_m also is referenced by other webpages p_p, p_q, the co-reference of p_i, p_j to p_m is not as significant as the co-reference to either p_k or p_n. This fact can be compensated by normalizing the weights on the in-bound links of a webpage.

where $D_{out} = \text{diag}(\mathbf{d}_{\text{out}})$ and $\mathbf{d}_{\text{out}} = (o_1, \cdots, o_n)^T$, is the vector of out-degrees. Thus LL^T is the sum of co-reference and outdegree, revealing the close relationship between hubs and co-references.

The average co-citation can be proved[4] to be $\langle C_{ij} \rangle = \mathbf{d}_{\text{in}}(i)\mathbf{d}_{\text{in}}(j)/(n-1)$ assuming that web graphs are fixed degree sequence random graphs. The average co-reference is $\langle R_{ij} \rangle = \mathbf{d}_{\text{out}}(i)\mathbf{d}_{\text{out}}(j)/(n-1)$. With these results, the solutions (principal eigenvectors) are further obtained in closed-form [4]. From that, it is shown that the authority ranking by HITS in average case is *identical* to the ranking by indegrees. Similarly, hub ranking in HITS is identical to the ranking by outdegrees.

In the following, we study co-citation and co-reference (see Fig.2), focusing on hyperline weight normalization which is a key issue in PageRank.

3 PageRank

In HITS, a webpage with a large number of out-going links will have a large influence on the final ranking, compared to a webpage with a smaller number of out-going links. In PageRank, each out-going hyperlinks from p_i is weighted by $1/o_i$, thus every webpage has the same total out-going weights. We may state this idea as *Internet Democracy*: each website (webpage) has a

total of one vote. The bibliographic reason for weight normalization is shown in Fig.2.

PageRank uses a web surfing model based on a random walk process,

$$\mathbf{x} = \mathcal{I}^{op}(\mathbf{x}) \tag{3.4}$$

where the $\mathcal{I}^{op}$ operation is defined to be

$$\mathcal{I}^{op}(\cdot) = L^T D_{out}^{-1} \equiv P^T. \tag{3.5}$$

This amounts to rescale the adjacency matrix L such that each row is sum-to-one. Thus $P = (P_{ij})$ is a stochastic matrix. At any moment, millions of people are using the web. The stationary distribution $\mathbf{x}$ is determined by $P^T\mathbf{x} = \lambda\mathbf{x}$.

PageRank models two types of random jumps on the Internet. (i) Link-tracking jump: a surfer often follows the hyperlinks of webpages by simply clicking on them; this is modeled by $L^T D_{out}^{-1}$. (ii) Link-interrupt jump: a surfer sometimes jumps to another webpage not hyperlinked by the current webpage. PageRank models such link-interrupt jump with a simple uniform distribution $(1-\alpha)/n$. The full transition probability is $P^T = \mathcal{I}^{op}(\cdot) = \alpha L^T D_{out}^{-1} + (1-\alpha)(1/n)\mathbf{e}\mathbf{e}^T$ where $\alpha = 0.9$ and $\mathbf{e} = (1, \cdots, 1)^T$.

3.1 Hubs in PageRank. We generalize the weight normalization idea to in-bound hyperlinks. This corresponds to normalization of each column of the adjacency matrix L to LD_{in}^{-1}.

There are two reasons for the in-link normalization for hub ranking. First, hub ranking is mostly an indication of co-references (§2.1). As illustrated in Fig.2, co-reference to a webpage with a large indegree is not as significant as co-reference to a webpage with a small indegree. For example, the fact that we all make reference to a highly referenced site such as New York Times homepage says little about whether we are similar. But if two person make reference to Knuth's *The Art of Computer Programming*, it is likely that both persons are interested in computer science.

Second, a rare or unique resource is sometimes pointed to by only a small number of hyperlinks and is thus difficult to be located whereas finding a highly popular website is an easy task. In-link normalization equalizes the efforts for finding a unique resource since the in-links of highly popular websites are weighted very low while the in-links of rare websites are weighted relative higher. This suggests that after the in-link normalization, websites still standing-out must be of special values. Remarkably, we found that the top hubs after in-link normalization are generally have large outdegrees, quite similar to the hubs without in-link normalization. This indication some *intrinsic* nature of these hub sites.

Scheme	$\mathcal{I}^{op}$	$\mathcal{O}^{op}$
HITS	L^T	L
PageRank	$L^T D_{out}^{-1}$	LD_{in}^{-1}
OnormRank	$L^T D_{out}^{-1/2}$	$D_{out}^{-1/2} L$
InormRank	$D_{in}^{-1/2} L^T$	$LD_{in}^{-1/2}$
SnormRank	$D_{in}^{-1/2} L^T D_{out}^{-1/2}$	$D_{out}^{-1/2} L D_{in}^{-1/2}$

Table 1: $\mathcal{I}^{op}$ and $\mathcal{O}^{op}$ operations for HITS, PageRank, the out-link normalized rank (OnormRank), the in-link normalized rank (InormRank), and the symmetrically normalized rank (SnormRank).

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

$$\mathbf{y} = \mathcal{O}^{op}(\mathbf{y}), \tag{3.6}$$

where $\mathcal{O}^{op}$ is defined as

$$\mathcal{O}^{op}(\cdot) = \alpha L D_{in}^{-1} + (1-\alpha)(1/n)\mathbf{e}\mathbf{e}^T \tag{3.7}$$

where LD_{in}^{-1} is the dominant part, and $\mathbf{e}\mathbf{e}^T$ accommodates the link-interrupt jump random surfing.

4 A Unified Framework

The most important feature of HITS is the mutual reinforcement (Eqs.2.1,2.2) between hubs and authorities, while the most important feature of PageRank is the hyperlink weight normalization (cf. Eqs.3.5,3.7). These features can be generalized and combined into a ranking framework with $\mathcal{I}^{op}, \mathcal{O}^{op}$ extended to

$$\mathcal{I}^{op}(\cdot) = D_{in}^{-p} L^T D_{out}^{-q}, \quad \mathcal{O}^{op}(\cdot) = \mathcal{I}^{op}(\cdot)^T. \tag{4.8}$$

As discussed in §2, D_{out}^{-q} describes the out-link normalization, and D_{in} describes the in-link normalization; $p, q \geq 0$ are constant parameters. In this unified framework, one can easily design new ranking algorithms. In this paper, we study three new *normalized* ranking algorithms within this framework. They are defined in Table 1. The key observation is that HITS and PageRank are two extreme ends of a wide spectrum of ranking algorithms within this unified framework. By studying these three intermediate ranking algorithms, we obtain the general conclusion that, to first order approximation, all these ranking algorithms lead to the same ranking.

In this paper, we also clarify and formalize two score computation schemes: (1) *similarity-mediated score propagation* and (2) *random surfing score propagation*.

5 Out-link normalized rank (OnormRank)

OnormRank extends the out-link weight normalization in PageRank for authority ranking. PageRank uses L_1 norm. In OnormRank, out-links are normalized using L_2 norm. $\mathcal{I}^{op}, \mathcal{O}^{op}$ operations are defined by

$$\mathcal{I}^{op}(\cdot) = L^T D_{out}^{-1/2}, \quad \mathcal{O}^{op}(\cdot) = D_{out}^{-1/2} L. \tag{5.9}$$

(see Table 1). OnormRank uses the mutual reinforcement of HITS. Because OnormRank combines both features of HITS and PageRank, the ranking produced by OnormRank is expected to be somewhere between the rankings produced by HITS and PageRank.

The authority scores are determined by the mutual re-inforcing iteration process, $\mathbf{x}^{(t+1)} = \mathcal{I}^{op}(\mathcal{O}^{op}(\mathbf{x}^{(t)}))$ with proper normalization. Authority scores are contained in the principal eigenvector of

$$A^{(O)}\mathbf{x} = \lambda \mathbf{x}, \quad A^{(O)} = L^T D_{out}^{-1} L. \tag{5.10}$$

Using explicit index, elements of authority matrix are

$$A^{(O)}_{ij} = \sum_k \frac{L_{ki} L_{kj}}{o_k}. \tag{5.11}$$

Note $\sum_k L_{ki} L_{kj} = C_{ij}$ is the co-citation between webpages p_i, p_j (see §2.1). Thus in $A^{(O)}_{ij}$ the co-citation is inversely weighted with the outdegree o_k, precisely the situation explained in Fig. 2.

Note that the positive and symmetric matrix $A^{(O)} = L^T D_{out}^{-1} L$ defines the *pairwise similarity* between two webpages. By Rayleigh-Ritz theorem, the principal eigenvector (the authority vector) is the solution to the maximization problem

$$\max_{\mathbf{x}} \frac{\mathbf{x}^T A^{(O)} \mathbf{x}}{\mathbf{x}^T \mathbf{x}}.$$

The similarity matrix $A^{(O)} = L^T D_{out}^{-1} L$ induces an undirected similarity graph $G(A^{(O)})$ among webpages, with $A^{(O)}$ as its adjacency matrix.

The induced similarity graph $G(A^{(O)})$ has following properties: (i) the node degree of the induced graph, $d_i(A^{(O)}) \equiv \sum_j A^{(O)}_{ij} = (L^T D_{out}^{-1} L \mathbf{e})_i = (L^T \mathbf{e})_i = b_i$, is equal to the indegree of the original web graph. We may write $D(A^{(O)}) \equiv \mathrm{diag}(d(A^{(O)})) = D_{in}$. (ii) $\sum_{ij} A^{(O)}_{ij} = \sum_{ij} L_{ij} = |E|$, where $|E|$ is the number of hyperlinks. (iii) The trace of A, $\sum_i A^{(O)}_{ii}$, is

$$\mathrm{Tr}(A^{(O)}) = \mathrm{Tr}(L^T D_{out}^{-1} L) = \mathrm{Tr}(D_{out}^{-1} D_{out}) = n.$$

Thus the diagonal elements of $A^{(O)}$ is 1 on average, in contrast to HITS authority matrix $L^T L$ whose diagonal elements are node indegree (see §2.1). This is another reason OnormRank is called *normalized* ranking.

We wish to compute the authority scores. We can compute them using Eq.5.10. Here we interpret Eq.5.10 in a new way: similarity-mediated score propagation on a similarity graph.

5.1 Similarity mediated score propagation.

Here we formalize the concept of *similarity mediated score propagation* scheme. Consider PageRank: a "good" authority should be pointed by or point to other "good" authorities. This idea translates into the iterative procedure of linearly *propagating* scores on the original directed web graph to an equilibrium state. In HITS, a good *authority* is pointed to by good *hubs* which by definition point to good authorities. We may combine the two-step process into one-step and view it as *similarity-mediated* authority score propagation on an undirected graph, where connection strength is the similarity between webpages, defined by the similarity matrix induced through the iterative mutual reinforcement Eq.(2.3). This is stated formally as

Definition. In similarity-mediated score propagation, scores are computed as the principal eigenvector of $A\mathbf{x} = \lambda \mathbf{x}$, where A contains pairwise similarities.

Remark. Mutual reinforcement on the original web graph is equivalent to similarity-mediated score propagation on the induced similarity graph.

5.2 Random surfing score propagation.

Besides similarity-mediated score propagation, we can adopt PageRank's random surfing on the similarity graph $G(A)$ to define authority scores. Here we only consider the link-tracking random surfing. The associated transition probability is directly proportional to the similarity between webpages, which is specified by the stochastic matrix $\widehat{A}$ obtained by inversely scaling each row of A such that the sum along each row is equal to one,

$$\widehat{A}^{(O)} = [D(A^{(O)})]^{-1} A^{(O)} = D_{in}^{-1} L^T D_{out}^{-1} L \tag{5.12}$$

The equilibrium distribution of random surfers is the solution to $(\widehat{A}^{(O)})^T \widehat{\mathbf{x}} = \widehat{\mathbf{x}}$. One can easily verify that

$$\widehat{\mathbf{x}}_1 = \mathbf{d}_{\mathrm{in}}/|E| = (b_1, \cdots, b_n)^T/|E|, \tag{5.13}$$

is the desired solution. We summarize all these results in the following theorem:

Theorem 5.1 For the authority similarity graph $G(A^{(O)})$, the node degree equals the indegree of the underlying web graph. The diagonal element of $A^{(O)}$ is 1 on average. Furthermore, random surfers on this graph will reach the equilibrium distribution of Eq.5.13.

6 In-link normalized rank (InormRank)

$\mathcal{I}^{op}(\cdot), \mathcal{O}^{op}(\cdot)$ operations are defined in Table 1. All results in §5 can be extended to here similarly.

7 Symmetric normalized rank (SnormRank)

For authority ranking in PageRank, out-links are normalized, i.e., L is replaced by $D_{out}^{-1}L$. For hub ranking in PageRank in-links are normalized, i.e., L is replaced by LD_{in}^{-1}. Here we normalize both in-links and out-links simultaneously in a symmetric fashion (note that HITS also treats in-link and out-link symmetrically). The mutual reinforcement operations are defined by

$$\mathcal{I}^{op}(\cdot) = D_{in}^{-1/2} L^T D_{out}^{-1/2}, \quad \mathcal{O}^{op}(\cdot) = D_{out}^{-1/2} L D_{in}^{-1/2}.$$

We consider the ranking through similarity-mediated score propagation. The authority scores are contained in the principal eigenvector of

$$A^{(S)}\mathbf{x} = \lambda \mathbf{x}, \quad A^{(S)} = D_{in}^{-1/2} L^T D_{out}^{-1} L D_{in}^{-1/2}.$$

Hub scores are contained in the eigenvector of

$$H^{(S)}\mathbf{y} = \lambda \mathbf{y}, \quad H^{(S)} = D_{out}^{-1/2} L D_{in}^{-1} L^T D_{out}^{-1/2}.$$

The principal eigenvectors of these equations have simple closed form solutions. For authority score, the eigenvector is

$$\mathbf{x}_1 = \mathbf{d}_{\text{in}}{}^{1/2} = (b_1^{1/2}, b_2^{1/2}, \cdots, b_n^{1/2})^T, \ \lambda_1 = 1. \quad (7.14)$$

For hub score, the eigenvector is

$$\mathbf{y}_1 = \mathbf{d}_{\text{out}}{}^{1/2} = (o_1^{1/2}, o_2^{1/2}, \cdots, o_n^{1/2})^T, \ \lambda_1 = 1, \quad (7.15)$$

(Both can be easily verified.) We summarize them as

Theorem 7.1 The authority ranking scores of the SnormRank are given in Eq.(7.14). They are exactly the ranking by indegrees. The hub ranking scores of the SnormRank are given in Eq.(7.15). They are exactly the ranking by outdegrees.

Thus SnormRank and OnormRank lead to the same authority ranking (the indegree ranking). By construction, OnormRank and SnormRank are intermediate between HITS and PageRank. From this, we conclude that authority rankings of HITS and PageRank will be close to these normalized rankings.

8 Experiments

The dataset is about the topic *Running* which contains a total of 13152 webpages. This dataset is a sub-category of a larger category *Fitness* which is obtained from the Open Directory Project (*www.dmoz.org*). We give HITS ranking (the modification [1] are adopted), PageRank ranking (Page), OnormRank ranking (OnmR) with similarity mediated score propagation (§5.1) and indegree ranking (IndR).

Hits	Page	OnmR	IndR	URL
1	1	1	2	www.runnersworld.com/
2	5	4	5	sunsite.unc.edu/drears/...
3	2	3	4	www.usatf.org/
4	3	2	1	www.coolrunning.com/
5	4	5	6	www.clark.net/pub/pribut/...
6	8	6	8	www.runningnetwork.com/
7	7	8	9	www.iaaf.org/
8	15	7	14	www.sirius.ca/running.html
9	12	9	12	www.wimsey.com/~dblaikie/
10	14	11	15	www.kicksports.com/
11	6	10	7	www.nyrrc.org/
12	17	12	18	www.usaldr.org/
13	24	13	20	www.halhigdon.com/
14	19	21	25	www.ontherun.com/
15	40	19	10	www.runningroom.com/
16	20	17	23	www.webrunner.com/webrun/...
17	26	18	22	www.doitsports.com/
18	33	26	21	www.arfa.org/
19	21	27	19	www.adidas.com/
20	11	22	11	www.uta.fi/~csmipe/sport/

Here HITS ranking agree with PageRank ranking, especially in top 10. OnormRank is intermediate between HITS and PageRank. They all correlate with the indegree ranking quite well. All major websites relating to *running* are represented in these top ranked webpages.

Lempel and Moran [6] define two Markov chains simultaneously on a bipartite graph, constructed from the original webgraph. Borodin et al [2] proposed two more refined random surfing models. Both these models are special cases of our unified ranking framework.

More detailed analyses, web graph experiments, references are given in the full paper.

References

[1] K. Bharat and M. R. Henzinger. Improved algorithms for topic distillation in a hyperlinked environment. *ACM Conf. on Research and Develop. in Info. Retrieval (SIGIR'98)*, 1998.

[2] A. Borodin, G.O. Roberts, J.S. Rosenthal, and P. Tsaparas. Finding authorities and hubs from link structures on the world wide web. *Proc. 10th WWW Conference*, 2001.

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

[4] 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.

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

[6] R. Lempel and S. Moran. SALSA: stochastic approach for link-structure analysis and the TKC effect. *ACM Trans. Information Systems*, 19:131–160, 2001.