PageRank, HITS and a Unified Framework for Link Analysis

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

LBNL Tech Report 49372, Nov 2001 (updated September 2002)

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. Rankings of webgraphs of different sizes and queries are presented to illustrate our analysis.

Keywords: mutual reinforcement, hyperlink normalization, similarity graph, score propagation.

1 Introduction

Given the vast amount of information on the World Wide Web, a typical short query of 1 - 3 words submitted to a search engine can easily retrieve tens of thousands webpages. Ranking the returned webpages such that the useful ones appear in the top of the ranked list is a critical task in the Web information retrieval (IR). Traditional IR content analysis is often not adequate here because the query is too short, webpages are created with varying qualities, web structure on a local site is not taken into account, and many other reasons.

This leads to the recent research of using information implicitly contained in the hyperlink structure of the web. Two popular ranking algorithms among the early developments are (i) the PageRank algorithm developed by Brin and Page [21, 6] and used in the search engine Google, and (ii) the HITS (Hypertext Induced Topic Selection) algorithm developed by Kleinberg[16]. 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 [4, 9, 19, 5, 20].

In this paper, we briefly discuss HITS with mutual reinforcement of hubs and authorities. We also emphasize the role of co-reference and co-citation (Fig.1), which provides the bibliographic rational for hyperlink weight normalization (Fig.2) as a key concept in link analysis. An indepth analysis of PageRank is also provided. We extend the weight normalization of in-bound links in PageRank to out-bound links and introduce the "hub" ranking for PageRank as well.

We then 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 are precisely given by node indegrees (see Eq.22), and similar results for hub ranking in InormRank, Using score propagation as in HITS, authorities scores are precisely given by square root of indegrees (Eq.27) and hub scores are given by square root of outdegrees (see Eq.28). 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

*NERSC Division, Lawrence Berkeley National Laboratory, University of California, Berkeley, CA 94720, {chqding,xhe,pjrhusbands,hdsimon}@lbl.gov.

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

by indegree. The difference between rankings produced by different algorithms reflects slightly different but useful emphasis. These results support the basic and universal notion that in web ranking, indegree and outdegree are of fundamental importance.

Webpage ranking using link topology tries to capture the notion of some *average* opinion of the webpage creators, and is specific to a webpage collection, such as those retrieved for a user query. The hyperlinks from these webpages form a directed Web graph $G = (V, E)$, where V is the set of nodes representing webpages, and E is the set of hyperlinks. The hyperlink topology of the web graph is contained in the *asymmetric* adjacency matrix $L = (L_{ij})$, where $L_{ij} = 1$ if $p_i \to p_j$ and $L_{ij} = 0$ otherwise.

In this study, we emphasize the role of indegree of a webpage p_i, given by $b_j = \sum_k L_{kj}$. In-degrees of all nodes form the vector $\mathbf{d}_{\rm in}$ and the diagonal matrix D_{in}:

$$\mathbf{d}_{\rm in} = (b_1, b_2, \cdots, b_n)^T,\ D_{in} = \mathrm{diag}(\mathbf{d}_{\rm in}). \qquad (1)$$

Similarly, outdegree of a webpage p_i is defined as $o_i = \sum_k L_{ik}$. Out-degrees form the vector $\mathbf{d}_{\rm out}$ and the diagonal matrix D_{out}:

$$\mathbf{d}_{\rm out} = (o_1, o_2, \cdots, o_n)^T,\ D_{out} = \mathrm{diag}(\mathbf{d}_{\rm out}). \qquad (2)$$

2 HITS Algorithm

In the HITS algorithm[16], 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,

$$\mathbf{x} = \mathcal{I}^{op}(\mathbf{y}), \quad \mathbf{y} = \mathcal{O}^{op}(\mathbf{x}). \qquad (3)$$

Here vectors $\mathbf{x} = (x_1, x_2, \cdots, x_n)^T$ and $\mathbf{y} = (y_1, y_2, \cdots, y_n)^T$ 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

$$\mathcal{I}^{op}(\cdot) = L^T, \quad \mathcal{O}^{op}(\cdot) = L. \qquad (4)$$

More explicitly, written in index notations,

$$x_i = \sum_{j:e_{ji}\in E} y_j, \quad y_i = \sum_{j:e_{ij}\in E} x_j$$

The final authority and hub scores of every webpage can be obtained through an iteratively updating 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

$$\begin{aligned} c\mathbf{x}^{(t+1)} &= \mathcal{I}^{op}(\mathcal{O}^{op}(\mathbf{x}^{(t)}) = L^T L\mathbf{x}^{(t)} \qquad (5)\\ c\mathbf{y}^{(t+1)} &= \mathcal{O}^{op}(\mathcal{I}^{op}(\mathbf{y}^{(t)}) = L L^T\mathbf{y}^{(t)} \end{aligned}$$

where, c is a normalization constant such that $||\mathbf{x}|| = ||\mathbf{y}|| = 1$. 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 [4] by suppressing the contribution from different webpages from same host (site or root in URL) is often adopted. Further developments and applications are discussed in [13, 8, 4, 18, 9, 19, 5, 20] (see §6).

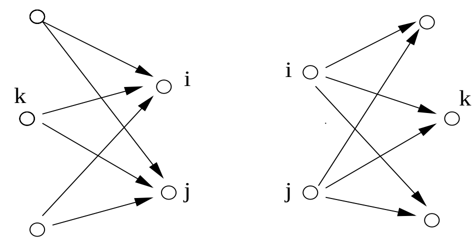

Figure 1: Left: webpages p_i, p_j are co-cited by webpage p_k. Right: webpages p_i, p_j co-reference webpage p_k.

2.1 Co-citation and co-reference

The authority and hub matrices have interesting connections [16] to co-citation and co-reference in the fields of citation analysis and bibliometrics. Here we discuss the relationships in further details and emphasize the important roles of in-degrees and out-degrees.

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 is a similarity measure[22], defined as the number of webpages that co-cite p_i, p_j: and is computed as $C_{ij} = \sum_k L_{ki}L_{kj} = (L^T L)_{ij}$, $i \neq j$. The self co-citation C_{ii} is not defined and is usually set to $C_{ii} = 0$. Also, $(L^T L)_{kk} = \sum_j L_{jk}L_{jk} = \sum_j L_{jk} = b_k$ is the indegree of p_k. This implies $\mathrm{diag}(L^T L) = D_{in}$. Therefore the authority matrix $L^T L$ is

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

which is the sum of co-citation and indegree [10]. 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

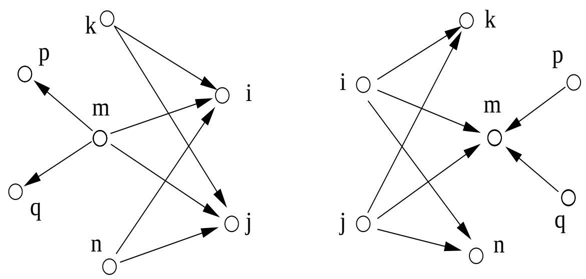

Figure 2: Importance of hyperlink weight normalization. **Left**: webpages p_i, p_j are co-cited by webpages p_k, p_m, p_n. However, since webpage p_m also cites webpages p_p, p_q, the co-citation of p_i, p_j by p_m is not as significant as the co-citation by either p_k or p_n. This fact can be compensated 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.

(also called bibliographic coupling[15]) 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 and is calculated as: $R_{ij} = \sum_k L_{ik}L_{jk} = (LL^T)_{ij},\ i \neq j$. The self-reference R_{ii} is not defined, and is set to $R_{ii} = 0$. Also $(LL^T)_{ii} = \sum_k L_{ik}L_{ik} = \sum_k L_{ik} = o_i$ is the outdegree of p_i. Thus $\text{diag}(LL^T) = D_{out}$ and the hub matrix LL^T can be expressed as

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

which is the sum of co-reference and outdegree, revealing the close relationship between hubs and co-references.

The average co-citation can be proved[10] 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 [2]. 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 [10]. 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 explore a different direction by carefully studying 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 an idea similar to HITS that a "good" webpage should connect to or be pointed to by other "good" webpages. However, instead of mutual reinforcement, it adopts a web surfing model based on a Markov process in determining the scores:

$$\mathbf{x} = \mathcal{I}^{op}(\mathbf{x}) \tag{6}$$

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

$$\mathcal{I}^{op}(\cdot) = L^T D_{out}^{-1} \equiv P^T. \tag{7}$$

This amounts to rescale the adjacency matrix L such that each row is sum-to-one. Thus $P = (P_{ij})$ is a stochastic matrix, since $\sum_v P_{uv} = 1,\ P_{uv} \geq 0$. P_{ij} represents the probability of a web surfer making a transition from webpage p_i to p_j. Starting from any webpage p_i, a surfer goes to any one of the hyperlinked webpages with equal probability $1/o_i$. At any moment, millions of people are using the web. We assume they follow the random surfing model. They will reach the equilibrium (stationary) distribution under general conditions. If a webpage has a high probability in the equilibrium distribution, that means more surfers will visit the webpage. Therefore, the equilibrium distribution of surfers on webpages is a measure of a webpage's "importance", which is the authority score in PageRank. The equilibrium distribution $\mathbf{x}$ is determined by

$$P^T\mathbf{x} = \lambda\mathbf{x}, \tag{8}$$

and x satisfies $\sum_k \mathbf{x}(k) = 1$. One can obtain the solution iteratively. Note that $\lambda = 1$ if the Markov process has an equilibrium distribution $\mathbf{x}$.

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-interupt jump: a surfer sometimes jumps to another webpage not hyperlinked by the current webpage. PageRank models such link-interupt jump with a simple uniform distribution $(1-\alpha)/n$.
The full stochastic matrix of transition probability is

$$P^T = \mathcal{I}^{op}(\cdot) = \alpha L^T D_{out}^{-1} + (1-\alpha)(1/n)\mathbf{e}\mathbf{e}^T \tag{9}$$

where $\alpha = 0.8 \sim 0.9$. Here $\mathbf{e} = (1, 1, \cdots, 1)^T$; thus $\mathbf{e}\mathbf{e}^T$ is a matrix of all 1's.

3.1 Further analysis of PageRank

Zero-outdegree webpages

PageRank has been generally considered as a "random walk" and the final importance score is the equilibrium distribution of random walkers. Here we point out this description is technically *incorrect*, i.e., the PageRank stochastic matrix of Eq.9 does not have a true equilibrium distribution.

Consider a webpage without out-links. A surfer currently looking at this webpage has no out-links to click, and thus will jump to a random webpage with probability $p = 1$, instead of probability $p = 1 - \alpha$ as specified in standard PageRank (Eq.9). In other words, for this zero-outdegree webpage p_i, $\sum_j P_{ij} = 1 - \alpha < 1$. Therefore, P is no longer a stochastic matrix. If we solve the eigenvalue problem Eq.(8), we will find that $\lambda < 1$, not $\lambda = 1$. In other words, there is no *equilibrium* distribution for the random surfers. For a webpage collection, there are a large number of zero-outdegree webpages (many good authorities have zero or small number of out-bound links). Thus the scores obtained in standard PageRank do not follow an equilibrium distribution: they are not true random walks.

To insure a true equilibrium distribution for random surfers, we increase the random surfer transition probability on these zero-outdegree webpages to $1/n$, i.e, we define

$$\mathbf{a}(j) = \begin{cases} 1/n & \text{if} \quad o_j = 0, \\ (1-\alpha)/n & \text{if} \quad o_j \geq 1, \end{cases} \tag{10}$$

and modify the transition probability as

$$P^T = \alpha L^T D_{out}^{-1} + \mathbf{a}\mathbf{e}^T.$$

With this modification, $\sum_j P_{ij} = 1$ for every webpage p_i. Thus P is now a correct stochastic matrix, and a true equilibrium distribution is ensured.

Zero-indegree webpages

Large number of webpages have no in-bound hyperlinks (zero-indegree). For these webpages, we have the following:

Theorem 3.1 In PageRank, the authority scores of zero-indegree webpages have the following properties: (i) they are zero if we use the original transition probability matrix $P^T = L^T D_{out}^{-1}$. (ii) they are all equal and smaller than any non-zero-indegree webpages if we use the full transition probability $P^T = \alpha L^T D_{out}^{-1} + (1-\alpha)(1/n)\mathbf{e}\mathbf{e}^T$.

Proof. For webpage p_j, zero indegree implies that all elements on the jth column in L are zero, so are $D_{out}^{-1} L$. Thus the jth row in $L^T D_{out}^{-1}$ are all zeroes. Starting with $\mathbf{x}^{(0)}$, the jth entry in $\mathbf{x}^{(1)} = L^T D_{out}^{-1} \mathbf{x}^{(0)}$ is zero, and will remain so in all subsequent iterations. This proves point (i). When using the full transition probability, at any iteration t with properly normalized $\mathbf{x}^{(t)}$, we have

$$\begin{aligned} \mathbf{x}^{(t+1)} &= \alpha L^T D_{out}^{-1} \mathbf{x}^{(t)} + (1-\alpha)(1/n)\mathbf{e}\mathbf{e}^T \mathbf{x}^{(t)} && (11) \\ &= \alpha L^T D_{out}^{-1} \mathbf{x}^{(t)} + (1-\alpha)(1/n)\mathbf{e}, && (12) \end{aligned}$$

since $\mathbf{e}^T \mathbf{x}^{(t)} = \sum_k \mathbf{x}^{(t)}(k) = 1$. The first term contributes zero to those zero-indegree webpages as in (i). The second term contributes identical amount to every webpage. In every iteration, this remains true. Thus all zero-indegree webpages have the same score. Note also that $L^T D_{out}^{-1} \mathbf{x}^{(t)}$ will contribute a positive quantity to all non-zero-indegree webpages; they will have larger scores than zero-indegree webpages. This proves point (ii). □

Thus for authority ranking, zero-indegree webpages are entirely *insignificant*: they all have a score smaller than any webpage with 1 or more in-bound link.

This fact can be used to speedup the computation of the ranking scores. In typical cases, because of the Zipf-law distribution of indegrees [7], a majority of webpages have zero indegrees. Neglecting them in the score computation, we can speedup the computation substantially.

3.2 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 $L D_{in}^{-1}$.

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

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

without in-link normalization. This indication some *intrinsic* nature of these hub sites.

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

$$\mathbf{y} = \mathcal{O}^{op}(\mathbf{y}), \tag{13}$$

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

$$\mathcal{O}^{op}(\cdot) = \alpha L D_{in}^{-1} + (1-\alpha)(1/n)\mathbf{e}\mathbf{e}^T \equiv Q^T \tag{14}$$

where $L D_{in}^{-1}$ is the main part as explained above, and $\mathbf{e}\mathbf{e}^T$ accommodates the link-interupt jump random surfing, same as that in defining authority scores. The hub scores are obtained through the iterations $\mathbf{y}^{(t+1)} = \mathcal{O}^{op}(\mathbf{y}^{(t)}) = Q^T\mathbf{y}^{(t)}$ which is the equation for equilibrium distribution of a Markov process $Q^T\mathbf{y} = \lambda\mathbf{y}$, where the eigenvalue $\lambda = 1$. To overcome the problem with zero-indegree webpages, we define

$$\mathbf{h}(j) = \begin{cases} 1/n & \text{if} \quad b_j = 0, \\ (1-\alpha)/n & \text{if} \quad b_j \geq 1. \end{cases} \tag{15}$$

and the modified full transition probability

$$Q^T = \alpha L D_{in}^{-1} + \mathbf{h}\mathbf{e}^T.$$

to insure a true equilibrium distribution for random surfers.

4 A unified framework

The most important feature of HITS is the mutual reinforcement (Eqs.3,4) between hubs and authorities, while the most important feature of PageRank is the hyperlink weight normalization (cf. Eqs.7,14). These features can be generalized and combined into a ranking framework with $\mathcal{I}^{op}, \mathcal{O}^{op}$ extended to

$$\mathcal{I}^{op}(\cdot) = D_{in}^{-p} L^T D_{out}^{-q}, \quad \mathcal{O}^{op}(\cdot) = \mathcal{I}^{op}(\cdot)^T. \tag{16}$$

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. Our purpose for introducing these three ranking algorithms is to prove this general conclusion. We do not claim that these normalized ranking algorithms are superior than HITS or PageRank.

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 (see Table 1). In OnormRank, out-links are normalized using L_2 norm. $\mathcal{I}^{op}, \mathcal{O}^{op}$ operations are defined by

$$\mathcal{I}^{op}(\cdot) = L^T D_{out}^{-1/2}, \quad \mathcal{O}^{op}(\cdot) = D_{out}^{-1/2} L. \tag{17}$$

(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 reinforcing 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

$$A^{(O)}\mathbf{x} = \lambda\mathbf{x}, \quad A^{(O)} = L^T D_{out}^{-1} L. \tag{18}$$

Writing out explicitly, the elements of authority matrix are

$$A_{ij}^{(O)} = \sum_k \frac{L_{ki}L_{kj}}{o_k}. \tag{19}$$

Note $\sum_k L_{ki}L_{kj} = C_{ij}$ is the co-citation between webpages p_i, p_j (see §2.1). Thus in $A_{ij}^{(O)}$ 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 the following properties:
(i) the node degree of the induced graph,

$$d_i(A^{(O)}) \equiv \sum_j A_{ij}^{(O)} = (L^T D_{out}^{-1} L\mathbf{e})_i = (L^T \mathbf{e})_i = b_i.$$

is equal to the indegree of the original web graph. We may write this as

$$D(A^{(O)}) \equiv \mathrm{diag}(d(A^{(O)})) = D_{in}. \qquad (20)$$

(ii) $\sum_{ij} A_{ij}^{(O)} = \sum_{ij} L_{ij} = |E|$, where $|E|$ is the number of hyperlinks.
(iii) The trace of A, $\sum_i A_{ii}^{(O)}$, 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 the 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.18. Here we interpret Eq.18 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 (cf Eq.8). 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.(5). 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 [see Eq.(20)]:

$$\widehat{A}^{(O)} = [D(A^{(O)})]^{-1} A^{(O)} = D_{in}^{-1} L^T D_{out}^{-1} L \qquad (21)$$

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

$$\widehat{\mathbf{x}}_1 = \mathbf{d}_{\mathrm{in}}/|E| = (b_1, \cdots, b_n)^T/|E|, \qquad (22)$$

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.22.

In OnormRank, if webpages are ranked using the random surfing score propagation, as in the case of PageRank, the importance of a webpage is directly proportional to its indegree. In general, ranking scores obtained by the similarity-mediated score propagation on a similarity graph will differ from those obtained by the random surfing score propagation on the same similarity graph; but they will be reasonably close (see experiments in §8). Based on this analogy, we conclude that the OnormRank scores will have a high correlation with the indegrees.

6 In-link normalized Rank (InormRank)

InormRank extends the in-link weight normalization for hub ranking in PageRank (cf §3.2). In InormRank, inbound links are normalized using L_2 norm. The mutual reinforcement $\mathcal{I}^{op}, \mathcal{O}^{op}$ operations are defined by

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

(see Table 1). The hub scores are determined by the mutual re-inforcing iteration process, $\mathbf{y}^{(t+1)} = \mathcal{O}^{op}(\mathcal{I}^{op}(\mathbf{y}^{(t)}))$

with proper normalization. Hub scores are contained in the principal eigenvector of

$$H^{(I)}\mathbf{y} = \lambda\mathbf{y}, \quad H^{(I)} = LD_{in}^{-1}L^T. \tag{23}$$

The hub matrix can be written as

$$H^{(I)}_{ij} = \sum_k \frac{L_{ik}L_{jk}}{b_k}. \tag{24}$$

$\sum_k L_{ik}L_{jk} = R_{ij}$ is the co-reference between webpages p_i, p_j (see §2.1). Thus in $H^{(I)}_{ij}$ the co-reference is inversely weighted with the indegree b_k, as shown in Fig. 2.

The hub scores are determined by either similarity-mediated score propagation or random surfing score propagation. All results for OnormRank (§5.1) can be similarly established here. (i) $d_i(H^{(I)}) = o_i$. (ii) $\sum_{ij} H^{(I)}_{ij} = |E|$. (iii) $\mathrm{Tr}(H^{(I)}) = n$. In particular, using random surfing score propagation, the hub scores are given by the equilibrium distribution

$$\mathbf{y}_1 = \mathbf{d}_{\text{out}}/|E| = (o_1, o_2, \cdots, o_n)^T/|E|,$$

i.e., hub ranking of a webpage is directly proportional to its out-bound degree.

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

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

(see Table 1). This symmetric-ranking is introduced to bridge the difference between HITS and PageRank and we expect the final scores will be somewhere between the scores from HITS and PageRank.

We consider the ranking through similarity-mediated score propagation[1]. The authority scores are contained in the principal eigenvector of

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

Using explicit index,

$$A^{(S)}_{ij} = \frac{1}{\sqrt{b_i}}\left(\sum_k \frac{L_{ki}L_{kj}}{o_k}\right)\frac{1}{\sqrt{b_j}}. \tag{25}$$

[1]For brevity, OnormRank, InormRank and SnormRank use similarity-mediated score propagation by default. When using random surfing score propagation, it will be explicitly mentioned.

Thus $A^{(S)}$ has similar out-link normalization as in OnormRank (see Eq.19).

The hub scores are contained in the principal eigenvector of

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

Using explicit index, we have

$$H^{(S)}_{ij} = \frac{1}{\sqrt{o_i}}\left(\sum_k \frac{L_{ik}L_{jk}}{d_k}\right)\frac{1}{\sqrt{o_j}}, \tag{26}$$

similar to the in-link normalization as in InormRank (see Eq.24).

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

$$\mathbf{x}_1 = \mathbf{d}_{\text{in}}{}^{1/2} = (b_1^{1/2}, b_2^{1/2}, \cdots, b_n^{1/2})^T, \ \lambda_1 = 1. \tag{27}$$

For hub score, the eigenvector is

$$\mathbf{y}_1 = \mathbf{d}_{\text{out}}{}^{1/2} = (o_1^{1/2}, o_2^{1/2}, \cdots, o_n^{1/2})^T, \ \lambda_1 = 1, \tag{28}$$

(Both can be easily verified.) We summarize them as **Theorem 7.1** The authority ranking scores of the SnormRank are given in Eq.(27). They are exactly the ranking by indegrees. The hub ranking scores of the SnormRank are given in Eq.(28). They are exactly the ranking by outdegrees.

Thus SnormRank and OnormRank lead to approximately same authority ranking (which is 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. We further conjecture that, to first order approximation, all ranking algorithms within the unified framework defined in §4 will have the same ranking. (We can reach similar results on hubs.) This is the central results of this paper.

8 Experiments

We give experimental results with HITS and PageRank, which serve to illustrate the analysis. Due to space limitation, we only list results for authorities, which are also the most important ones in seeking information on the Web.

For the new ranking schemes, we list only OnormRank with similarity-mediated score propagation. (As the analysis in §5-7, OnormRank with random surfing score propagation is identical to the ranking by indegree;

SnormRank produces ranking identical to ranking by indegree; InormRank is mainly about Hub ranking. These ranking are thus omitted here.)

Our main theme in this paper is that HITS and PageRank provide same ranking in first order approximation. The three new ranking schemes are mainly introduced as intermediate ranking schemes to prove this theme; they are not considered as improved ranking schemes.

Experiment 1. This dataset 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. The table below shows the list of the top 20 authorities. with HITS, PageRank (Page) and ranking-by-indegree (IndR).

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

We see that the HITS ranking agrees well with the PageRank ranking, especially in top 10. They are also highly correlated with the ranking by indegree.

Two types of webpages that deviate substantially from indegree and are worthwhile to mention: (a) highly ranked authority webpages, but with relatively smaller indegrees; (b) webpages with large indegrees, but ranked low. For type (b) website *www.linkexchange.com*, is ranked 13 by indegree, but ranked 111 by HITS. This website have very large indegrees, but very small outdegrees; it is rank *sinks*. For type (a) website *sunsite.unc.edu*, is ranked 13 in HITS, but is ranked 33 by indegree.

Experiment 2. This 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*). Note that in this and all later experiments, the full URL is retained instead of site-level URL in experiment 1; the modification [4] of HITS are adopted. We also give the ranking by OnormRank (OnmR) as discussed in §5.1.

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.

Experiment 3. This dataset is web neighborhood graph for query word *star*. First, among all the retrieved webpages we choose top 120 as the root set. The root set is then expanded into a one level neighborhood graph [16] with 3504 webpages (URLs). We list the top ranked URLs for each ranking method.

HITS	URL
1	www.starwars.com/
2	www.lucasarts.com/
3	www.jediknight.net/
4	www.sirstevesguide.com/
5	www.paramount.com/
6	www.surfthe.net/swma/
7	insurrection.startrek.com/
8	www.startrek.com
9	www.fanfix.com/
10	www.physics.usyd.edu.au/.../starwars

OnmR	URL
1	www.starwars.com/
2	www.lucasarts.com/
3	www.jediknight.net/
4	www.paramount.com/
5	www.sirstevesguide.com/
6	www.surfthe.net/swma/
7	insurrection.startrek.com/
8	www.fanfix.com/
9	shop.starwars.com/
10	www.physics.usyd.edu.au/.../starwars

Page	URL
1	www.starwars.com/
2	www.lucasarts.com/
3	www.paramount.com/
4	www.4starads.com/romance/
5	www.starpages.net/
6	www.dailystarnews.com/
7	www.state.mn.us/mainmenu.html
8	www.star-telegram.com/
9	www.starbulletin.com/
10	www.kansascity.com/
..	......
19	www.jediknight.net/
21	insurrection.startrek.com/
23	www.surfthe.net/swma/
34	www.fanfix.com/
35	www.physics.usyd.edu.au/.../starwars

This webgraph is dominated by *Star Wars*; its related interesting webpages all show up in three rankings. However, PageRank also show `starpages.net` (gossip on stars) and `star-telegram` (news media with *star* as part of its name) on top.

Experiment 4. This webgraph contains retrieved webpages for query *amazon*. The neighborhood webgraph is constructed same as Experiment 3. In this graph, there are 17 connected components, the largest one has 2181 webpages (URLs). Ranking is performed on the largest component. We list the top ranked URLs for each ranking method.

HITS	URL
1	www.amazon.com/
2	www.amazon.co.uk/
3	www.amazon.de
4	www.amazoncity.com/
5	www.echostation.com/
6	www.amazonfembks.com/
7	www.amazonrecords.com/
8	www.amazon.com/exec/obidos/subst/help/
9	www.igc.apc.org/women/bookstores/
10	w3.one.net/~jhoffman/sqltut.htm
11	www.amazon.org/
12	radio.amazoncity.com/

OnmR	URL
1	www.amazon.com/
2	www.amazon.co.uk/
3	www.amazon.de
4	www.amazoncity.com/
5	www.echostation.com/
6	www.amazonfembks.com/
7	www.amazon.com/exec/obidos/subst/help/
8	www.amazonrecords.com/
9	www.igc.apc.org/women/bookstores/
10	w3.one.net/~jhoffman/sqltut.htm
11	www.amazon.org/
12	radio.amazoncity.com/

Page	URL
1	www.amazon.com/
2	www.amazon.co.uk/
3	www.amazoncity.com/
4	www.echostation.com/
5	www.amazon.org/
6	www.amazonfembks.com/
7	www.amazon.de
8	radio.amazoncity.com/
9	www.amazoncityradio.com/
10	s1.amazon.com/exec/varzea/subst/home/home
11	www.ethnobotany.org/
12	www.science.org/amazonassociate.html
13	www.amazonrecords.com/

This query is dominated by `amazon.com` and its European affiliates. Another theme is on lesbian issues and is represented by `amazon.org, amazoncity.com`. `echostation.com` is a *Star Wars* related site; it promotes `amazon.com` as one of its top vendors. All these show up in three rankings.

Experiment 5. This webgraph is retrieved for query *apple*. The query neighborhood web graph is constructed same as Experiment 3. In this graph, there are 21 connected components. The largest one has 2456 webpages (URLs), for which ranking is performed. We list the top ranked URLs for each ranking method.

HITS	URL
1	www.apple.com/
2	www.apple.com/support/
3	www.info.apple.com
4	www.apple.com/quicktime/
5	www.apple.com/education/
6	www.apple.com/hotnews/

```
  7  www.apple.com/developer/
  8  quicktime.apple.com/
  9  www.euro.apple.com/
 10  www.apple.com/find/
 ..  ......
 17  itools.mac.com/WebObjects/Tools
 19  www.macweek.com/
 29  www.claris.com/
 55  www.next.com/

OnmR URL
  1  www.apple.com/
  2  www.claris.com/
  3  www.info.apple.com
  4  www.next.com/
  5  www.pelagius.com/AppleRecon/
  6  www.apple.de/
  7  www.apple.com/support/
  8  www.macos.apple.com/
  9  www.apple.com/quicktime/
 10  www.macweek.com/
 ..  ......
 30  itools.mac.com/WebObjects/Tools

Page URL
  1  www.apple.com/
  2  quicktime.apple.com/sw/sw3.html
  3  www.jokewallpaper.com/
  4  apple.com/
  5  the-tech.mit.edu/Macmade/
  6  www.apple.de/
  7  www.claris.com/
  8  www.apple.ca/
  9  quicktimevr.apple.com/
 10  www.apple.com.au/
 11  www.next.com/
 12  itools.mac.com/WebObjects/Tools
 ..  ......
 25  www.macweek.com/
```

This query is dominated by `apple.com` and its subdomains. The related useful sites such as `claris.com, next.com, macweek.com` clearly show in PageRank and OnormRank.

Several observations and comments on these 5 experiment results are given below. (1) HITS and PageRank lead to very similar rankings. These rankings are highly correlated with indegrees. (2) OnormRank gives ranking quite close to that of HITS in experiments 2-4. Also, OnormRank ranking is closer to PageRank ranking than HITS ranking is close to PageRank. This confirms our theoretical arguments in §4-5 that OnormRank is somewhere in-between HITS and PageRank.

(3) When there are more than one connected components in the webgraph, HITS and OnmRank mostly focus on the large connected component, while PageRank can still pick up significant sites from smaller connected components, due to the random link-interupt jump term in Eq.9. In this sense, PageRank is more stable than HITS. Ng et al [20] extend this random part to HITS.

(4) Although the major webpages in a webgraph (defined by large indegree) mostly show up on the top in all three rankings, different ranking schemes occasionally bring some (different) useful webpages to top, such as in experiments 3 and 5.

(5) Rankings of a query graph (experiments 3-5) is often over-represented by webpages relating to the dominant topic (*star wars* dominates query *star*). Webpages of non-dominant topics are often ranked low (say, 41th or lower) and outside the range that most users browse; and these non-dominant but interesting topics are often missed by users. One way to overcome this difficulty is to use webpage clustering on the retrieved webpages (see [14] for example) so that non-dominant topics are separated into different clusters and then use ranking method to pick up the top webpages on each topic. One may also uses web trawling [18] or network-flow model [12] to identify the web communities and then pick up top webpages for each community.

9 Related work

There are many further development following HITS and PageRank. There are refined ranking schemes [4], probablistic extension of HITS [9], stability study [20]. and rank aggregation [11]. HITS is used for locating web community [13, 8]. Web models are discussed in [17, 1, 3, 2].

We discuss in some detail the work related to the ranking issues in our analysis. In SALSA [19], Lempel and Moran define two Markov chains simultaneously on a bipartite graph, constructed from the original webgraph. An edge (b_h, k_a) is a transition between hub node i and authority node k. The stochastic matrix used in the Markov chain for authority between authority nodes i, j are

$$(\hat{A})_{ij} = \sum_{k|(k_h,i_a),(k_h,j_a)\in\hat{G}} \frac{1}{deg(i_a)} \frac{1}{deg(k_h)} \tag{29}$$

They show that the equilibrium distribution of this random walk is proportional to node indegrees, i.e, SALSA ranking is equivalent to ranking by indegree. A random walk model for hub score is similarly constructed there.

Borodin et al [5] proposed two more refined random surfing models, based on $(BF)^k$ and $(FB)^k$ ($B(i)$ is the set of arcs pointing to i and $F(i)$ is the set of edges pointing away from i, and $k \geq 1$ is a fixed integer). They show that the equilibrium distribution for authority is

$$(a_1, \cdots, a_n), \quad a_i = \sum_j (L^T L)^k_{ij} / ||(L^T L)^k||_F,$$

and similar result for hub scores. The interesting point is that the HITS ranking is recovered as $k \to \infty$, since only the principle eigenvector survive in that limit.

These refined random walk models are interesting extensions of the HITS (and PageRank); but they do not really have the same *mutual reinforcement* as in HITS because authority scores and hub scores are not related by the $\mathcal{I}^{op}, \mathcal{O}^{op}$. Their random walk score propagation differs from HITS which uses similarity-mediated score propagation (§5.1). In these models, a surfer can jump from webpage p_i to webpage p_j even though there is no hyperlink pointing from p_i to p_j, and the link-interupt jumps are absent; thus they do not directly simulate the behavior of web surfers, while PageRank does.

Our approach is different in that we directly extend $\mathcal{I}^{op}, \mathcal{O}^{op}$ operations, as shown in Table 1, that combine mutual reinforcement of HITS with hyperlink weight normalization of PageRank. Thus our approach more closely resembles HITS and PageRank. Within this ranking framework, there is a large space to define $\mathcal{I}^{op}, \mathcal{O}^{op}$ operations — HITS and PageRank are two extreme situations in this space. Our main point is that all ranking algorithms in this framework, to first order approximation, lead to same ranking which is ranking by indegree.

An interesting question is whether SALSA and that of Borodin et al are contained in this ranking framework. In SALSA, $(\hat{A})_{ij}$ in Eq.(29) can be written as $(\hat{A})_{ij} = (D_{in}^{-1} L^T D_{out}^{-1} L)_{ij}$ in our notation, which is identical to the transition matrix of Eq.21 in OnormRank. Thus OnormRank (using the random surfing score propagation) have the same final ranking as SALSA. In this sense, SALSA is contained in the ranking framework. For the models of Borodin et al, we can obtain the same final ranking scores if we define[2] $\mathcal{I}^{op}(\cdot) = \mathcal{O}^{op}(\cdot)^T = (L^T L)^{k/2}$, and use the random surfing score propagation. Thus the models of Borodin et al are also contained in this ranking framework.

10 Discussions and conclusion

We studied similarity and difference between HITS and PageRank. In practical applications, PageRank is applied to the entire web graph, while HITS is usually applied to a subgraph related to a query. This difference is not essential: PageRank can be equally well application to a subgraph and HITS can be equally applied to entire web graph.

An important characteristics of web page ranking is that we strongly emphasize the top ranked webpages, say, those in top 20, because in general a user seldomly browse beyond these webpages. This is in sharp contrast to other rankings, say college ranking: whether a college is ranked 50th or 90th makes a big difference, because most students will attend colleges which are not top ranked, and a college ranked 50th is more attractive than a college ranked 90th. On the web, every surfer chooses the No. 1 site he/she wants to go to. Therefore in assessing ranking effectiveness or comparing different ranking, only the top ranked webpages are considered. A ranking scheme improving ranking at 50th place and below will make no difference.

In this paper, we emphasize the role of co-reference in defining authorities, and the role of co-reference in defining hubs. This provides rational for hyperlink weight normalization as used in PageRank to justify the random surfing model. We generalize and combine hyperlink weight normalization and the mutual reinforcement, together with similarity-mediated and random surfing score propagation schemes, to form a unified framework for link analysis. We analyze three normalized ranking algorithms (OnormRank, InormRank, SnormRank) within this framework. Closed-form solutions are obtained which show that indegree and outdegree are of fundamental importance in all ranking algorithms. Good authorities and good hubs should always have high in/outdegrees.

Two types of exceptions to this generic rule are: (a) Highly ranked (HITS or PageRank) authority webpages, but with relatively small indegrees and (b) Webpages with large indegrees, but ranked low by HITS or PageRank. These webpages would have been incorrectly ranked if we simply count indegrees; they represent the net improvements brought by HITS or PageRank. We note that in literature citation, (a) corresponds to some highly regarded papers which are not popularly cited. and (b) corresponds to some frequently cited but not highly regarded papers. We have discussed several cases in §5.

As mentioned above, strong emphasis on top ranked webpages is a key characteristics in web ranking schemes. For most queries, the number of retrieved webpages is usually very large, easily be larger than 10,000 webpages. Ranking the webpages such that the most informative webpages are placed within top 20 is therefore a truly challenging task. We have shown that in-degree ranking is a first order approximation to HITS and PageRank. Indegree ranking is also a good ranking in the sense most connected (and thus probably most visited) webpages are listed in the top. We have seen that sometimes HITS and PageRank can bring a relatively low-ranked but *use-*

[2] If we further expand the type of $\mathcal{I}^{op}, \mathcal{O}^{op}$ operations beyond the hyperlink weight normalizations defined in Eq.(16).

ful webpage to within top 20. In other words, HITS or PageRank effectively gives a slightly different but useful perturbation (deviation) from the indegree ranking. For this reason, we believe a search engine should provide multiple ranking schemes for user to choose from. Browsing through 60 webpages from a given ranking is probably not as effective as browsing through 20 top webpages on three different rankings. [3] Some webpages are likely to remain on top in all three different rankings because of their large indegrees. Some other useful webpages are likely to appear on top in one of the rankings. A user may first browse through those consistently top-ranked webpages, and move on to those top-ranked pages which are brought into top 20 by a single ranking scheme because the webpage fits well to a particular heuristics emphasized in a ranking scheme.

As for practical applications, our results suggest that in additional to current search engine ranking techniques, one may (1) combine ranking results from multiple schemes in some way; and/or (2) cluster the webpages and apply ranking algorithm to each cluster in order to capture non-dominant topics.

Our results also suggest that new ranking algorithm development should emphasize their deviations from indegree ranking and explore the significance of that deviations.

Acknowledgement. This work is supported by the U.S. Department of Energy (Office of Science, Office of Laboratory Policy and Infrastructure through LDRD) under contract DE-AC03-76SF00098.

References

[1] D. Achlioptas, A. Fiat, A.R. Karlin, and F. McSherry. Web search via hub synthesis. *Proc. Symp. on Foundations of Computer Science*, 2001.

[2] W. Aiello, F. Chung, and L. Lu. A random graph model for massive graphs. *ACM Symposium on Theory of Computing*, pages 171–180, 2000.

[3] A.-L. Barabasi and R. Albert. Emergence of scaling in random networks. *Science*, 286:509–512, 1999.

[4] 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.

[5] 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.

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

[7] A. Broder, R Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. Graph structure in the web. *Proc. 9th International World Wide Web Conference*, 2000.

[8] S. Chakrabarti, B. E. Dom, P. Raghavan, S. Rajagopalan, D. Gibson, and J. Kleinberg. Automatic resource compilation by analyzing hyperlink structure and associated text. *Computer Networks and ISDN Systems*, 30:65–74, 1998.

[9] D. Cohn and H. Chang. Learning to probabilistically identify authoritative documents. *Proc. ICML 2000. pp.167-174.*, 2000.

[10] 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 (www.nersc.gov/∼cding/hits.ps)*, 2001.

[11] C. Dwork, R. Kumar, M. Naor, and D. Sivakumar. Rank aggregation methods for the web. *Proc. 10th WWW Conference*, 2001.

[12] G. W. Flake, S. Lawrence, and C. L. Giles. Efficient identification of web communities. *Proc. Int'l Conf. Knowledge Kiscovery and Data Mining (KDD)*, pages 150–159, 2000.

[13] D. Gibson, J. Kleinberg, and P. Raghavan. Inferring web communities from link topology. In *Proc. 9th ACM Conference on Hypertext and Hypermedia (HYPER-98)*, pages 225–234, 1998.

[14] X. He, C. Ding, H. Zha, and H.D. Simon. Automatic topic identification using webpages clustering. *Proc. IEEE Int'l Conf. Data Mining. San Jose, CA*, pages 195–202, 2001.

[15] M. Kessler. Bibliographic coupling between scientific papers. *American documentation*, 14:10–25, 1963.

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

[17] R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins, and E. Upfal. Stochastic models for the web graph. *Proc. Symposium on Foundations of Computer Science (FOCS)*, 2000.

[18] R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. Trawling the Web for emerging cyber-communities. *Computer Networks*, 31:1481–1493, 1999.

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

[20] A.Y. Ng, A.X. Zheng, and M.I. Jordan. Stable algorithms for link analysis. *Proc. ACM Conf. on Research and Develop. Info. Retrieval (SIGIR)*, 2001.

[21] L. Page, S. Brin, R. Motowani, and T. Winograd. PageRank citation ranking: bring order to the web. *Stanford Digital Library working paper 1997-0072*, 1997.

[22] H. Small. Co-citation in the scientific literature: A new measure of the relationship between two documents. *J. Am. Soc. for Info. Sci.*, 24(4):265–269, 1973.

[3] This is somewhat similar to different ordering in a catalog system: one may switch among the time-ordered list, the price-ordered list, the brand-name ordered list, etc.