LBNL-47847



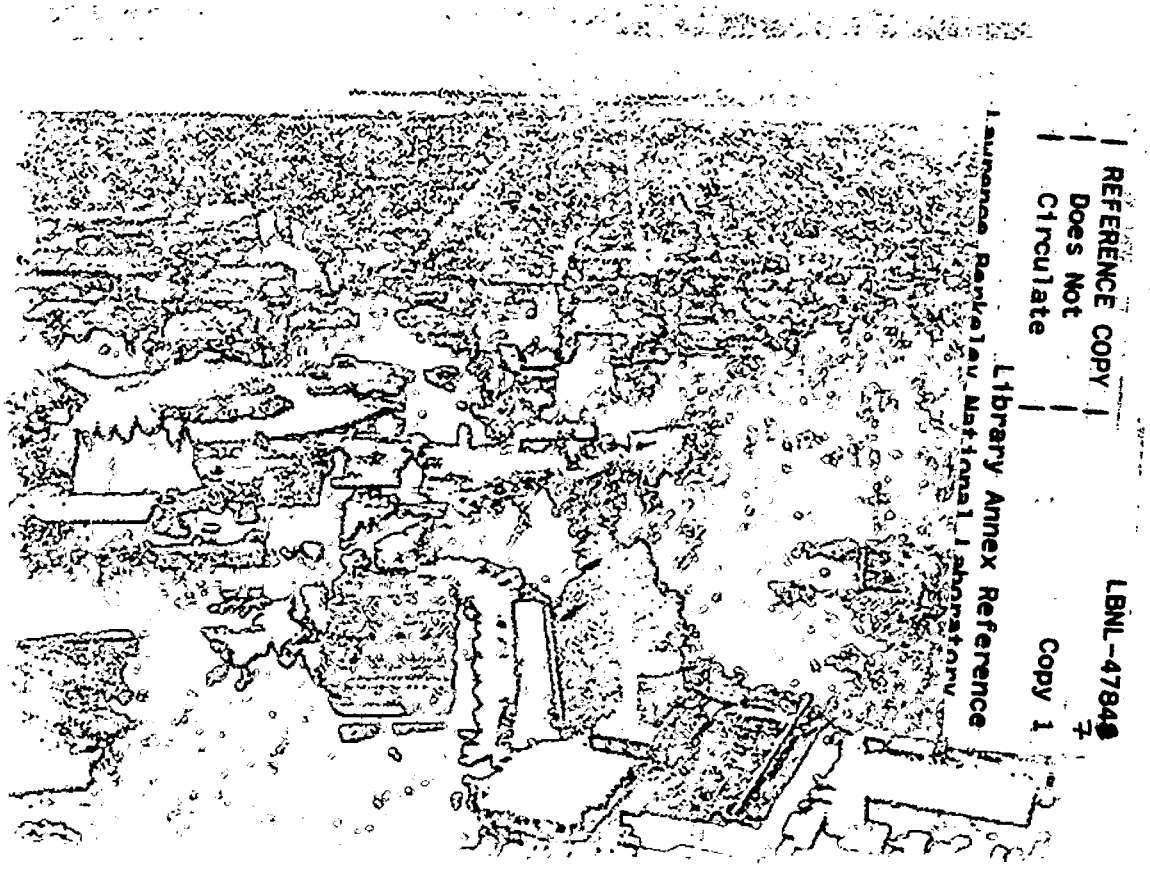
ERNEST ORLANDO LAWRENCE
BERKELEY NATIONAL LABORATORY

Analysis of Hubs and Authorities on the Web

Chris Ding, Hongyuan Zha, Xiaofeng He,
Parry Husbands, and Horst Simon

**National Energy Research
Scientific Computing Division**

May 2001


REFERENCE COPY
Does Not
Circulate
Library Annex Reference
LBNL-47847
Copy 1

DISCLAIMER

This document was prepared as an account of work sponsored by the United States Government. While this document is believed to contain correct information, neither the United States Government nor any agency thereof, nor the Regents of the University of California, nor any of their employees, makes any warranty, express or implied, or assumes any legal responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by its trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or the Regents of the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof or the Regents of the University of California.

LBNL-47847

Analysis of Hubs and Authorities on the Web

Chris Ding,[a] Hongyuan Zha,[b] Xiaofeng He,[a,b]
Parry Husbands,[a] and Horst Simon[a]

[a]National Energy Research Scientific Computing Division
Ernest Orlando Lawrence Berkeley National Laboratory
University of California
Berkeley, California 94720

[b]Department of Computer Science and Engineering
Pennsylvania State University
University Park, PA 16802

May 2001

This work was supported in part by the Director, Office of Science, Office of Laboratory Policy and Infrastructure, of the U.S. Department of Energy under Contract No. DE-AC03-76SF00098, and by National Science Foundation Grant No. CCR-9001986.

Analysis of Hubs and Authorities on the Web

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

May 7, 2001

Abstract

Ranking tens of thousands of retrieved webpages for a user query on a internet search engine so that most informative (authoritative) or popular (hub) webpages are on the top 10 or 20 is an key information retrieval technology. Kleinberg's HITS algorithm represents a major advance in relevance ranking algorithms. It explores the reinforcing interplay between authority and hub webpages on a particular topic by taking into account the structure of the web graphs formed by the hyperlinks bewteen the webpages. In this paper, we give a detailed analysis of the HITS algorithm using a unique combination of matrix algebra and probabilistic analysis. In particular, we show that in the avarage case the ranking given by the HITS algorithm is the same as the ranking obtained by using in-bound and out-bound hyperlink counts. Using web graphs of different sizes, we also provide experimental results to illustrate our analysis.

1 Introduction

The rapidly growing World Wide Web now contains more than two billion webpages of text, images and various multimedia information. While this vast amount of information has the potential to benefit all aspects of our society, finding the relevant webpages to satisfy a user's information need still remains to be a very important and challenging task. Many commercial search engines have been developed and used by millions of people all over the world. However, the relevancy of webpages returned in search engine result sets is still lacking, and further research and development is needed to really make search engines a ubiquitous information-seeking tool.

A distinct feature of the web is the proliferation of hyperlinks between webpages which allow a user to surf from one webpage to another with a simple click. This hyperlink structure contains very useful information. If webpage p_i has a link pointing to webpage p_j, it indicates that the creator of p_i considers p_j containing relevant information for webpage p_i. Such thoughtful and often unbiased (exceptions do exist) opinions are therefore registered in the form of hyperlinks. A webpage pointed to by a large number of hyperlinks (the degree of in-bound hyperlinks, usually refered to as the in-degree) is probably more valuable or informative than another webpage pointed to by a smaller number of hyperlinks. Thus in-degree is in general a good indication of the quality of a webpage.

*NERSC Division, Lawrence Berkeley National Laboratory, University of California, Berkeley, CA 94720, {chqding,pjrhusbands,hdsimon}@lbl.gov. This work is supported in part by Office of Science, Office of Laboratory Policy and Infrastructure, of the U.S. Department of Energy under contract number DE-AC03-76SF00098 through an LBL LDRD grant.

†Department of Computer Science and Engineering, The Pennsylvania State University, University Park, PA 16802, {zha,xhe}@cse.psu.edu. The work was supported in part by NSF grant CCR-9901986.

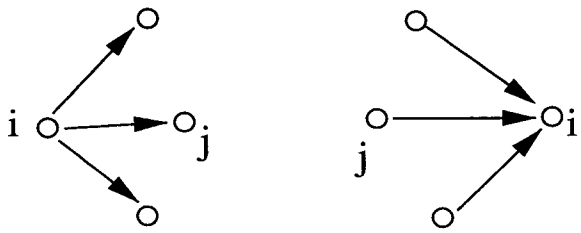


Figure 1: Left: hub webpage p_i has many out-bound hyperlinks. Right: authority webpage p_i has many in-bound hyperlinks.

Kleinberg developed a very popular ranking algorithm: the *Hypertext Induced Topic Selection* (HITS) algorithm that explores the hyperlink information in order to improve search engine retrieval relevancy [15]. HITS utilizes the directionality of the hyperlinks and makes the crucial distinction of *hubs* and *authorities*. Intuitively, within a set of webpages dealing with a particular topic, an *authority* webpage is one with a large number of webpages in the set pointing to it and a *hub* page is one that points to a large number of webpages in the set. The HITS algorithm, however, improves on this basic idea: it assigns respective scores to hubs and authorities, and computes them in a mutually reinforcing way: an authority must be pointed to by several good hubs (i.e., webpages with large hub scores) while a hub must be pointed by several good authorities (i.e., webpages with large authority scores). This learning process is iterated several times to rearch equilibrium scores for the hubs and authorities. Further improved versions are also developed [9, 3, 6]. The ranking giving by the HITS algorithm and that obtained directly using inbound and outbound link counts are closely related as has been observed by [15, 3]. Some authors even advocate directly using some weighted version of the inbound and outbound link counts without any iterations [16]. The goal of this paper is to give a detailed analysis of the HITS algorithm.

2 HITS Algorithm

In the HITS algorithm, each webpage p_i is assigned 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* and a good *hub* points to many good authorities. This mutually reinforceing relationship is represented as,

$$x_i = \sum_{j:e_{ji}\in E} y_j, \quad y_i = \sum_{j:e_{ij}\in E} x_j \tag{1}$$

E is the set of hyperlinks (edges in the web graph). Iteratively update the authority and hub scores of every web page, using Eq.(1), and sort the web pages in decreasing order according to their authority and hub weights, respectively, we can obtain the authorities and hubs of the webpage set.

A matrix and vector representation better describe the process. The link information is obtained directly from the link graph. The set of webpages forms a directed graph $G = (V, E)$, where a webpage p_i is a node ($p_i \in V$) and a hyperlink e_{ij} is an edge ($e_{ij} \in E$). The link matrix L of the directed graph is defined to be: $L_{ij} = 1$ if $e_{ij} \in E$, 0 otherwise. L is also called adjacency matrix of the graph. The authority scores on all n nodes form vector $x = (x_1, x_2, \cdots, x_n)$ and the hub scores

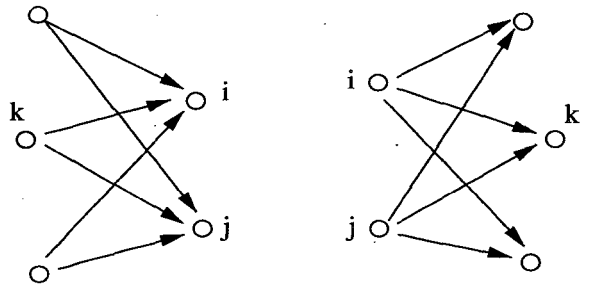


Figure 2: Left: webpages p_i, p_j are co-cited by webpage p_k. Right: webpages p_i, p_j co-reference webpage p_k.

form vector $y = (y_1, y_2, \cdots, y_n)$. With these notations, Eqs.(1) can be cast into

$$x = L^T y, \quad y = Lx.$$

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

$$cx^{(t+1)} = L^T L x^{(t)}, \quad cy^{(t+1)} = L L^T y^{(t)} \tag{2}$$

starting with

$$x^{(0)} = y^{(0)} = (1, 1, 1, \cdots, 1).$$

where, c is a normalization constant so that $||x|| = ||y|| = 1$. For this reason, we call $L^T L$ the authority matrix and LL^T the hub matrix. One can easily see that the final solution x^*, y^* are the principal eigenvectors of the symmetric positive definite matrices $L^T L$ and LL^T : $L^T L x^* = \lambda x^*$ and $LL^T y^* = \lambda y^*$, i.e., we seek to find the largest triplet $\{\lambda, x^*, y^*\}$ of L. It is clear that the HITS iteration process (2) is just the *power method* for computing the largest singular value triplet of L. Once we obtain x^* and y^*, we can rank the webpages according to their hub scores and authority scores, returning to the user a list of hubs and authorities [15].

3 Authority vs. Hub and Co-citation vs. Co-reference

We analyze the structure of the Authority and Hub matrices and derive several interesting analytical results defined in the above section and make connections to two important concepts in bibliometrics: co-citation and co-reference. Co-citation and bibliographic coupling (we will refer it as co-reference) are first proposed in the fields of citation analysis and bibliometrics as fundamental metrics to characterize the similarity between two documents [18, 14]. We concentrate on clarifying the relationship between authority matrix and co-citation, and that between hub matrix and co-reference.

If two distinct webpages i, j are co-cited by many other webpages, as in Fig. 2, webpages i, j are likely to be related in some way. Thus co-citation is a similarity measure. It is defined as the number of webpages that co-cite webpages i, j. The co-citation between two webpages p_i, p_j can be calculated as

$$C_{ij} = \sum_k L_{ki} L_{kj} = \sum_k (L^T)_{ik} L_{kj} = (L^T L)_{ij} \tag{3}$$

Note that the self-citation C_{ii} is not defined, and is usually set to $C_{ii} = 0$. Also, C_{ij} is symmetric, $C_{ij} = C_{ji}$.

Let us count the in-degree of webpage p_i. It is given by

$$d_i = \sum_k L_{ki} = \sum_k L_{ki} L_{ki} = (L^T L)_{ii} \tag{4}$$

because $L_{ki} = L_{ki}^2$, since $L_{ki} = 0, 1$. Let D be the diagonal matrix of in-degrees,

$$D = \text{diag}(d_1, d_2, \cdots, d_n) \tag{5}$$

we see that the link structure of $L^T L$ is

$$L^T L = D + C. \tag{6}$$

Thus the authority matrix is the sum of co-citation and in-degree. This result is a mathemetical statement on the close relationship between authority and co-citation, and reveals the important role of in-degree which is further examined in later sections. One can also see that

$$\max(0, d_i + d_k - n) \leq C_{ik} \leq \min(d_i, d_k). \tag{7}$$

As in Fig. 2, the fact that two distinct webpages p_i, p_j co-reference many other webpages indicates that p_i, p_j have certain commonality. Co-reference (bibliometric 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 two webpages i, j, calculated as (see Fig. 2),

$$R_{ij} = \sum_k L_{ik} L_{jk} = \sum_k L_{ik} (L^T)_{kj} = (L L^T)_{ij} \tag{8}$$

The self-reference R_{ii} is not defined, and is set to $R_{ii} = 0$. The out-degree of node p_i is

$$o_i = \sum_k L_{ik} = \sum_k L_{ik} L_{ik} = (L L^T)_{ii}. \tag{9}$$

Let $O = \text{diag}(o_1, o_2, \cdots, o_n)$, we have

$$L L^T = O + R, \tag{10}$$

the hub matrix is the sum of co-reference and out-degree, revealing the close relationship between hubs and co-references. We also have the inequality

$$\max(0, o_i + o_k - n) \leq R_{ik} \leq \min(o_i, o_k). \tag{11}$$

It is interesting to note the duality relationship between hubs and authorities, between co-citations and co-references.

4 Probabilistic analysis

We seek to analyze the structures of the authority and hub matrices in finer granularity. Our results of Eq.(6) suggests an interesting and useful observation on the relationship of co-citations and in-degree: in general, nodes with large in-degrees will have large co-citations with other nodes, simply

because they have more in-links. Conversely, large co-citations are directly related to the in-degrees of the nodes involved. This intuition can be made more precise by using some probablistic analysis. Specifically, we prove that

$$\langle C_{ik} \rangle = d_i d_k/(n-1), \tag{12}$$

where $\langle C_{ik} \rangle$ is the *average* of C_{ik} under random distribution. This is consistent with Eq.7. Suppose $d_i \geq d_k$. There are at most d_k nonzero terms in Eq.3, which is the product of elements in i^{th} and k^{th} columns of adjacency matrix L. Consider the case where q^{th} row in k^{th} column is one (not zero). The probability that the corresponding position in i^{th} column being 1 is

$$P(L_{qi} = 1) = C_{n-2}^{d_i - 1}/C_{n-1}^{d_i} = d_i/(n-1). \tag{13}$$

Here $C_{n-1}^{d_i}$ is the total number of possible patterns for d_i ones in i^{th} column, and $C_{n-2}^{d_i-1}$ is the total number of possible patterns given that there is a one at row q. Thus $\langle C_{ik} \rangle = \sum_q \langle L_{qi} L_{qk} \rangle = \sum_q^{d_k} \langle L_{qi} \rangle = d_k \cdot P(L_{qi} = 1)$, we have Eq.12.

From these analysis, we see that node i with large in-degree d_i will have large co-citations with other nodes, compared to node j with a smaller in-degree d_j. i.e., if $d_i > d_j$, we have

$$\langle C_{ik} \rangle \; > \; \langle C_{jk} \rangle, \quad \forall k, k \neq i, k \neq j. \tag{14}$$

From this probabilistic equation, C_{ik} is larger than C_{jk} most of time, but not necessarily true in every cases. For convenience, we often say that the inequality $C_{ik} \gtrsim C_{jk}$ holds *on average.*

The same probablistic analysis can be applied to out-degree and co-reference for hub matrix LL^T . We have

$$\langle R_{ik} \rangle = o_i o_k/(n-1). \tag{15}$$

If $o_i > o_j$, we have $\langle R_{ik} \rangle \; > \; \langle R_{jk} \rangle$, which we say that $R_{ik} \gtrsim R_{jk}$ *on average.*

5 Average case analysis

Generally speaking, the web graph is a random graph — millions of individials, organizations, develop their webpages for different purposes. For this reason, we perform analysis for the average case, i.e., the hub and authority matrices are replaced by their average values: $< L^T L >=< D > + < C >$. Using Eq.12, we have the average case authority matrix

$$< L^T L >= \hat{D} + < C >= \text{diag}(\hat{d}_1, \ldots, \hat{d}_n) + dd^T/(n-1),$$

where $\hat{d}_i = d_i - d_i^2/(n-1)$ and $d = (d_1, d_2, \cdots, d_n)^T$. Now $< L^T L >$ is the sum of a diagonal matrix and a rank-one matrix whose eigendecomposition is known (Theorem 8.5.3 in Golub and van Loan [11]). Theorem 8.5.3 requires that $\hat{d}_1 > \hat{d}_2 > \cdots > \hat{d}_n$. This is satisfied if we index the webpages according to their in-degrees, $d_1 > d_2 > \ldots > d_n$, and make the assumption that

$$d_i + d_j < n - 1 \tag{16}$$

for all i and j.[1] Then it follows $\hat{d}_i - \hat{d}_j = (d_i - d_j)(1 - (d_i + d_j)/(n-1)) > 0$ if $i < j$. Theorem 8.5.3 has the following two main results:

[1]This will be satisfied if $d_i < (n-1)/2$ for all i, which is reasonable if the set of webpages is large enough.

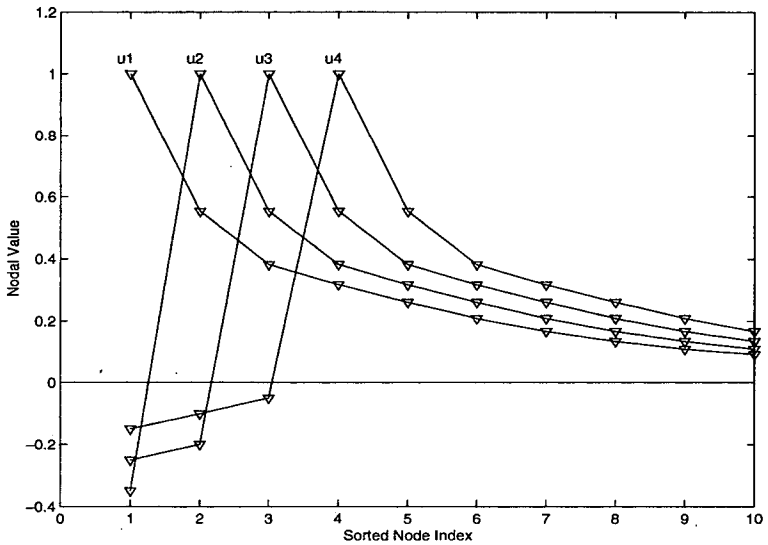


Figure 3: Eigenvectors of Eq.(18).

1. The n eigenvalues have the following interleave relation,

$$\lambda_1 > \hat{d}_1 > \lambda_2 > \hat{d}_2 > \cdots > \hat{d}_n, \tag{17}$$

2. The kth eigenvector is

$$\mathbf{u}_k = (\frac{d_1}{\lambda_k - \hat{d}_1}, \frac{d_2}{\lambda_k - \hat{d}_2}, \ldots, \frac{d_n}{\lambda_k - \hat{d}_n})^T. \tag{18}$$

6 Properties of HITS Algorithm

Several interesting consequences follow from the above analysis:

1. **Webpage ordering.** The authority ranking is approximately the same as a ranking according to webpage in-degrees. To see this, we show that authority scores (nodal values of the principal eigenvector) are monotonically decreasing. In fact,

$$u_1(i) - u_1(j) = \frac{d_i}{\lambda_1 - \hat{d}_i} - \frac{d_j}{\lambda_1 - \hat{d}_j} = \frac{(d_i - d_j)[\lambda_1 - d_i d_j/(n-1)]}{(\lambda_1 - \hat{d}_i)(\lambda_1 - \hat{d}_j)} > 0,$$

since

$$\lambda_1 - d_i d_j/(n-1) > \hat{d}_i - d_i d_j/(n-1) = d_i(1 - (d_i + d_j)/(n-1)) > 0,$$

using Eq.16, all other 3 factors are positive. It is interesting that the idea of mutual reinforcement between hubs and authorities leads to this conclusion. Indeed, this feature is highly consistent with our intuition. A good authority should have a large number of webpages pointing to it, just as a seminal paper is often cited by a large number of later research papers.

The eigenvectors behave fairly regularly, as illustrated in Figure 3. $\mathbf{u}_1$ is always positive. For $\mathbf{u}_2$, the first node is negative, turning positive from the second node. For $\mathbf{u}_3$, the first 2 nodes are negative, turning positive from the third node and so on.

2. **Uniqueness.** If d_1 is larger than d_2, then the principal eigenvector of L^TL is unique, and quite different from second principal eigenvector (see Figure 3). If we start HITS iteration with an arbitrary initial vector, we are guaranteed to converge to the principal eigenvector.

3. **Convergence.** The convergence for HITS can be rather fast. In fact, using the starting vector $\mathbf{x}^{(0)} = (1, 1, \cdots, 1)$ which has very little overlap with non principal eigenvectors $(\mathbf{x}^{(0)} \cdot \mathbf{u}_k, k > 1)$ because they all contain negative node values (see Figure 3). Using the spectral expansion of $L^TL = \lambda_1 \mathbf{u}_1 \mathbf{u}_1^T + \lambda_2 \mathbf{u}_2 \mathbf{u}_2^T + \cdots$, after t iterations, we have

$$x^{(t)} = c_1 \lambda_1^t \mathbf{u}_1 + c_2 \lambda_2^t \mathbf{u}_2 + \cdots \tag{19}$$

   where $c_2 \ll c_1$ because of small overlap between $\mathbf{x}^{(0)}$ and $\mathbf{u}_2$. For random graphs, it is well-known that in-degrees and out-degrees follow power-law distributions [10, 5]. Also, the eigenvalues typically follows a Zipf distribution [8]. This implies the ratio $\lambda_2/\lambda_1 \simeq 1/2$. Thus the iteration converges rapidly. Typically 5-10 iterations are sufficient.

4. **Web communities.** An important aspect of HITS algorithm is to identify multiple web communities using different eigenvectors [15, 9]. The principal eigenvector defines a dominant web community. Each non-principal eigenvector defines two communities, one with non-negative values $\{i | \mathbf{u}_k(i) \geq 0\}$ and the other with negative values $\{i | \mathbf{u}_k(i) < 0\}$.

   From the pattern of eigenvectors in our solutions (see Fig. 3), the negative region of eigenvector $\mathbf{u}_k$ have large overlap with the negative region of another eigenvector $\mathbf{u}_\ell$. This happens for positive regions as well. Therefore, we believe this method to identify multiple communnities is not as instructive. This difficulty is also noticed in practical applications [3].

   A better way to identify web communities is to use unsupervised learning techniques such as clustering. In this case, a similarity metric is necessary to define the goal or objective function of clustering. As discussed above, the co-citation matrix could serve as the similarity metric. Other information could be incorporated as well. In a recent study[13], we found that incorporating additional link structure and text information improve the qualify of clustering substantially.

7 Experimental results

In our experiments with HITS, we first noticed the high correlation between HITS rankings and the rankings by degree which motived this analytical study. Here we give experimental results on running HITS on two webpage datasets. We note that this high correlations are also noticed in the study of Amento, et al,[1] and mentioned implicitly in Bharat and Henzinger[3].

Experiment 1. This dateset was supplied by the Internet Archive [2] 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 principal eigenvectors were obtained using PARPACK [17] on NERSC's IBM SP computer. The table below shows the list of the top 20 authorities, ranked by HITS (1st column) and by in-degree (2nd column).

`Authority Ranking`

Hits	In	URL
1	4	www.yahoo.com
2	3	www.geocities.com
3	1	www.microsoft.com
4	6	members.aol.com
5	2	home.netscape.com
6	10	www.excite.com
7	11	www.lycos.com
8	9	members.tripod.com
9	15	ourworld.compuserve.com
10	5	www.netscape.com
11	20	www.cnn.com
12	28	www.webcom.com
13	33	sunsite.unc.edu
14	7	www.adobe.com
15	35	www.teleport.com
16	17	www.altavista.digital.com
17	25	www.w3.org
18	19	www.infoseek.com
19	18	www.angelfire.com
20	21	www.hotbot.com
...	..	.....
111	13	www.linkexchange.com
137	14	ad.linkexchange.com
174	17	member.linkexchange.com

In general, one see that the HITS ranking and in-degree rank are highly correlated, as expected from our analytical results. For these reasons, we consider as *normal* those webpages highly ranked by HITS that also have high in-degree. There are two types of webpages that deviate from this general pattern and are theoretically interesting : (a) those highly ranked authority webpages by HITS, but with relatively smaller in-degrees, and (b) those webpages with large in-degrees, but ranked low by HITS. These webpages would have been incorrectly ranked if we simply count in-degrees, thus represents the net improvements brought by HITS algorithm.

As for type (b) webpages, we note that three websites *www.linkexchange.com*, *ad.linkexchange.com*, and *member.linkexchange.com* that ranked high by in-degree (rank 13, 14, 16 respectively). They ranked quite low by HITS (rank 111, 137, 174 respectively). All three sites have very large in-degrees, but also very small out-degrees; they are all *sinks*: many sites point to them, but they do not point to anywhere. The mutually re-inforcing nature of the HITS algorithm ranked them low, because there are no good hubs pointing to them. These anomalies demonstrate the effectiveness of the HITS algorithm.

As for type (a) webpages, we mention two websites: (1) *sunsite.unc.edu*, which is ranked 13 in HITS, but is ranked 33 by in-degree. This site holds many software repositories, but few out-bound links. Its higher HITS ranking is likely because more top sites such as microsoft point to it. (2)

www.teleport.com, which is ranked 15 in HITS, but is ranked 35 by in-degree. This site has a large number of out-links, and more top sites point to it. The following table lists the top hubs, ranked by HITS (1st column) and by out-degree (2nd column).

Hub Ranking

Hits	Out	URL
1	4	www.yahoo.com.au
2	5	www.yahoo.co.uk
3	3	dir.yahoo.com
4	7	www.yahoo.com.sg
5	8	www.yahoo.ca
6	9	www2.aunz.yahoo.com
7	1	members.aol.com
8	2	www.geocities.com
9	6	members.tripod.com
10	10	ispc.yahoo.co.uk
11	11	y3.yahoo.ca
12	12	y4.yahoo.ca
13	13	www6.yahoo.co.uk
14	16	tv.yahoo.com.au
15	17	www.yahoo.co.nz
16	19	soccer.yahoo.com.au
17	18	www.yahoo.com.my
18	21	www.aunz.yahoo.com
19	20	203.103.130.22
20	23	206.222.66.43

Here one see very high correlation between the HITS ranking and out-degree ranking, indicating that our approximate analytical results are fairly accurate in this case.

We note, however, that the distinction between hubs and authorities are sometime blurred. Good examples are *members.aol.com, www.geocities.com*, etc. they are ranked very high in both authority list and hub list. Although they are not authoritative on any particular subject, careful content selection and organization on these websites make them valuable, almost like authoritative figures. This also happens in the bibliometrics domain, some good survey papers/books (hubs) become as valuable or important as the original seminar papers (authorities), especially because these good surveys are written by authoritative people in the field, and the additional insights they provide in the survey documents.

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(ODP) *www.dmoz.org*. Under each category of the ODP, there is a relatively focused topic. The data file from the ODP contains the hierarchical structure of these webpages. A *Perl* program is used to generate the linkgraph of these webpages. We form the linkgraph of subcategory *Running* by extracting from the *Fitness* linkgraph the document IDs of those webpages

under *Running* sub-category. The table below shows the list of the top 20 authorities, ranked either by HITS (1st column) or by in-degree (2nd column).

Authority Ranking

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

Here the correlation between the HITS ranking and the in-degree ranking is high. If we organize the results in top 10, second top 10, etc., as done by many internet search engines, the match within top 10, and second top 10 are fairly close. The following table lists the top hubs, ranked by HITS (1st column) and by out-degree (2nd column).

Hub Ranking

Hits	Out	URL
1	3	www.fix.net/~doogie/links.html
2	1	www.gbtc.org/whatelse.html
3	4	www.usateamsports.com/running.htm
4	15	home1.gte.net/gregtrrc/links.htm
5	17	www.afn.org/~ftc/othlinks.html
6	19	www.grainnet.com/rdraces/websites.html
7	14	www.runner.org/links.htm
8	20	directory.netscape.com/Health/Fitness/Running
9	21	www.dmoz.org/Health/Fitness/Running/
10	20	directorysearch.mozilla.org/Health/Fitness/Running/
11	15	dmoz.org/Health/Fitness/Running

12	25	www.cajuncup.com/links.htm
13	11	www.rrm.com/sites.html
14	18	www.doitsports.com/guides/running.html
15	20	www.webcrawler.com/kids_and_family/hobbies/outdoors/running
16	20	magellan.mckinley.com/lifestyle/hobbies_and_recreation/outdoors/...
17	28	www.webfanatix.com/running_resources.htm
18	28	www.webfanatix.com/_vti_bin/shtml.exe/running_resources.htm/map
19	25	www.isp.nwu.edu/~brianw/running.html
20	23	www.geocities.com/HotSprings/Resort/5457/

For the hub ranking, correlation between the HITS ranking and the in-degree ranking is not as high as for the authority. but still apparent, especially if we look at top 3.

8 Discussions and summary

Although the HITS algorithm is motivated by the mutual reinforcement between hubs and authorities, we can arrive at the HITS algorithm from a different perspective. Note that the authority matrix is essentially a similarity metric between different authority webpages. In determining the weight for each webpage, we use the following weight-propagation idea similar to that used in PageRank [4]. If webpage p_i is pointed to by a good webpage p_j (with large authority score), p_i is likely to be valuable. On the other hand, if webpage p_i is pointed to by a poor webpage p_j (with small authority score), p_i is not likely to be valuable. The connection strength between p_i, p_j is their similarity, $(L^T L)_{ij}$. Thus the weight-propagation equation is

$$cx^{(t+1)} = L^T L x^{(t)}$$

c is a normalization constant. This is exactly Eq.2, the HITS algorithm.

Besides finding hubs and authorities on the web, HITS algorithm is also used in finding authoritative documents in document databases[7, 12]. Our results apply there too, since the underlying theory is based entirely on the analysis of the directed graph which are identical in these domains.

Our results have implications on current search engine technology. Instead of building a web subgraph among the retrieved webpages for a user query and then run the HITS algorithm, one can simply count the in-degree and out-degree and returned the webpages ranked by the degrees. The method of [16] is the a reasonable approach. Note that this ranking is query dependent, which differs from the static global ranking. This appears to be implemented by Google (although the PageRank score in Google may differ from authority score). Of course, it's possible that some interesting webpages will not rank high as they would in HITS ranking.

In summary, we have analyzed the HITS algorithm and derived several insightful relationships between hubs and co-reference and between authorities and co-citations. We perform probablistic analysis and average case analysis which shed much lights into the HITS algorithm.

References

[1] B. Amento, L. Terveen, and W. Hill. Does *authority* mean quality? predicting expert quality ratings of web documents. *Proc. ACM Conf. on Research and Development in Information Retrieval (SIGIR'00)*, pages 296–303, 2000.

[2] Internet Archive. *http://www.archive.org/.*

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

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

[5] 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.

[6] 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.

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

[8] C.H.Q. Ding. A probabilistic model for dimensionality reduction in information retrieval and filtering. *Submitted to Journal of the ACM*, 2000.

[9] 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.

[10] S. Glassman. A caching relay for the world wide web. *Comput. Networks ISDN System*, 27:165–175, 1994.

[11] G. Golub and C. V. Loan. *Matrix Computations, 3rd edition.* Johns Hopkins, Baltimore, 1996.

[12] A.K. McCallum H. Chang, D. Cohn. Learning to create customized authority lists. *Proc. ICML 2000. pp.127-134.*, 2000.

[13] X. He, H. Zha, C. Ding, and H.D. Simon. Web document clustering using hyperlink structures. *Tech Report CSE-01-006. Submitted*, 2001.

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

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

[16] R. Lempel and S. Moran. The stochastic approach for link-structure analysis (salsa) and the tkc effect. *Proceedings of WWW9*, 2000.

[17] K. J. Maschhoff and D. C. Sorensen. A portable implementation of arpack for distributed memory parallel computers. In *Preliminary Proceedings of the Copper Mountain Conference on Iterative Methods*, 1996.

[18] 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.

Ernest Orlando Lawrence Berkeley National Laboratory
One Cyclotron Road Berkeley, California 94720

Prepared for the U.S. Department of Energy under Contract No. DE-AC03-76SF00098