Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2014, Article ID 418605, 6 pages
http://dx.doi.org/10.1155/2014/418605

Research Article

A Model Based on Cocitation for Web Information Retrieval

Yue Xie and Ting-Zhu Huang

School of Mathematical Sciences, University of Electronic Science and Technology of China, Sichuan, Chengdu 611731, China

Correspondence should be addressed to Ting-Zhu Huang; tingzhuhuang@126.com

Received 16 August 2013; Revised 13 January 2014; Accepted 13 January 2014; Published 27 February 2014

Academic Editor: Masoud Hajarian

According to the relationship between authority and cocitation in HITS, we propose a new hyperlink weighting scheme to describe the strength of the relevancy between any two webpages. Then we combine hyperlink weight normalization and random surfing schemes as used in PageRank to justify the new model. In the new model based on cocitation (MBCC), the pages with stronger relevancy are assigned higher values, not just depending on the outlinks. This model combines both features of HITS and PageRank. Finally, we present the results of some numerical experiments, showing that the MBCC ranking agrees with the HITS ranking, especially in top 10. Meanwhile, MBCC keeps the superiority of PageRank, that is, existence and uniqueness of ranking vectors.

1. Introduction

In the past, search engines ranked pages by using word frequency or similar measures. However, the relevancy of webpages returned by this traditional web information retrieval is still lacking, because the webpages are created with varying qualities. Recently, some new algorithms have been created that greatly improve rankings. One of the popular ideas is to use hyperlinks to determine the value of different webpages. This hyperlink graph contains useful information: if webpage i has a link pointing to webpage j, it usually indicates that the creator of i considers j to contain relevant information for i. Such useful opinions and knowledge are therefore registered in the form of adjacency matrix which is denoted by L. $L_{ij} = 1$ if there is a link from i to j, or 0, otherwise.

Two most popular ranking algorithms based on hyperlink analysis are the PageRank algorithm [1, 2] and the HITS (Hyper-text Induced Topic Selection) algorithm [3]. Generally, PageRank considers the hyperlink weight normalization and the equilibrium distribution of random surfers as the citation score. For more information about the calculation methods of PageRank refer to [4–6]. HITS makes the distinction between hubs and authorities and then computes them in a mutually reinforcing way. For each of these two algorithms, the ranking vector is the dominant eigenvector of some matrix describing the network. How this matrix is defined differs in each method. There are other works which have recognized that the hyperlink structure can be very valuable for locating information [3, 7, 8].

This paper is organized as follows. In Section 2, we introduce the PageRank and HITS algorithms and briefly discuss the limitations in HITS. Then in Section 3, we emphasize the role of cocitation (Figure 1) and provide a hyperlink weighting scheme to describe the strength of the relevancy between any two webpages. In order to ensure the existence of solutions and uniqueness of solutions in the new model (MBCC), we also combine ideas from PageRank. In Section 4, some experiments are presented. The result shows that the MBCC ranking is close well to the HITS ranking. Conclusions are given in Section 5.

2. PageRank and HITS

We treat the web as a directed graph $D = (V, E)$: the nodes in V correspond to the pages, and a directed edge $(i, j) \in E$ indicates the existence of a link from i to j. We say that the out-degree of a node i denoted by $d_{\text{out}}(i)$ is the number of nodes it has links point to, and the in-degree of i denoted by $d_{\text{in}}(i)$ is the number of nodes that have links point to it. We also denote that

$$d_{\text{out}} = \left(d_{\text{out}}(1), \ldots, d_{\text{out}}(n)\right)^T,$$
$$d_{\text{in}} = \left(d_{\text{in}}(1), \ldots, d_{\text{in}}(n)\right)^T. \quad (1)$$

2.1. Review of PageRank. PageRank [1, 2] uses a web surfing model based on a random walk process. Suppose there is a link from page i to page j; that is, $(i, j) \in E$. Consider a random surfer visiting page i at time t. Then at the next time $t+1$, the surfer lands at page j with probability $1/d_{\text{out}}(i)$. Once the above is done, the PageRank algorithm assigns a rank value x_i for the page i as a function of the rank of the pages that point to it:

$$x_i = \sum_{(j,i)\in E} \frac{r_i}{d_{\text{out}}(j)}. \tag{2}$$

If the page i has no outlink, that is, $d_{\text{out}}(i) = 0$, then, at time $t + 1$, the surfer chooses any page with probability $1/n$. Thus, we replace $d_{\text{out}}(i) = 0$ with $d_{\text{out}}(i) = n$. Then the stationary distribution x is determined by the following matrix form:

$$x = P^T x, \quad P^T = \left(L + de^T\right)^T D_{\text{out}}^{-1}. \tag{3}$$

Here $x = (x_1, \ldots, x_n)^T$, L is the adjacency matrix of the directed web graph, $D_{\text{out}} = \text{diag}(d_{\text{out}})$, and $e = (1, \ldots, 1)^T$. In the vector d, the element $d_i = 1$ if the ith row of L corresponds to a dangling node ($d_{\text{out}}(i) = 0$), or 0, otherwise.

In order to calculate the above recursive equation and get a unique stationary probability distribution, it is important to guarantee that (3) is convergent. This problem can be solved if the directed graph D is strongly connected, which is generally not the case for the directed graph. In the context of computing PageRank, the standard way of ensuring this property is to add a new set of complete outgoing transitions, with small transition probabilities (in this work, we set each of them as $1/n$), to all nodes in D. Then the modified transition probability called Google matrix is

$$G^T = \alpha\left(L + de^T\right)^T D_{\text{out}}^{-1} + (1 - \alpha)\left(\frac{1}{n}\right) ee^T, \tag{4}$$

where $\alpha = 0.8 \sim 0.9$. Here $e = (1, \ldots, 1)^T$; thus ee^T is a matrix of all 1's. The PageRank algorithm is to solve the eigenvector of the Google matrix G^T

$$x = G^T x, \qquad \sum_{i=1}^{n} x_i = 1, \tag{5}$$

where G^T is stochastic and irreducible.

PageRank models two types of random jumps on the Internet. With probability $1 - \alpha$ a surfer randomly chooses a new page. Otherwise, the surfer follows one of directed edges from the present node.

2.2. Review of HITS. In the HITS algorithm [3], each webpage i has both a hub score y_i (based on the links going from the page) and an authority score x_i (based on the links going to the page). Let $x = (x_1, \ldots, x_n)^T$ denote the vector of all authority weights, let $y = (y_1, \ldots, y_n)^T$ denote the vector of all hub weights, and let L be the adjacency matrix of the directed web graph. In HITS, there are two operations at each iteration. One is defined as operation $\mathcal{I}$ which sets the authority vector to $x = L^T y$. It indicates that a good authority is pointed by many good hubs. Another is defined as operation $\mathcal{O}$ which sets the hub vector to $y = Lx$. It indicates that a good hub points to many good authorities. This mutually reinforcing relationship can be written in the following matrix representations:

$$\begin{aligned} x &= \frac{1}{\lambda^*} L^T L x, \qquad \sum_{i=1}^{n} x_i^2 = 1, \\ y &= \frac{1}{\lambda^*} L L^T y, \qquad \sum_{i=1}^{n} y_i^2 = 1. \end{aligned} \tag{6}$$

The final authority and hub scores are the principal eigenvectors of $L^T L$ and LL^T which are corresponding to the dominant eigenvalue λ^*. Since $L^T L$ and LL^T determine the authority ranking and hub ranking, we call $L^T L$ the authority matrix and LL^T the hub matrix.

In the fields of citation analysis and bibliometrics, it has shown that the authority matrix has interesting connections to cocitation [3]. Here cocitation is defined as the number of webpages that cocite i, j [9]. In the authority matrix, $(L^T L)_{ii} = \Sigma_{k=1}^{n} L_{ki} L_{ki} = \Sigma_{k=1}^{n} L_{ki}$ is the in-degree of page i; that is,

$$\left(L^T L\right)_{ii} = d_{\text{in}}(i). \tag{7}$$

This implies that

$$\text{diag}\left(L^T L\right) = D_{\text{in}}. \tag{8}$$

For $i \neq j$, $(L^T L)_{ij} = \Sigma_{k=1}^{n} L_{ki} L_{kj}$ is the number of webpages that cocite i, j that is denoted by C_{ij}. Therefore the authority matrix $L^T L$ is the sum of in-degree and cocitation [10, 11]

$$L^T L = D_{\text{in}} + C. \tag{9}$$

The self cocitation C_{ii} in C is not defined and is usually set to 0.

2.3. Existence and Uniqueness of Ranking Vectors. In this section, we present the existence and uniqueness of ranking vectors in the above two algorithms.

Since the Google matrix G^T in (4) is stochastic and irreducible, for the PageRank algorithm, the PageRank ranking vector exists, and it is unique and positive. See the equivalent theorem in [12, Theorem 3.8]. For the HITS algorithm, it has been proved that the hub and authority ranking vectors exist but may not be unique. In [12], they show that the HITS algorithm badly behaved on certain networks, meaning that (i) it can return ranking vectors that are not unique but depend on the initial seed vector or (ii) it can return ranking vectors that inappropriately assign zero weights to parts of the network.

There are also other limitations for HITS; see [12, 13]. Thus, to address these limitations, a modification for HITS is needed, for example, exponentiated input method in [12]. In the next section, we combine both features of HITS and PageRank. The ranking produced by the new model is expected to be unique and close to the HITS ranking.

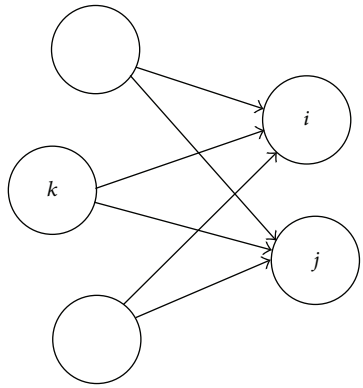

Figure 1: Webpages i, j are cocited by webpage k.

3. A Model Based on Cocitation (MBCC)

In HITS, according to (9), the authority ranking value x_i can be expressed as

$$x_i = \frac{1}{\lambda^*} d_{\text{in}}(i)\, x_i + \frac{1}{\lambda^*} \sum_{j=1}^{n} C_{ij} x_j, \tag{10}$$

revealing the close relationship between authorities and cocitations. It also implies that, if two distinct webpages i, j are cocited by many other webpages k as shown in Figure 1, then i, j are likely to be related in some sense. In this paper, we present a property for HITS corresponding to (10).

Property 1 (relationship between authority value and cocitation). If the number of webpages that cocite webpages i and j, that is, C_{ij}, is larger, the page i could receive more authority value from the page j, even though there are no links between i and j.

The fact that the webpages cocite two distinct webpages i and j indicates that i, j have certain commonality. Therefore, we say that the number of cocitations represents the relevancy among the pages. Then, in the following, we focus on the use of cocitation for analyzing the relevancy among the pages.

Note that, in Section 2.1, the rank of a page in PageRank is divided among its forward links evenly; see (2); that is, a web surfer could chose the forward outlinks randomly. However, this process of dividing the rank equally may seem unrealistic; that is, a web surfer may have a priori idea of the value of pages, favoring pages from the relevant sites. Since it shows that the number of cocitations could represent the relevancy among the pages, we say that the number of cocitations between two pages can impact the behavior of web surfers. Therefore, we define a new hyperlink weighting scheme based on cocitation as follows:

Definition 2 (hyperlink weighting scheme based on cocitation). Let Q_{ij} be the number of webpages that cocite two webpages i, j. Specially, $Q_{ii} = d_{\text{in}}(i)$, and $d_{\text{in}}(i)$ is the in-degree of webpage i. Then we define the following function as the value of j which will receive form i:

$$W_{ij} = \frac{Q_{ij}}{\sum_{k \in V} Q_{ik}} = \frac{Q_{ij}}{Q_i}, \tag{11}$$

where $Q_i = \sum_{k \in V} Q_{ik}$.

Table 1: The data of Example 1.

	(j_1, j_1)	(j_1, j_2)	(j_1, j_3)
Q	3	3	1
W	3/7	3/7	1/7

Under this assignment method, the rank value for the page j is determined by

$$x_j = \sum_{m \in V} W_{mj} x_m. \tag{12}$$

The matrix form of above equation is $x = Wx$, where $x = (x_1, \ldots, x_n)^T$. The problem is that, if at least one page has zero in-degree, that is, no in-links and $Q_i = 0$, then the matrix W is absorbing and its dominant eigenvector does not exist. In order to resolve this, similarly to PageRank, we assume that, if the page i has no link that points to it, then at time $t + 1$, the page i divides its value equally to any other page with probability $1/n$. The modified matrix W is given by

$$W = \left(L'L + ve^T\right)^T D_Q^{-1} = \left(L'L + ev^T\right) D_Q^{-1}, \tag{13}$$

where we replace $Q_i = 0$ with $Q_i = n$, $D_Q = \text{diag}(\widetilde{Q})$ and $\widetilde{Q} = (Q_1, \ldots, Q_n)^T$. $\widetilde{Q}$ can be computed as

$$\widetilde{Q} = \left(L'L + ve^T\right) e. \tag{14}$$

In the vector v, the element $v_i = 1$ if the i-th row of L corresponds to a page with no in-degree, or 0, otherwise. Therefore, the modified matrix W becomes a stochastic matrix, that is, each column in W sum to 1.

In order to get a unique stationary probability distribution, it is important to guarantee that W is strongly connected. Similarly to PageRank, we add a new set of complete outgoing transitions. The final transition probability matrix based on using cocitation as a hyperlink weighting scheme is

$$W = \beta \left(L'L + ev^T\right) D_Q^{-1} + (1 - \beta) \left(\frac{1}{n}\right) ee^T, \tag{15}$$

where $0 < \beta < 1$ and $e = (1, \ldots, 1)^T$. The model based on cocitation (MBCC) is to solve the following function:

$$x^* = \left\{\beta \left(L'L + ev^T\right) D_Q^{-1} + (1 - \beta) \left(\frac{1}{n}\right) ee^T\right\} x^*, \quad \sum_{i=1}^{n} x_i = 1. \tag{16}$$

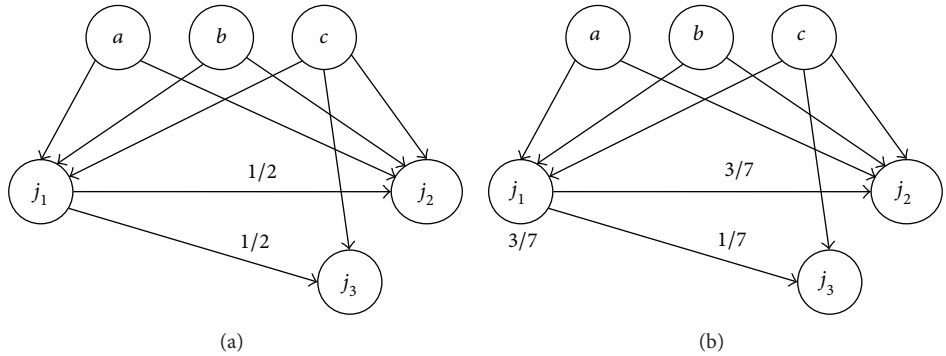

Figure 2: (a) Link structure of PageRank. (b) Resigned link structure of MBCC.

Table 2: HITS authority ranking, MBCC authority ranking, and PageRank ranking in top 20.

HITS	MBCC	PageRank	URL
1	1	9	http://www.ics.uci.edu/~eppstein/geom.html
2	2	10	http://www.geom.uiuc.edu/software/cglist/
3	3	6	http://www.cs.uu.nl/CGAL
4	4	18	http://www.ics.uci.edu/~eppstein/junkyard
5	6	14	http://www.scs.carleton.ca/~csgs/resources/cg.html
6	7	20	http://www.geom.uiuc.edu/
7	5	>20	http://www.ics.uci.edu/~eppstein
8	8	>20	http://www.mpi-sb.mpg.de/LEDA/leda.html
9	10	>20	http://www.inria.fr/prisme/personnel/bronnimann/cgt
10	9	>20	http://www.cs.sunysb.edu/~algorith/
11	13	>20	http://graphics.lcs.mit.edu/~seth/pubs/taskforce/techrep.html
12	11	>20	http://www.cs.smith.edu/~orourke
13	12	>20	http://www.cs.brown.edu/people/rt
14	18	>20	http://www.geom.uiuc.edu/~nina/
15	>20	>20	http://compgeom.cs.uiuc.edu/~jeffe/compgeom
16	14	>20	http://www.cs.princeton.edu/~chazelle
17	>20	>20	http://www.dcc.unicamp.br/~guialbu/geompages.html
18	16	8	http://www.yahoo.com
19	>20	>20	http://www.ics.uci.edu/~eppstein/gina/authors.html
20	>20	>20	http://www.cs.brown.edu/people/rt/sdcr/report.html

We assume that the solution of (16) denoted by $x^* = (x_1^*, \ldots, x_n^*)$ is the MBCC authority ranking vector, and $y = LL^T x^*$ is the MBCC hub ranking vector. Since the matrix W in (15) is stochastic and irreducible, just like the Google matrix G in PageRank, the solution of (16) exists, and it is unique and positive.

4. Numerical Experiments

First, we present an example to describe the assignment process in Definition 2.

Example 1. Suppose that there are six webpages $V = (j_1, j_2, j_3, a, b, c)$, and the directed graph is shown in Figure 2. The conclusion can be found from Table 1 and Figure 2. In Table 1, $Q(i, j)$ is the number of webpages that cocite webpages i and j; $W(i, j)$ is obtained by (11). In Figure 2, the left one is the original link structure of PageRank where the value of the page j_1 is divided equally to the pages that it points to, and the right one divides the value of j_1 based on cocitation.

Then, we compare the MBCC model with HITS and PageRank, experimenting with dataset from http://www.cs.toronto.edu/~tsap/experiments/datasets/. The dataset is about the topic computational geometry which contains a total of 1100 webpages. We set $\beta = 0.9$. Meanwhile, we use $\tau = 10^{-10}$ as the convergence tolerance and measure the convergence rates of the three algorithms using the L1 norm of the residual vector. Table 2 shows the list of the top 20 authorities with HITS, MBCC, and PageRank. Table 3 shows the list of the top 20 hubs with HITS and MBCC. It shows that MBCC authority ranking is closer to HITS

Table 3: HITS hub ranking and MBCC hub ranking in top 20.

HITS	MBCC	URL
1	2	http://mother.lub.lu.se/ae/bytitle/043501-043550.html
2	1	http://compgeom.cs.uiuc.edu/~jeffe/compgeom/preprehistory.html
3	3	http://www.dcc.unicamp.br/~guialbu/geompages.html
4	4	http://www.softlab.ece.ntua.gr/~cfrag
5	5	http://corelab.cs.ntua.gr/courses/compgeom
6	6	http://compgeom.cs.uiuc.edu/~jeffe/compgeom/direct.html
7	7	http://mother.lub.lu.se/ae/bydomain/010101-010150.html
8	9	http://www-sop.inria.fr/prisme/personnel/bronnimann/cgt/WWW.html
9	10	http://members.tripod.com/~GeomWiz/develop.html
10	11	http://geomwiz.tripod.com/develop.html
11	8	http://www.scs.carleton.ca/~csgs/resources/cg.html
12	13	http://www.ams.sunysb.edu/~jsbm/hotlist.html
13	14	http://www.graphics.lcs.mit.edu/~fredo/Book/geoAlgo.html
14	16	http://www.cs.uwaterloo.ca/~yganjali/r_links.html
15	12	http://cs.smith.edu/~streinu/bookmarks.html
16	15	http://www.cs.umn.edu/scg98
17	>20	http://forum.swarthmore.edu/library/topics/comp_geom
18	19	http://www.geom.umn.edu/~mucke/GeomDir/people.html
19	>20	http://cis.poly.edu/~aronov
20	20	http://intra.cmkos.cz/~honza/osvr/odkazy.html

Table 4: Comparison between MBCC and HITS ranking vectors, for example, top 10 represents a ranking vector agreeing with another ranking vector in top 10.

	Top 10	Top 20	Top 30	Top 40	Top 50
Authority (HITS MBCC)	100%	80%	96.7%	85%	90%
Hub (HITS MBCC)	90%	90%	90%	92.5%	96%

authority ranking than PageRank ranking which is close to HITS authority ranking. The comparison between MBCC and HITS ranking vectors in Table 4 indicates that MBCC ranking agrees well with HITS ranking, especially in top 10.

5. Conclusion

In this work, we emphasize the role of cocitation in defining authorities. First, we observe that, in the HITS algorithm, if two distinct webpages i, j are cocited by many other webpages k, then i, j are likely to be related in some sense or have certain commonality. According to this close relationship, we come to the conclusion that the higher the number of webpages that cocite webpages i and j, the stronger the relevancy between the two pages. The page j with stronger relevancy should obtain more values from page i. Therefore, we develop a hyperlink weighting scheme for extracting information from the link structure. Then we combine hyperlink weight normalization and random surfing schemes as used in PageRank to justify the model.

The experimental results show that the MBCC authority (hub) ranking is close well to the HITS authority (hub) ranking in top 20, and in general a surfer seldomly browses beyond these webpages in top 20 [11]. Moreover, MBCC keeps the superiority of PageRank: the authority vector of MBCC in (16) exists, and it is unique and positive, while the authority and hub vectors of HITS may not be unique. Therefore, we can use the authority (hub) ranking vector of MBCC as the authority (hub) ranking vector of HITS.

Conflict of Interests

The authors declare that they have no conflict of interests regarding the publication of this paper.

Acknowledgment

This research is supported by NSFC (61370147, 61170309), Chinese Universities Specialized Research Fund for the Doctoral Program (20110185110020).

References

[1] S. Brin, L. Page, R. Motwami, and T. Winograd, "The PageRank citation ranking: bringing order to the web," Tech. Rep. 1999-0120, Computer Science Department, Stanford University, Stanford, Calif, USA, 1999.

[2] S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine," *Computer Networks and ISDN Systems*, vol. 30, no. 1–7, pp. 107–117, 1998.

[3] J. M. Kleinberg, "Authoritative sources in a hyperlinked environment," *Journal of the ACM*, vol. 46, no. 5, pp. 604–632, 1999.

[4] K. Avrachenkov, N. Litvak, D. Nemirovsky, and N. Osipova, "Monte Carlo methods in PageRank computation: when one

iteration is sufficient," *SIAM Journal on Numerical Analysis*, vol. 45, no. 2, pp. 890–904, 2007.

[5] D. F. Gleich, A. P. Gray, C. Greif, and T. Lau, "An inner-outer iteration for computing RageRank," *SIAM Journal on Scientific Computing*, vol. 32, no. 1, pp. 349–371, 2010.

[6] R. S. Wills and I. C. F. Ipsen, "Ordinal ranking for Google's PageRank," *SIAM Journal on Matrix Analysis and Applications*, vol. 30, no. 4, pp. 1677–1696, 2009.

[7] A. N. Langville and C. D. Meyer, "A reordering for the PageRank problem," *SIAM Journal on Scientific Computing*, vol. 27, no. 6, pp. 2112–2120, 2006.

[8] S. D. Kamvar, T. H. Haveliwala, C. D. Manning, and G. H. Golub, "Exploiting the block structure of the web for computing PageRank," Tech. Rep. 2003-17, Stanford University, Stanford, Calif, USA, 2003.

[9] H. Small, "Co-citation in the scientific literature: a new measure of the relationship between two documents," *Journal of the American Society for Information Science*, vol. 24, no. 4, pp. 265–269, 1973.

[10] C. Ding, H. Zha, X. He, P. Husbands, and H. Simon, "Analysis of hubs and authorities on the web," Tech. Rep. 47847, Lawrence Berkeley National Laboratory, 2001.

[11] C. Ding, X. He, P. Husbands, H. Zha, and H. Simon, "PageRank, HITS and a unified framework for link analysis," Tech. Rep. 49372, Lawrence Berkeley National Laboratory, 2001.

[12] A. Farahat, T. Lofaro, J. C. Miller, G. Rae, and L. A. Ward, "Authority rankings from HITS, PageRank, and SALSA: existence, uniqueness, and effect of initialization," *SIAM Journal on Scientific Computing*, vol. 27, no. 4, pp. 1181–1201, 2006.

[13] A. N. Langville and C. D. Meyer, "A survey of eigenvector methods for web information retrieval," *SIAM Review*, vol. 47, no. 1, pp. 135–161, 2005.