PageRank, HITS and Impact Factor for Journal Ranking

Su Cheng, Pan YunTao, Yuan JunPeng, Guo Hong, Yu ZhengLu, Hu ZhiYu
Institute of Scientific and Technical Information of China (ISTIC), Beijing (P. R. China)

Abstract

Journal citation measures are one of the most widely used bibliometric tools. The most well-known measure is the ISI Impact Factor, under the standard definition, the impact factor of journal j in a given year is the average number of citations received by papers published in the previous two years of journal j. However, the impact factor has its "intrinsic" limitations, it is a ranking measure based fundamentally on a pure counting of the in-degrees of nodes in the network, and its calculation does not take into account the "impact" or "prestige" of the journals in which the citations appear. Google's PageRank algorithm and Kleinberg's HITS method are webpage ranking algorithm, they compute the scores of webpages based on a combination of the number of hyperlinks that point to the page and the status of pages that the hyperlinks originate from, a page is important if it is pointed to by other important pages. We demonstrate how popular webpage algorithm PageRank and HITS can be used ranking journal, and we compared ISI impact factor, PageRank and HITS for journal ranking, and with PageRank and HITS compute respectively including self-citation and non self-citation, and discussed the merit and shortcomings and the scope of application that the various algorithms are used to rank journal.

1. Introduction

Journal citation measures are one of the most widely used bibliometric tools. They are used in information retrieval, scientific information, library science and research evaluation[1]. The main source of journal citation measures is the annually appearing Journal Citation Report (JCR) which has now become the most important publication of Institute for Scientific Information (ISI) .The most well-known measure is the Impact Factor (Garfield, 1979)[2]. Under the standard definition, the impact factor of journal j in a given year is the average number of citations received by papers published in the previous two years of journal j. Because of its comprehensibility and its fast availability, the impact factor became very quickly popular and widely used. However, the impact factor has its "intrinsic" limitations, it is a ranking measure based fundamentally on a pure counting of the in-degrees of nodes in the network, and its calculation does not take into account the "impact" or "prestige" of the journals in which the citations appear[3][4][5].

Google's PageRank is a popular webpage ranking algorithm, it views a hyperlink as a recommendation[6]. Thus a page with more recommendations (which are realized through inlinks) must be important than a page with a few inlinks. However, similar to other recommendation systems such as journal citations, the status of the recommender is also important. A page is important if it is pointed to by other important pages. Google's PageRank algorithm computes the scores of webpages based on a combination of the number of hyperlinks that point to the page and the status of pages that the hyperlinks originate from[7].

Kleinberg's HITS method for ranking webpages is very similar to PageRank, but is uses both inlinks and outlinks to create two popularity scores for each page. HITS defines hubs and authorities. HITS has a pair of interdependent circular theses: a page is a good hub (and therefore deserves a high hub score) if it points to good authorities. And a page is a good authority if it is pointed to by a good hub[8].

Although there are differences between Journal Citation network and web's hyperlink, we can define Journal Citation network and web's hyperlink as a directed graph. The nodes in the graph represent journals or webpages and the directed arcs or links represent the citations or hyperlinks.

So it is reasonable that we adopt PageRank and HITS for journal ranking. PageRank and HITS, as an alternative to the ISI's impact factor, undoubtedly presents considerable advantages, such as its calculation take into account the "impact" or "prestige" of the journals in which the citations appear.

We use the dataset of the 2006 China Scientific and Technical Papers and Citations database (CSTPC) to compare the ISI Impact Factor, PageRank and HITS ranking of journals, and with PageRank and HITS compute respectively including self citation and

978-0-7695-3507-4/08 $25.00 © 2008 IEEE
DOI 10.1109/CSIE.2009.351

excluding self citation, and discussed the merit and shortcomings and the scope of application that the various algorithms are used to rank journal.

2. Data and method

2.1 Data

Institute of Scientific and Technical Information of China (ISTIC), an institute under the ministry of Science and Technology of China, has built China Scientific and Technical Papers and Citations database (CSTPC) in 1987[9], CSTPC is based on representative domestic Scientific and Technical journals. In 1988, CSTPC covered 1189 journals, while in 2006, the source journals totaled 1723.

We used the dataset of 2006 CSTPC, a Journal Citation Network was constructed on the basis of the 2006 CSTPC data set which contains 2006 journal citations to 2005 and 2004 publications. This journal citation information was represented as a 1723×1723 matrix in which both rows and columns represent journals, and in which cells represent the amount of times a journal in a row cites a journal in a column.

2.2 Journal citation network and Webpages hyperlink struture

In webpages hyperlink structure, there exists at most an edge from node *i* to *j*, In the Journal Citation Network, however, not all edges are created equal; some journals are connected by more citations than others. So there can exists *n* edge from journal *i* to *j*. webpage hyperlink structure can be written:

$$Lij = \begin{cases} 1, \text{if there exists an edge from node } i \text{ to } j \\ 0, \text{otherwise} \end{cases} \qquad (2.2.1)$$

journal citation network can be written:

$$Lij = \begin{cases} n, \text{if there exists } n \text{ edge from node } i \text{ to } j \\ 0, \text{otherwise} \end{cases} \qquad (2.2.2)$$

In Journal Citation Network, including self-citation(citations from the Journal to the Journal) and non self-citation (citations from other journals to the Journal), Almost all journals have self citations, in CSTPCD 2006, the average rate of self-citation is 19.89%. In hyperlink structure, a webpage has no connection self.

In short, there are two differences between the matrix of web's hyperlink and Journal citation network, one is in Journal citation network one journal can have more citations than others. The other is Journal citation network have self-citation. So the PageRank or HITS equation when applied to journal citation networks should therefore be adapted to take into account journal citation frequencies and self-citation in its transfer of PageRank or HITS values.

2.3 Algorithms for journal ranking

2.3.1 The ISI Impact factor

Interest in classifying or "measuring" scientific research is not a recent phenomenon: one of the first classifications was proposed by Gross (1927). Nevertheless, the criterion of measuring the "impact" of scientific publications was suggested by Garfield (1955), and published in the journal science, and "impact factor" was first used for quantifying publications in the 1963 edition of the Science Citation Index (SCI). This index was initially published in a supplement of the SCI, under the name Journal Citation Reports (JCR), and it has now become the most important publication of Institute for Scientific Information (ISI)[3].

Journal citation measures are one of the most widely used bibliometric tools. The main source of journal citation measures is the annually appearing Journal Citation Report (JCR) which has now become the most important publication of Institute for Scientific Information (ISI) .The most well-known measure is the Impact Factor (Garfield, 1979) [2]. The impact factor for the journal J in the year n is defined as the ratio

$$IF_n(J) = \frac{c_n}{p_{n-1} + p_{n-2}} \qquad (2.3.1.1)$$

Where cn is the number of citations received in the year n by papers published in the journal *J* in the years *n-1* and *n-2* and the total number of source items ($p_{n-1}+p_{n-2}$) published in the journal J in these two years (*n-1 and n-2*). We observe that the ISI impact factor is a ranking measure based fundamentally on a pure counting of the in-degrees of nodes in the network.

2.3.2 PageRank for Journal citation network

Brin and Page, the inventors of PageRank, began with a simple summation equation, The PageRank of a page P_i, denoted $r(P_i)$, is the sum of the PageRanks of all pages pointing into P_i[6].

$$r(P\,i) = \sum_{Pj \in BPi} \frac{r(Pj)}{|Pj|} \qquad (2.3.2.1)$$

Where B_{P_i} is the set of pages pointing into P_i, and $|P_j|$ is the number of outlinks from page P_j, notice that the PageRank of inlinking pages $r(P_j)$ in equation (2.3.2.1) is tempered by the number of recommendations made by P_j, denoted $|P_j|$. The problem with equation (2.3.2.1) is that $r(P_j)$ values, the PageRanks of pages inlinking to page P_i, are unknown. To sidestep this problem, Brin and Page used an iterative procedure. That is, they assumed that, in the beginning, all pages have equal PageRank. Now the rule in equation (2.3.2.1) is followed to computer $r(P_i)$for each page P_i in the index. The rule in equation (2.3.2.1) is successively applied, substituting the value of the previous iterate into $r(P_j)$. Iterative procedure is repeated with the hope that the PageRank scores will eventually converge to some final stable values.

Equation (2.3.2.1) compute PageRank one page at a time. Using matrices, we replace the tedious Σ symbol, and at each iteration, compute a PageRank vector, which uses a single 1×n vector to hold the PageRank values for all pages in the index.

$$\boldsymbol{\pi}^{(k+1)T} = \boldsymbol{\pi}^{(k)T}\ \mathbf{H} \qquad (2.3.2.2)$$

Equation (2.3.2.2) have several problems, for example, there is the problem of rank sinks, those pages that accumulate more and more PageRank at each iteration, monopolizing the scores and refusing to share. So Google's adjusted PageRank method is[7]

$$\boldsymbol{\pi}^{(k+1)T} = \boldsymbol{\pi}^{(k)T}\ (\alpha\mathbf{H} + (\alpha a + (1-\alpha)e)\ 1/n\ e^T \qquad (2.3.2.3)$$

$\pi^{(k)T}$ PageRank vector at the kth iteration
H very sparse, raw substochastic hyperlink matrix
α scaling parameter between 0 and 1
a binary dangling node vector
e^T the row vector of all 1s

Since logically the prestige obtained by a journal is the result of the prestige obtained by its articles, so it could be compared the prestige average per article without having in mind other factors like the frequency of each journal, the number articles. For journal citation network, the PageRank scores for the journal *J* in the year *n* is defined as the ratio

$$PR_n(J) = \frac{\pi_n^{(k)T}}{p_{n-1} + p_{n-2}} \qquad (2.3.2.4)$$

Where $\pi_n^{(k)T}$ is the PageRank vector at the kth iteration, $p_{n-1}+p_{n-2}$ is the papers published in the journal *J* in the year *n-1* and *n-2*.

2.3.3 HITS for Journal citation network

HITS, which is an acronym for Hypertext Induced Topic Search, was invented by Jon Kleinberg in 1998. HITS, like PageRank, uses the web's hyperlink structure to create popularity scores associated with webpages. However, HITS has some important differences. Whereas the PageRank method produces one popularity score for each page, HITS produce two. Whereas the PageRank is query-independent, HITS is query-dependent. HITS thinks of webpages as authorities and hubs. An authority is a page with many inlinks, and a hub is a page with many outlinks[7].

$$x_i^{(k)} = \sum_{j:e_{ji}\in E} y_j^{(k-1)}$$
$$y_i^{(k)} = \sum_{j:e_{ij}\in E} x_j^{(k)}$$

for $k = 1, 2, 3, \ldots$ (2.3.3.1)

These equations, which were Kleinberg's original equations, can be written equation (2.2.1) in matrix form with the help of the adjacency matrix L of the directed web graph.

In matrix notation, the equation in (2.3.3.1) assume the form

$$x^{(k)}=L^T y^{(k-1)} \text{ and } y^{(k)}=Lx^{(k)} \qquad (2.3.3.2)$$

Where x(k) and y(k) are n×1 vectors holding the approximate authority and hub scores at each iteration.

The equation (2.3.3.2) can be simplified by substitution to

$$x^{(k)}=L^T L x^{(k-1)}$$
$$y^{(k)}= LL^T y^{(k-1)}$$

These two new equations define the iterative power method for computing the dominant eigenvetor for the matrices $L^T L$ and LL^T. The matrix $L^T L$ determines the authority scores, the matrix LL^T determines the hub scores.

In Journal citation network, the journal includes self-citation (citations from the Journal to the Journal) and non self-citation (citations from other journals to the Journal), and in the Journal Citation Network, however, not all edges are created equal; some journals are connected by more citations than others. So the matrix *J* of Journal citation network can be written equation (2.2.2). For journal citation network, the HITS authority and hub scores for the journal J in the year n is defined as the ratio

$$HITS_{xn}(J) = \frac{x_n^{(k)}}{p_{n-1} + p_{n-2}}$$

$$HITS_{yn}(J) = \frac{y_n^{(k)}}{p_{n-1} + p_{n-2}} \qquad (2.3.3.3)$$

Where $x_n^{(k)}$ and $y_n^{(k)}$ are the approximate authority and hub scores of journal *J* in the year *n*, $p_{n-1}+p_{n-2}$ is the papers published in the journal *J* in the years *n-1* and *n-2*.

3. Results

In this section, we use equation (2.3.1.1), equation (2.3.2.4) and equation (2.3.3.3) to compute the scores of ISI IF, PageRank and HITS, and compare these three algorithms for Journal Citation Network. In order to inspect the influence of the journal self citation to the journal ranking, we compute respectively the value of journal including self-citation and non self-citation.

3.1 Comparing the ISI IF,PageRank,HITS

Table 1 shows the top 10 ranking journals according to ISI IF, PageRank and PageRank non self-citation, Table 2 shows the top 10 ranking journals according to HITS authority score, HITS authority score non self-citation, HITS hub score and HITS hub score non self citation. Clearly, the rankings diverge significantly, There is not one journal to have appeared in the all lists. Only two journals, “PEDOSPHERE” and “Acta Geographica Sinica” are represented in IF and PR lists. Only two journals, “Power System Technology” and “Proceedings of the Chinese Society for Electrical Engineering” are represented in IF and HITS authority score lists. No journal is represented in PR and HITS authority score lists.

Analyzing the value of PageRank and HITS including self-citation and non self-citation shows that the sensitivity of PageRank and HITS with respect to self-citation. PageRank is very sensitive to self-citation, but the HITS is not sensitive to self-citation. From table 1 we can find that the average rate of self-citation of the top 10 journal PageRank scores including self-citation is 26.9%, the average rate of self-citation of the top 10 journal PageRank scores non self-citation is 7.9%, from table 2 we can find that the average rate of self-citation of the top 10 journal HITS authority score non self-citation is 36.3%, the average rate of self-citation of the top 10 journal HITS authority score including self-citation is 35.4%. Generally speaking, the rate of self-citation of top 10 journal according to PageRank scores are less than top 10 journal according to HITS.

We also can detect that the top 10 journal according to PageRank include some subject, for example, Medicine, Computer science, Geology, mathematics, geography. But the top 10 journal according to HITS are nearly electricity, this is a disadvantage of HITS, because in journal citation network include a very authoritative electricity journal, This very authoritative journal can carry so much weight that it and its neighboring journals dominate the relevant ranked list.

Table 1: the top 10 ranking journals according to ISI IF, PageRank and PageRank non self-citation

	IF			PR			PR non self-citation		
Rank	Value	The rate of self-citation	Journal	Value ($\times10^5$)	The rate of self - citation	Journal	Value ($\times10^6$)	The rate of self-citation	Journal
1	2.86	0.54	Power System Technology	1.24	0.54	ASIAN JOURNAL OF ANDROLOGY	3.01	0.03	Prog in Phys
2	2.65	0.41	Acta Petrologica Sinica	1.13	0.34	Acta Anthropologica Sinica	2.88	0.07	J Syst Engi
3	2.59	0.76	Exp Tech Mana	1.03	0.06	Journal of Software	2.85	0.09	J Syst Manag
4	2.54	0.48	Proc Chin Soc Elect Engi	0.99	0.28	PEDOSPHERE	2.75	0.07	Acta Geogr Sin
5	2.46	0.49	J Desert Res	0.97	0.11	Ocean et Limno Sin	2.51	0.06	Journal of Software
6	2.44	0.32	China Journal of Highway and Transport	0.89	0.49	New Carbon Materials	2.49	0.06	Sci in China(D)
7	2.33	0.28	PEDOSPHERE	0.87	0.51	Vertebrata Palasiatica	2.47	0.21	Ann Shanghai Astron Obser Chin Acad Sci
8	2.33	0.19	Acta Geol Sin	0.85	0.2	ACTA MATH SIN ENG SER	2.26	0.08	Biodiversity Sci
9	2.3	0.07	Acta Geogr Sin	0.82	0.09	Acta Meteo Sin	2.19	0.04	Adv In Mecha
10	2.21	0.23	Chin J Geology	0.81	0.07	Acta Geogr Sin	2.18	0.08	Math Numer Sin

Table 2:The top 10 ranking journals according to HITS authority score, HITS authority score non self-citation, HITS hub score and HITS hub score non self-citation

HITS authority score	HITS authority score non self citation	HITS hub score	HITS hub score non self citation

rank	Value $\times 10^5$	the rate of self citation	Journal	Value $\times 10^5$	the rate of self citation	Journal	Value $\times 10^5$	the rate of self citation	Journal	Value $\times 10^5$	the rate of self citation	Journal
1	37.83	0.54	Power System Technology	38.49	0.48	Proceedings of the Chinese Society for Electrical Engineering	38.25	0.54	Power System Technology	42.22	0.54	Power System Technology
2	36.39	0.48	Proceedings of the Chinese Society for Electrical Engineering	10.68	0.42	Automation of Electric Power Systems	28.78	0.48	Proceedings of the Chinese Society for Electrical Engineering	8.81	0.15	Transactions of China Electrotechnical Society
3	6.03	0.42	Automation of Electric Power Systems	7.28	0.54	Power System Technology	5.14	0.42	Automation of Electric Power Systems	7.36	0.1	Modern Electric Power
4	2.69	0.67	High Voltage Engineering	5.33	0.67	High Voltage Engineering	4.42	0.67	High Voltage Engineering	7.05	0.34	Proceedings of the Chinese Society of Universities
5	2.13	0.15	Transactions of China Electrotechnical Society	4.50	0.37	Advanced Technology of Electrical Engineering and Energy	3.84	0.15	Transactions of China Electrotechnical Society	6.81	0.67	High Voltage Engineering
6	2.05	0.37	Advanced Technology of Electrical Engineering and Energy	3.62	0.27	Power System Protection and Control	3.52	0.34	Proceedings of the Chinese Society of Universities	6.78	0.28	Journal of North China Electric Power University
7	1.47	0.19	Chinese Journal of Power Engineering	3.51	0.11	Electric Power	3.40	0.27	Power System Protection and Control	5.96	0.27	Power System Protection and Control
8	1.36	0.11	Electric Power	3.10	0.34	Proceedings of the Chinese Society of Universities	3.38	0.1	Modern Electric Power	5.68	0.28	Electric Power Automation Equipment
9	1.23	0.27	Power System Protection and Control	2.63	0.28	Electric Power Automation Equipment	3.19	0.28	Journal of North China Electric Power University	5.57	0.42	Automation of Electric Power Systems
10	1.14	0.34	Proceedings of the Chinese Society of Universities	2.60	0.15	Transactions of China Electrotechnical Society	3.12	0.28	Electric Power Automation Equipment	3.99	0.37	Advanced Technology of Electrical Engineering and Energy

4. Discussions and conclusion

Journal ranking is similar to webpage ranking, PageRank and HITS can be adapted to rank journal. From the result of ranking journal we can detect that PageRank is fit for whole graph and HITS is fit for neighborhood graph. For HITS, if a subject has one or two very authoritative journal, this journal can carry so much weight that it and its neighboring journals dominate the relevant ranked list. So HITS is not fit for whole graph of journal.

Generally speaking, the rate of self-citation of top 10 journal according to PageRank is less than top 10 journal according to Impact factor. If compute respectively the PageRank values including self-citation and non self-citation, we can find that the rankings diverge significantly,

In short, PageRank and HITS, as an alternative to the ISI's impact factor, undoubtedly presents considerable advantages, such as they can measure the "prestige" of journal better than Impact factor, but we must use these algorithms carefully.

5. Acknowledge

This work is supported by the Ministry of Science and technology of China under contract 2006BAH03B05 and National Natural Science Foundation of China (70673019). The authors thank Xiong Ning for her assistance in polishing article.

6. References

[1] W. Glanzel, *BIBLIOMETRICS AS A RESEATCH FIELD,* 2003

[2] Eugene Garfield. *Citation Indexing: Its Theory and Application in Science, Technology, and Humanities.* John Wiley and Sons, New York, 1979.

[3] Gualberto Buela-Casal. Assessing the Quality of Articles and Scientific Journals: Proposal for Weighted Impact Factor. *Psychology in Spain*,8(1):60-76,2004

[4] Amin, M. & Mabe, M. (2000). Impact factor: use and abuse. *Perspectives in Publishing,* 1, 1-6.

[5] Johan Bollen, Marko A. Rodriguez, Herbert V. de Sompel, Journal Status, *http://arxiv.org/abs/cs.DL/0601030*

[6] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems,* 30(1–7):107–117, 1998.

[7] A. Langville, C. Meyer, *Google's PageRank and Beyond: The Science of Search Engine Rankings.* PRINCETON UNIVERSITY PRESS, 2006

[8] Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. *In Proceedings of the 9th ACMSIAM Symposium on Discrete Algorithms,* pp 668–677, Baltimore, MD, 1998.

[9] Yishan Wu,Yuntao Pan, Yuhua Zhang, Zheng Ma, Jingan Pang, Hong Guo, Bo Xu, Zhiqing yang. China Scientific and Technical Papers and Citations (CSTPC): History, impact and outlook. *Scientometrics,* 60(3):385-397, 2004