Emerald Insight

Aslib Proceedings

ArticleRank: a PageRank-based alternative to numbers of citations for analysing citation networks

Jiang Li Peter Willett

Article information:

To cite this document:
Jiang Li Peter Willett, (2009),"ArticleRank: a PageRank-based alternative to numbers of citations for analysing citation networks", Aslib Proceedings, Vol. 61 Iss 6 pp. 605 - 618
Permanent link to this document:
http://dx.doi.org/10.1108/00012530911005544

Downloaded on: 04 February 2015, At: 10:53 (PT)
References: this document contains references to 21 other documents.
To copy this document: permissions@emeraldinsight.com
The fulltext of this document has been downloaded 523 times since 2009*

Users who downloaded this article also downloaded:

Olga Tsepilova, (2007),"Forging Change in a Contaminated Russian City: A Longitudinal View of Kirishi", Research in Social Problems and Public Policy, Vol. 14 pp. 31-46

Access to this document was granted through an Emerald subscription provided by 240104 []

For Authors

If you would like to write for this, or any other Emerald publication, then please use our Emerald for Authors service information about how to choose which publication to write for and submission guidelines are available for all. Please visit www.emeraldinsight.com/authors for more information.

About Emerald www.emeraldinsight.com

Emerald is a global publisher linking research and practice to the benefit of society. The company manages a portfolio of more than 290 journals and over 2,350 books and book series volumes, as well as providing an extensive range of online products and additional customer resources and services.

Emerald is both COUNTER 4 and TRANSFER compliant. The organization is a partner of the Committee on Publication Ethics (COPE) and also works with Portico and the LOCKSS initiative for digital archive preservation.

*Related content and download information correct at time of download.

Downloaded by Yale University At 10:53 04 February 2015 (PT)

The current issue and full text archive of this journal is available at
www.emeraldinsight.com/0001-253X.htm

ArticleRank: a PageRank-based alternative to numbers of citations for analysing citation networks

Received 4 June 2009
Revised 4 July 2009
Accepted 21 August 2009

Jiang Li
Department of Information Management, Nanjing University, Nanjing, China, and
Peter Willett
Department of Information Studies, University of Sheffield, Sheffield, UK

Downloaded by Yale University At 10:53 04 February 2015 (PT)

Abstract

Purpose – The purpose of this paper is to suggest an alternative to the widely used Times Cited criterion for analysing citation networks. The approach involves taking account of the natures of the papers that cite a given paper, so as to differentiate between papers that attract the same number of citations.

Design/methodology/approach – ArticleRank is an algorithm that has been derived from Google's PageRank algorithm to measure the influence of journal articles. ArticleRank is applied to two datasets – a citation network based on an early paper on webometrics, and a self-citation network based on the 19 most cited papers in the *Journal of Documentation* – using citation data taken from the Web of Knowledge database.

Findings – ArticleRank values provide a different ranking of a set of papers from that provided by the corresponding Times Cited values, and overcomes the inability of the latter to differentiate between papers with the same numbers of citations. The difference in rankings between Times Cited and ArticleRank is greatest for the most heavily cited articles in a dataset.

Originality/value – This is a novel application of the PageRank algorithm.

Keywords Bibliographies, Reference services

Paper type Research paper

Introduction

The citations in an article, book, or report indicate those items from the published literature that the author believes are of importance, in that they are related to, support, illustrate, or elaborate on what the author has to say (Bar-Ilan, 2008; Borgman and Furner, 2002; Garfield, 1979). Citations have long been thought to represent authoritativeness (Gilbert, 1977), intellectual influence (Zuckerman, 1987), and high quality (Cole and Cole, 1971), and it is hence normally assumed that the greater the number of citations that an item receives, then the greater the impact (or influence, importance, authoritativeness) of that item within its particular research community. However the use of the numbers of times that an item is cited (Times Cited) as a means of comparing different items makes the assumption that all citations contribute equally to the impact of a cited article; instead, it has been argued that not all citations are of equal importance (Sidiropoulos and Manolopoulos, 2006). In particular, Times Cited is

Aslib Proceedings: New Information Perspectives
Vol. 61 No. 6, 2009
pp. 605-618
© Emerald Group Publishing Limited
0001-253X
DOI 10.1108/00012530911005544

Downloaded by Yale University At 10:53 04 February 2015 (PT)

unable to differentiate between the impact of some papers, P_a, and that of another paper, P_b, when both of them are cited the same number of times, irrespective of the nature of the items citing P_a and P_b. For example, most of the citations to P_a may come from well-established, highly cited scientists, while most of the citations to P_b may come from beginners in the field; we refer to this phenomenon subsequently as the P_a-P_b problem.

When carrying out a search on the Web of Knowledge (WOK) database, it is possible to sort the results by Times Cited (as well as by other criteria such as source title or publication year). Studies of retrieval behaviour suggest that most searchers want the most important items matching their search criteria to appear on the first few pages (and ideally the very first page) of the search output; however, the P_a-P_b problem means that the use of Times Cited as a ranking criterion may not result in the most influential papers appearing first in a citation search. One way in which this problem might be addressed would be to take account not just of the number of citations but also of the journal impact factor (JIF) of each citing item; here, we suggest an alternative approach based on an application of the PageRank algorithm that is used to rank search outputs in the Google search engine (Brin and Page, 1998; Page *et al.*, 1998).

The PageRank algorithm takes account of the influence of web pages when carrying out a subject search on the web, where the influence of a source page is taken to be the number of other web pages linked to it. Thus, source pages with different levels of influence will, in general, make different contributions to the scores that are used to rank target pages, an idea first suggested in the context of citation analysis many years ago by Pinski and Narin (1976). The links between web pages are clearly analogous to the links between citing and cited items, and this has led to several previous applications of PageRank-like procedures to citation analysis: Bollen *et al.* (2006) discussed the use of their *Y*-factor to rank journals, using a weighted combination of the journal impact factor and the PageRank value; Fiala *et al.* (2008) used a modification of PageRank to rank authors using citation and collaboration networks; Jezek *et al.* (2008) used a modified PageRank score to analyse the degree of cooperation between citing and cited authors; Ma *et al.* (2008) used PageRank to study the influence of different countries' research in the fields of biochemistry and molecular biology; and Liu *et al.* (2005) discussed the use of their AuthorRank algorithm to analyse co-authorship networks (rather than citation networks) based on digital library conferences. Here we describe a modification of PageRank to determine the influence of academic journal articles, and to address the P_a-P_b problem.

The ArticleRank algorithm

Simple examples of a citation network and of a link network are shown in Figures 1 and 2, respectively. These networks are examples of *graphs* (Diestel, 2000; Thelwall, 2004; Wilson, 1996). In graph theory, a *directed graph* is made up of a collection of objects called *nodes* or *vertices* and a collection of connections between nodes, called *edges*, *arcs*, or *arrows*. Nodes and arrows can represent articles and the corresponding citation relationships (as in Figure 1), or web pages and the corresponding link relationships (as in Figure 2). Generally, an arrow from A to B indicates that A cites or links to B. The *indegree* of a node is the number of arrows that point to it, and its *outdegree* is the number of arrows that originate at the node.

Downloaded by Yale University At 10:53 04 February 2015 (PT)

Figures 1 and 2 exemplify directed graphs but there are at least three differences between them. First, the arrows in Figure 1 are unidirectional, but in Figure 2 they are bidirectional; this is because citation relationships exist within a time sequence that is absent from link relationships. Second, articles cannot cite themselves, whereas this is quite common for web pages (e.g. node 5 in Figure 2). Third, citation relationships will change only when new nodes and/or arrows are added to the network; link relationships, conversely, can vary substantially over quite short timescales as pages disappear, links become redirected, and the scope of web crawlers change. There is a further, non-structural difference between the two types of network; counts of citation links derived from a source such as the Web of Knowledge or Scopus are not inflated by the additional web links that can be obtained by sources of self-generated or mutually self-generating links such as those obtained using link farms.

The Google PageRank algorithm was designed to process link networks such as that shown in Figure 2 (Brin and Page, 1998; Page *et al.*, 1998). At the heart of the algorithm is the equation:

$$\text{PageRank}(A) = 1 - d + d \times \sum_{i=1}^{n} \frac{\text{PageRank}(P_i)}{O(P_i)},$$

where PageRank(A) denotes the PageRank value of Page(A); d is a damping factor between 0 and 1, usually 0.85; P_i ($1 <= i <= n$) is one of the n pages that link to Page(A); and $O(\mathrm{P_i})$ is the outdegree of Page(P_i) in the link network. Thus the PageRank value of Page(A) is recursively defined by the PageRank values of Page(A)'s indegree pages, with the value of PageRank(A) depending on the indegree of Page(A), the

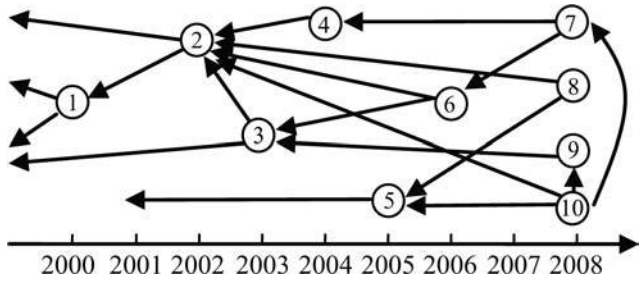

Figure 1.
Citation network

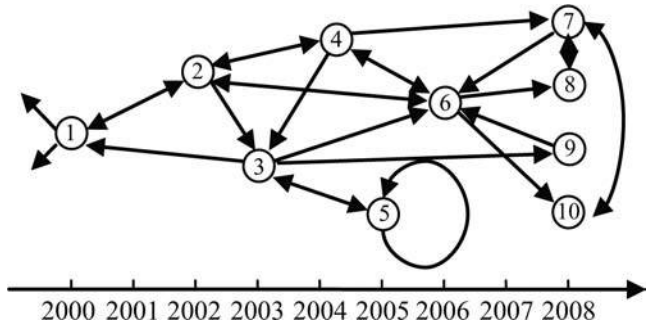

Figure 2.
Link network

Downloaded by Yale University At 10:53 04 February 2015 (PT)

PageRank values of Page(A)'s neighbouring nodes, and the outdegrees of these neighbouring nodes (taking account of the fact that a site's outdegree may be less than the actual number of outlinks since some of the latter may have become broken or be unavailable to a search engine's spider, as mentioned previously).

The obvious definition of ArticleRank for evaluating articles in citation networks is:

$$\text{ArticleRank}(A) = 1 - d + d \times \sum_{i=1}^{n} \frac{\text{ArticleRank}(Pi)}{NR(Pi)}, \quad (1)$$

where ArticleRank(A) denotes the ArticleRank score of a paper A; d is a damping factor as before (and set to 0.85 in all of our experiments, as with PageRank); P_i ($1 <= i <= n$) is one of the n papers that cite A; and $NR(P_i)$ is the number of references for P_i in the citation network. However, equation (1) has an inherent bias in that – other things being equal – a paper with very few references (i.e. a low value of $NR(P_i)$) will make a greater contribution to other papers' ArticleRank scores than will a paper with many references. This can yield counter-intuitive results, as we discovered in our initial experiments and as discussed further below. We hence considered modifications of equation (1) that would retain the basic PageRank methodology but that would not encode the bias that we have noted. Modifications to the denominator that were considered included $\sqrt{NR(Pi)}$, $\log_2 NR(P_i)$ and $\max\{NR(P_i)\} - \min\{NR(P_i)\}$. We finally settled on the form shown in equation (2) below, where $N\bar{R}$ is the mean value of NR when averaged over all of the papers in the network:

$$\text{ArticleRank}(A) = 1 - d + d \times N\bar{R} \times \sum_{i=1}^{n} \frac{\text{ArticleRank}(P_i)}{N\bar{R} + NR(P_i)}. \quad (2)$$

We thus replace the factor $1/NR(P_i)$ in equation (1) by the factor $(N\bar{R}/(N\bar{R} + NR(P_i))$ in equation (2). This form was chosen because it has the following characteristics: if $NR(P_i)$ is very small, then the factor approaches unity; if $NR(P_i)$ is typical of papers in the network, then the factor tends to one-half; and if $NR(P_i)$ is very large, then the factor approaches zero. To put this into perspective using the 343 papers in dataset-1 (as discussed in the next section), the value of $N\bar{R}$ is 35.6 (to three significant figures), and the minimum and maximum values of $NR(P_i)$ are 1 and 310: the factor $N\bar{R}/(N\bar{R} + NR(P_i))$ hence takes values between 0.973 and 0.103, respectively. It should be noted that the PageRank(A) values for a set of documents, and hence by analogy a set of ArticleRank(A) values, can be normalised so that they sum to unity; in this work, we have considered only the raw, unnormalised values.

Experimental details

The ArticleRank algorithm has been tested using two datasets derived from the ISI WOK database (see www.isiknowledge.com/) produced by Thomson Reuters and covering *c.* 8,700 of the world's leading journals in science, technology, the social sciences, arts and humanities.

The first dataset was a citation network based on the paper by Björneborn and Ingwersen (2001) entitled "Perspectives of webometrics", one of the earliest review articles in the field of webometrics. The citation network was created by taking all papers that had cited this starting paper, then taking all papers that had cited any of

Downloaded by Yale University At 10:53 04 February 2015 (PT)

Code	Title	Author	Year of publication	Number of references	Times cited
P001	Extracting macroscopic information from Web links	Thelwall, M.	2001	65	66
P002	*Perspectives of webometrics*	*Bjorneborn, L. and Ingwersen, P.*	*2001*	*58*	*61*
P003	Scholarly use of the web: what are the key inducers of links to journal web sites?	Vaughan, L. and Thelwall, M.	2003	45	43
P004	Conceptualising documentation on the web: an evaluation of different heuristic-based models for counting links between university web sites	Thelwall, M.	2002	47	42
P005	The history and meaning of the journal impact factor	Garfield, E.	2006	25	42
P006	Current concepts review – understanding the limitations of the journal impact factor	Kurmis, A.P.	2003	42	30
P007	Linguistic patterns of academic web use in Western Europe	Thelwall, M., Tang, R. and Price, L.	2003	48	22
P008	The relationship between the WIFs or inlinks of Computer Science Departments in UK and their RAE ratings or research productivities in 2001	Li, X.M., Thelwall, M., Musgrove, P. and Wilkinson, D.	2003	26	19
P009	A microscopic link analysis of academic institutions within a country – the case of Israel	Bar-Ilan, J.	2004	16	18
P010	Motivations for academic web site interlinking: evidence for the web as a novel source of information on informal scholarly communication	Wilkinson, D., Harries, G., Thelwall, M. and Price, L.	2003	49	16
...	... 331 other articles	...	...	...	...
P342	Why are websites co-linked? The case of Canadian universities	Vaughan, L., Kipp, M.E.I. and Gao, Y.J.	2007	19	0
P343	Women in paediatrics: recommendations for the future. Women chairs of the Association of Medical School Paediatric Department Chairs	Felice, M.E.	2007	28	0

Notes: The Times Cited values were collected from WOK, 15-31 December 2007. P002, the starting point for the network, is shown in italics

Table I.
The 343 papers in the citation network for dataset-1 (part only)

Downloaded by Yale University At 10:53 04 February 2015 (PT)

the papers that cited the starting paper, and so on till all papers had been included that could reach the starting paper by a citation path of whatever length. By definition, the terminal nodes in the resulting graph were papers that were uncited (as of 31 December 2007). The final network contained 343 papers (nodes) and 819 citation linkages. Some of the papers, ranked in order of decreasing Times Cited and coded P001-P343, are shown in Table I, with some of the corresponding citation links shown in Table II. Table II contains columns for just those 142 papers that had been cited at least once (i.e., the papers P143-P343 were uncited and thus formed the terminal nodes in the network). A "1" in the *XY*th cell of Table II means that the paper in row *X* cited the paper in column *Y*, while an empty cell means that paper *X* did not cite paper *Y*.

The second dataset was based on the 19 most cited papers from the *Journal of Documentation*, all of which had at least 100 citations in WOK on 31 October 2008. These papers are shown in Table III, coded by their positions when the papers are ranked in decreasing Times Cited order. The network here is based on journal self-citations, i.e. on citations that came from papers also published in the *Journal of Documentation*. The resulting self-citation network contained 354 papers, 195 of which had been cited at least once, and 752 citation linkages, these being encoded in a table analogous to that shown in Table II.

The calculation of a set of ArticleRank(*A*) values is iterative, with the calculations being repeated multiple times before the papers' ArticleRank(*A*) values stabilise. The calculations here were carried out in Microsoft Excel, as illustrated in Figure 3 for dataset-1. The first column of Figure 3 contains the iteration count and the first row the codes for the first 16 papers in the network. The initial values are 0.15 for all of the nodes (i.e. the factor $(1 - d)$ from equation (2) with d set to 0.85), and this continues to be the value for the 201 uncited nodes in the network. For the rest, the AR values change according to equation (2), converging (using a precision of 10^{-10}) in the fortieth iteration.

Results and discussion

The ArticleRank values for the 142 cited papers in dataset-1 were computed as described above, and the resulting values compared with the corresponding Times Cited values (from WOK) in Table IV, together with the corresponding rankings, where

Code		P001	P002	P003	P004	P005	P006	P007	P008	P009	…	P142
	NR	65	58	45	47	25	42	48	26	16		64
P001	65		1								…	
P002	58										…	
P003	45	1	1								…	
P004	47	1	1								…	
P005	25										…	
P006	42										…	
P007	48	1	1		1						…	
P008	26	1	1	1							…	
P009	16	1			1						…	
…		…	…	…	…	…	…	…	…	…	…	…
P343	28										…	

Table II.
The 819 citation relationships in the citation network for dataset-1 (part only)

Downloaded by Yale University At 10:53 04 February 2015 (PT)

Table III.
The 19 starting papers in the self-citation network for *Journal of Documentation*

Code	Title	Author	Year of publication	Number of references	Times Cited
P001	A behavioral approach to information retrieval system-design	Ellis, D.	1989	123	24
P002	ASK for information-retrieval. Part I. Background and theory	Belkin, N.J. Oddy, R.N. and Brooks, H.M.	1982	24	21
P003	Cognitive perspectives of information retrieval interaction: elements of a cognitive IR theory	Ingwersen, P.	1996	99	20
P004	Models in information behavior research	Wilson, T.D.	1999	40	19
P005	The derivation and application of the Bradford-Zipf distribution	Brookes, B.C.	1968	8	17
P006	Progress in documentation-empirical hyperbolic distributions (Bradford-Zipf-Mandelbrot) for bibliometric description and prediction	Fairthor, R.A.	1969	63	16
P008	Bradford distribution	Leimkuhl, F.F.	1967	8	13
P009	Progress in documentation – obsolescence and changes in the use of literature with time	Line, M.B. and Sandison, A.	1974	201	13
P011	On user studies and information needs	Wilson, T.D.	1981	28	12
P013	A theoretical basis for the use of co-occurrence data in information-retrieval	van Rijsbergen, C.J.	1977	26	11
P014	ASK for information retrieval. Part II. Results of a design study	Belkin, N.J., Oddy, R.N. and Brooks, H.M.	1982	13	11
P023	The calculation of web impact factors	Ingwersen, P.	1998	7	8
P024	Informetric analyses on the world wide web: methodological approaches to "webometrics"	Almind, T.C. and Ingwersen, P.	1997	22	8
P025	A statistical interpretation of term specificity and its application in retrieval	Sparck Jones, K.	1972	10	8
P035	The probability ranking principle in IR	Robertson, S.E.	1977	13	6
P036	Using probabilistic models of document retrieval without relevance information	Croft, W.B. and Harper, D.J.	1979	10	6
P037	Information retrieval through man-machine dialogue	Oddy, R.N.	1977	32	6
P052	Statistical bibliography or bibliometrics	Pritchard, A.	1969	8	4
P067	On the specification of term values in automatic indexing	Salton, G. and Yang, C.S.	1973	6	3

Note: The Times Cited values were collected from WOK, 21-31 October 2008

	P001	P002	P003	P004	P005	P006	P007	P008	P009	P010	P011	P012	P013	P014	P015	P016
	0.15000	0.15000	0.15000	0.15000	0.15000	0.15000	0.15000	0.15000	0.15000	0.15000	0.15000	0.15000	0.15000	0.15000	0.15000	0.15000
1	3.29180	2.94171	3.47238	2.13642	1.92708	2.39806	1.13809	1.05201	1.10061	0.75082	0.79714	0.80561	0.78035	0.87128	0.83763	0.85570
2	10.60976	10.19116	4.73589	5.72060	5.09082	2.49043	2.18327	1.66203	1.76557	1.18556	1.43786	2.28791	1.66967	1.48427	2.29752	3.21480
3	21.03202	20.44134	4.97305	8.75471	7.84102	2.49043	2.80894	1.81112	1.94850	1.52498	1.62191	3.38286	2.03432	1.79233	3.81464	4.03015
4	28.55033	29.40541	4.98937	9.88845	9.10335	2.49043	3.03845	1.85789	1.95486	1.66338	1.67216	3.70556	2.11389	1.86081	4.81551	4.22115
5	31.72902	34.09768	4.98937	10.18459	9.53543	2.49043	3.11467	1.87360	1.95486	1.70547	1.68672	3.76899	2.13793	1.87448	5.25193	4.23415
6	32.75714	35.89529	4.98937	10.25853	9.69265	2.49043	3.14343	1.88061	1.95486	1.71984	1.69373	3.78747	2.14671	1.87925	5.41358	4.23415
7	33.10226	36.50948	4.98937	10.28574	9.75798	2.49043	3.15635	1.88390	1.95486	1.72604	1.69702	3.79477	2.15092	1.88132	5.47868	4.23415
8	33.24033	36.74178	4.98937	10.29814	9.78723	2.49043	3.16247	1.88549	1.95486	1.72900	1.69861	3.79825	2.15291	1.88231	5.50741	4.23415
9	33.30206	36.84082	4.98937	10.30404	9.80063	2.49043	3.16539	1.88623	1.95486	1.73040	1.69935	3.79990	2.15386	1.88278	5.52038	4.23415
10	33.33059	36.88578	4.98937	10.30685	9.80683	2.49043	3.16677	1.88659	1.95486	1.73107	1.69971	3.80069	2.15431	1.88300	5.52635	4.23415
11	33.34385	36.90652	4.98937	10.30818	9.80970	2.49043	3.16743	1.88676	1.95486	1.73138	1.69988	3.80106	2.15452	1.88311	5.52910	4.23415
12	33.35005	36.91616	4.98937	10.30881	9.81105	2.49043	3.16774	1.88684	1.95486	1.73153	1.69996	3.80124	2.15462	1.88316	5.53037	4.23415
13	33.35295	36.92066	4.98937	10.30912	9.81167	2.49043	3.16789	1.88687	1.95486	1.73161	1.70000	3.80132	2.15467	1.88318	5.53097	4.23415
14	33.35432	36.92276	4.98937	10.30926	9.81197	2.49043	3.16796	1.88689	1.95486	1.73164	1.70001	3.80136	2.15469	1.88319	5.53125	4.23415
15	33.35496	36.92375	4.98937	10.30933	9.81211	2.49043	3.16799	1.88690	1.95486	1.73166	1.70002	3.80138	2.15470	1.88320	5.53138	4.23415
16	33.35527	36.92422	4.98937	10.30936	9.81217	2.49043	3.16801	1.88691	1.95486	1.73166	1.70003	3.80139	2.15471	1.88320	5.53144	4.23415
17	33.35541	36.92444	4.98937	10.30938	9.81221	2.49043	3.16802	1.88691	1.95486	1.73167	1.70003	3.80139	2.15471	1.88320	5.53147	4.23415
18	33.35548	36.92454	4.98937	10.30938	9.81222	2.49043	3.16802	1.88691	1.95486	1.73167	1.70003	3.80139	2.15471	1.88320	5.53148	4.23415
19	33.35551	36.92459	4.98937	10.30939	9.81223	2.49043	3.16802	1.88691	1.95486	1.73167	1.70003	3.80140	2.15471	1.88320	5.53149	4.23415
20	33.35553	36.92462	4.98937	10.30939	9.81223	2.49043	3.16802	1.88691	1.95486	1.73167	1.70003	3.80140	2.15471	1.88320	5.53149	4.23415
21	33.35554	36.92463	4.98937	10.30939	9.81223	2.49043	3.16802	1.88691	1.95486	1.73167	1.70003	3.80140	2.15471	1.88320	5.53149	4.23415
22	33.35554	36.92463	4.98937	10.30939	9.81223	2.49043	3.16803	1.88691	1.95486	1.73167	1.70003	3.80140	2.15471	1.88320	5.53149	4.23415
23	33.35554	36.92464	4.98937	10.30939	9.81223	2.49043	3.16803	1.88691	1.95486	1.73167	1.70003	3.80140	2.15471	1.88320	5.53149	4.23415
24	33.35554	36.92464	4.98937	10.30939	9.81223	2.49043	3.16803	1.88691	1.95486	1.73167	1.70003	3.80140	2.15471	1.88320	5.53149	4.23415
25	33.35554	36.92464	4.98937	10.30939	9.81223	2.49043	3.16803	1.88691	1.95486	1.73167	1.70003	3.80140	2.15471	1.88320	5.53149	4.23415

Figure 3.
The iterative computation of ArticleRank values for dataset-1 (part only)

R_{TC} and R_{AR} are the papers' ranks when they are ranked in descending order of Times Cited and ArticleRank values, respectively.

We illustrate the effect of the ArticleRank calculations by reference to papers P004 and P005. P004 has Times Cited and ArticleRank values of 34 and 10.30939 ($R_{TC} = 4.5$ and $R_{AR} = 3$), while the corresponding values for P005 are 34 and 9.81223 ($R_{TC} = 4.5$ and $R_{AR} = 4$). Eleven papers cited both P004 and P005 and hence make the same contribution to their ArticleRank values; the contributions of the remaining citing documents are shown in Table V, where the right-hand column in each part of the table contains values for the contribution, C, to the overall ArticleRank value, where $C = AR/(NR + N\bar{R})$. It will be seen that factors causing P004 to be ranked above P005 include: P004 is cited by six uncited papers (i.e. papers with the default ArticleRank value of 0.15) where as P005 is cited by seven uncited papers; one of the papers citing P004 is P016, which makes a very high contribution of 0.05049.

The R_{AR} rankings shown in Table IV are very different from those obtained using equation (1), i.e. the standard PageRank algorithm when applied to citation data. For example, P123 has the disparate rankings $R_{TC} = 123.5$ and $R_{AR} = 31.5$: it is cited by just a single paper P279 that has not been cited and that has $NR(279) = 3$. With the initial PageRank(A) values set to 0.15, this paper hence makes a contribution of $0.85 \times 0.15/3$, i.e., 0.042505, to P123, which is sufficient to give the latter a reasonably high ranking: thus the observed discrepancy between the two sets of rankings arises not from P123 being cited by important papers but from it being cited by a paper that has very few references. This behaviour is typical of some of the other outlier papers that were observed when equation (1) was used: for example, the twice-cited P089 with $R_{TC} = 91$ and $R_{AR} = 6$, and the thrice-cited P075 with $R_{TC} = 69.5$ and $R_{AR} = 7$. Reference to Table IV will show that the R_{TC} and R_{AR} ranks are much less discordant for these three articles when equation (2) is used; P075 with $R_{TC} = 69.5$ and $R_{AR} = 62$; P089 with $R_{TC} = 91$ and $R_{AR} = 72$; and P123 with $R_{TC} = 123.5$ and $R_{AR} = 103.5$. Similar behaviour was observed with many of the papers in dataset-2 when equation (1) was used.

The extent of the statistical correlation between the sets of R_{TC} and R_{AR} values was investigated using the Kendall τ coefficient (Kendall and Gibbons, 1990; Siegel and

Downloaded by Yale University At 10:53 04 February 2015 (PT)

Code	*TC*	R_{TC}	*AR*	R_{AR}	Code	*TC*	R_{TC}	*AR*	R_{AR}	Code	*TC*	R_{TC}	*AR*	R_{AR}
P001	53	1	33.35554	2	P049	5	50	0.55184	58	P097	2	91	0.26563	105
P002	47	2	36.92464	1	P050	5	50	0.55597	57	P098	2	91	0.34080	86
P003	39	3	4.98937	7	P051	5	50	0.58421	56	P099	2	91	0.29163	98
P004	34	4.5	10.30939	3	P052	5	50	0.73040	50	P100	2	91	0.31781	90.5
P005	34	4.5	9.81223	4	P053	5	50	0.53752	59	P101	2	91	0.27218	102
P006	27	6	2.49043	16	P054	4	57.5	0.42037	71	P102	2	91	0.31781	90.5
P007	18	7	3.16803	11	P055	4	57.5	0.53112	60	P103	2	91	0.23730	118
P008	15	8	1.88691	21	P056	4	57.5	0.38659	75	P104	2	91	0.25426	113
P009	14	9	1.95486	20	P057	4	57.5	0.39014	73	P105	1	123.5	0.21919	128
P010	13	12.5	1.73167	24	P058	4	57.5	0.52432	61	P106	1	123.5	0.45670	65
P011	13	12.5	1.70003	26	P059	4	57.5	0.79993	44	P107	1	123.5	0.26462	106.5
P012	13	12.5	3.80140	9	P060	4	57.5	2.94009	13	P108	1	123.5	0.23706	119
P013	13	12.5	2.15471	19	P061	4	57.5	1.68602	27	P109	1	123.5	0.25911	110
P014	13	12.5	1.88320	22	P062	3	69.5	0.42850	68	P110	1	123.5	0.24954	116
P015	13	12.5	5.53149	6	P063	3	69.5	0.45347	67	P111	1	123.5	0.20241	134.5
P016	12	17	4.23415	8	P064	3	69.5	1.22665	32	P112	1	123.5	0.22137	126
P017	12	17	2.76597	14	P065	3	69.5	0.32010	89	P113	1	123.5	0.18117	141
P018	12	17	2.39696	17	P066	3	69.5	0.39008	74	P114	1	123.5	0.21522	130
P019	11	19	2.76156	15	P067	3	69.5	0.37365	79	P115	1	123.5	0.31103	93
P020	10	20.5	8.28088	5	P068	3	69.5	0.30498	94	P116	1	123.5	0.25177	114
P021	10	20.5	2.20307	18	P069	3	69.5	0.67915	52	P117	1	123.5	0.26462	106.5
P022	9	24	1.12020	36	P070	3	69.5	0.42637	70	P118	1	123.5	0.20241	134.5
P023	9	24	0.93380	38	P071	3	69.5	0.29153	99	P119	1	123.5	0.22251	125
P024	9	24	1.29026	31	P072	3	69.5	0.45386	66	P120	1	123.5	0.82624	42
P025	9	24	1.14588	34	P073	3	69.5	0.33068	87	P121	1	123.5	0.23164	122
P026	9	24	1.52910	29	P074	3	69.5	0.66765	54	P122	1	123.5	0.19512	137
P027	8	31	2.95710	12	P075	3	69.5	0.51089	62	P123	1	123.5	0.26759	103.5
P028	8	31	1.70052	25	P076	3	69.5	0.42817	69	P124	1	123.5	0.20522	133
P029	8	31	1.14544	35	P077	3	69.5	0.35729	82	P125	1	123.5	0.26759	103.5
P030	8	31	1.00754	37	P078	2	91	0.27627	101	P126	1	123.5	0.25655	112
P031	8	31	3.47569	10	P079	2	91	0.23686	120	P127	1	123.5	0.20182	136
P032	8	31	0.73394	48	P080	2	91	0.28445	100	P128	1	123.5	0.19468	138
P033	8	31	0.87282	40	P081	2	91	0.36200	81	P129	1	123.5	0.30351	95
P034	8	31	0.67254	53	P082	2	91	0.24623	117	P130	1	123.5	0.22616	124
P035	8	31	1.37709	30	P083	2	91	0.34284	84	P131	1	123.5	0.21617	129
P036	7	37	0.90228	39	P084	2	91	0.34229	85	P132	1	123.5	0.17936	142
P037	7	37	0.82136	43	P085	2	91	0.25016	115	P133	1	123.5	0.25911	110
P038	7	37	0.73390	49	P086	2	91	0.36608	80	P134	1	123.5	0.25911	110
P039	6	42.5	0.71420	51	P087	2	91	0.29619	97	P135	1	123.5	0.31500	92
P040	6	42.5	1.19783	33	P088	2	91	0.38057	77	P136	1	123.5	0.20563	131.5
P041	6	42.5	1.62248	28	P089	2	91	0.40817	72	P137	1	123.5	0.23629	121
P042	6	42.5	0.84672	41	P090	2	91	0.30236	96	P138	1	123.5	0.19218	140
P043	6	42.5	0.79213	45	P091	2	91	0.33004	88	P139	1	123.5	0.22026	127
P044	6	42.5	0.62780	55	P092	2	91	0.37373	78	P140	1	123.5	0.20563	131.5
P045	6	42.5	0.74867	46	P093	2	91	0.26453	108	P141	1	123.5	0.19381	139
P046	6	42.5	1.86090	23	P094	2	91	0.38102	76	P142	1	123.5	0.22880	123
P047	5	50	0.73791	47	P095	2	91	0.35321	83					
P048	5	50	0.49924	63	P096	2	91	0.46224	64					

Notes: *AR*, Article Rank; *TC*, Times Cited; AR values are to five decimal places

Table IV.
AR and *TC* values and corresponding rankings for the 142 cited papers in dataset-1

Downloaded by Yale University At 10:53 04 February 2015 (PT)

Downloaded by Yale University At 10:53 04 February 2015 (PT)

Code	Papers citing P004 *AR*	*NR*	$C \times 10^3$	Code	Papers citing P005 *AR*	*NR*	$C \times 10^3$
P009	1.95486	35	27.68929	P007	3.16803	48	37.89504
P010	1.73167	26	28.11154	P008	1.88691	16	36.56802
P016	4.23415	49	50.04908	P017	2.76597	49	32.69472
P022	1.12020	40	14.81746	P024	1.29026	191	5.69398
P037	0.82136	19	15.04320	P032	0.73394	99	5.45276
P038	0.73390	50	8.57360	P051	0.58421	49	6.90557
P039	0.71420	26	11.59417	P053	0.53752	33	7.83561
P048	0.49924	33	7.27755	P064	1.22665	43	15.60618
P082	0.24623	70	2.33169	P075	0.51089	55	5.63898
P093	0.26453	31	3.97197	P080	0.28445	72	2.64360
P095	0.35321	21	6.24053	P094	0.38102	36	5.32154
P100	0.31781	41	4.14902	P097	0.26563	67	2.58901
P104	0.25426	62	2.60516	P101	0.27218	59	2.87716
P105	0.21919	23	3.74048	P115	0.31103	65	3.09175
P111	0.20241	45	2.51133	P122	0.19512	33	2.84430
P118	0.20241	63	2.05287	P135	0.31500	41	4.11228
P139	0.22026	4	5.56220	P170	0.15	26	2.43506
P160	0.15	51	1.73210	P173	0.15	60	1.56904
P184	0.15	64	1.50602	P182	0.15	58	1.60256
P319	0.15	41	1.95822	P198	0.15	66	1.47638
P328	0.15	48	1.79426	P252	0.15	65	1.49105
P336	0.15	68	1.44788	P291	0.15	12	3.15126
P337	0.15	23	2.55973	P341	0.15	72	1.39405
Total			207.31937	Total			190.88989

Table V.
AR values and *NR* of the papers citing P004 or P005 in dataset-1

Notes: *AR*, ArticleRank; *NR*, number of references; $C = AR/(NR + N\overline{R})$

Castellan, 1988), a non-parametric coefficient that measures the degree of correlation between two rankings of the same set of objects (i.e. the set of 142 cited papers in the present context). The computed value for τ was 0.517, which corresponds to a highly significant statistical correlation ($p < 0.001$) between the sets of R_{TC} and R_{AR} values. The correlation is shown in the scatter diagram of Figure 4, where points above the diagonal indicate papers with higher R_{TC} and/or lower R_{AR} values, and where points below the diagonal imply the converse. Some of the outlier papers have been circled, and these demonstrate clearly the effect of being cited by prestigious papers. For example, P106 and P120 are in the shadowed circle, and both of them were cited only once by an important paper: P106 ($R_{TC} = 123.5$, $R_{AR} = 65$) was cited by P046, whose *AR* value is 1.86090 ($R_{AR} = 23$); and P120 ($R_{TC} = 123.5$, $R_{AR} = 42$) was cited by P040, whose *AR* value is 1.19783 ($R_{AR} = 33$). The vertical sets of points in this figure provide a graphical representation of what we have referred to previously as the P_a-P_b problem and indicate the extent of this problem.

The procedures described previously for dataset-1 were then applied to dataset-2, the *Journal of Documentation* self-citation network. To save space, we have not included a table analogous to that shown in Table IV listing all of the 195 sets of data; however, the scatter plot in Figure 5 shows similar behaviour to that observed previously, with a value for τ between the sets of R_{TC} and R_{AR} values of 0.519 ($p < 0.001$).

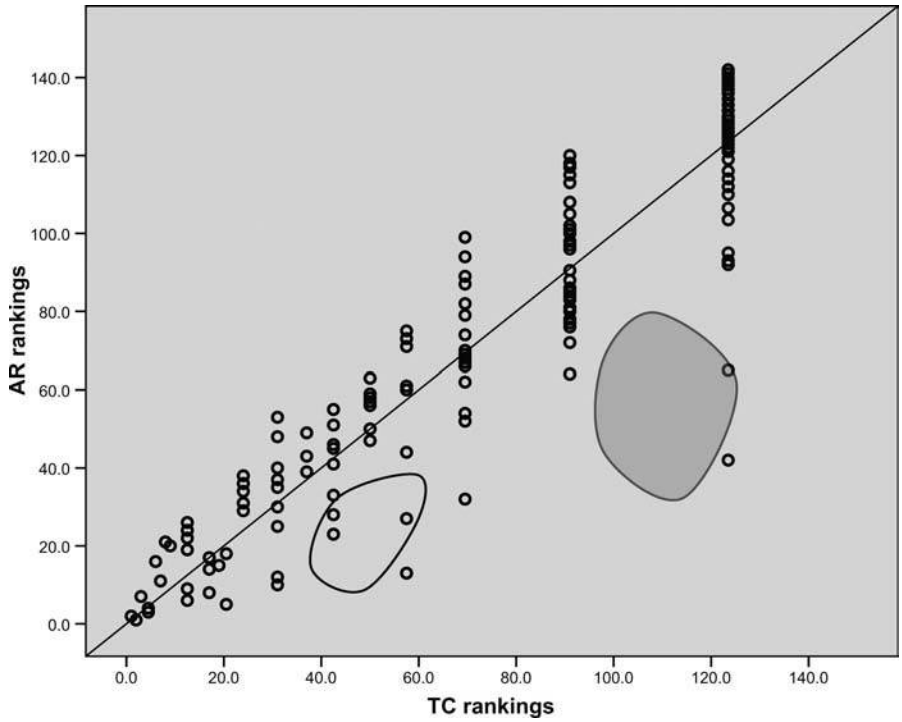

Figure 4.
Scatter plots of R_{AR} and R_{TC} for the 142 cited papers in dataset-1

The differences in ranking here are often more marked than in the case of dataset-1. For example, papers P013, P014 and P015 were all cited 13 times, with $R_{TC} = 14$, but have very different R_{AR} values of 1, 7 and 77, respectively. The resolution of this P_a-P_b problem is explained by the data in Table VI, which is analogous to that discussed previously in Table V. Specifically, both P014 and P015 were cited by a P001with an *AR* value of 10.59889 ($R_{AR} = 3$), with P014 also being cited by three other papers (P003, P019 and P007) whose *AR* values were in excess of 3; conversely, the highest *AR* value amongst the papers citing P013 was as low as 0.90971 (paper P038 with $R_{AR} = 60$).

The correlation values for the scatter plots in Figures 4 and 5 show statistically significant correlations between the sets of R_{TC} and R_{AR} values (although we have already noted that there are some outlier papers whose rank positions change considerably when the *AR* values are computed). An inspection of the ranked lists suggests that the less-cited papers tend to remain near to the bottom of the ranking. For example, P078-P0142 in dataset-1 comprise the 65 papers with TC = 1 or 2; only seven of these appear outside the bottom 65 positions when the papers are ranked using R_{AR}, with positions 42, 64, 65, 72, 76, 77 and 78. For the corresponding 108 papers with TC = 1 or 2 in dataset-2, only 23 appear outside the bottom 108 positions when ranked using R_{AR}. Conversely, the more highly cited papers tend to show a greater degree of movement when ranked using *AR*, as we demonstrate using the top 19 papers from Table IV, i.e. those with TC > 10. These were ranked using the *TC* and *AR* values in Table IV so that each had a rank in the range 1-19. The resulting scatter

Downloaded by Yale University At 10:53 04 February 2015 (PT)

Downloaded by Yale University At 10:53 04 February 2015 (PT)

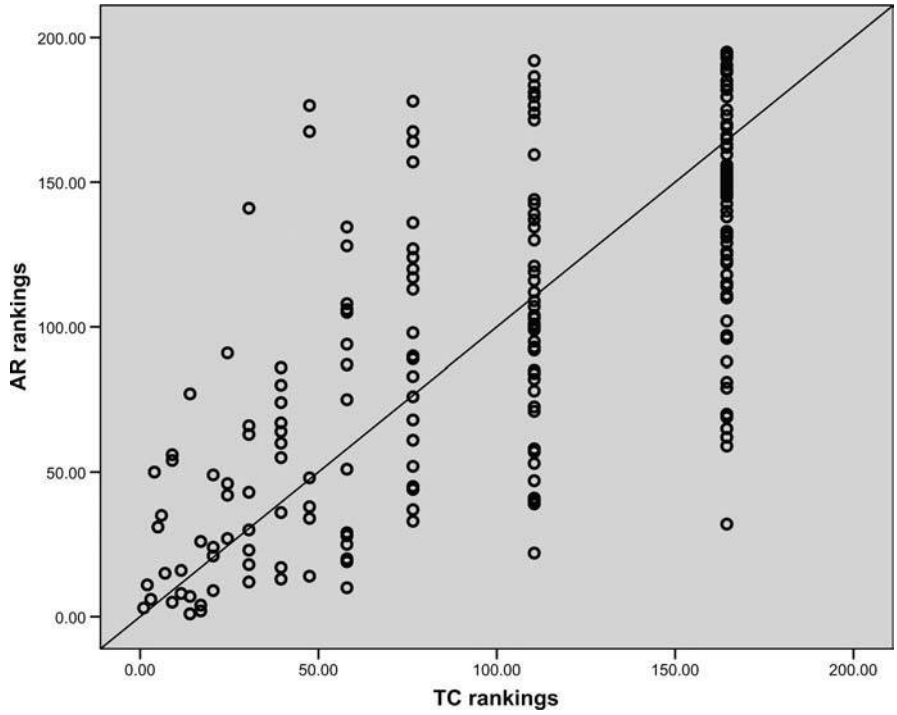

Figure 5.
Scatter plot of R_{AR} and R_{TC} for the 195 cited papers in dataset-2

Papers citing P013				Papers citing P014				Papers citing P015			
Code	*AR*	*NR*	$C \times 10^3$	Code	*AR*	*NR*	$C \times 10^3$	Code	*AR*	*NR*	$C \times 10^3$
P038	0.90971	19	16.97221	P001	10.59889	123	67.25184	P001	10.59889	123	67.25184
P036	0.87759	10	19.67691	P003	7.71829	99	57.77163	P092	2.32164	13	48.77388
P072	0.70384	8	16.52213	P019	5.36514	45	67.40119	P053	1.87894	116	12.47634
P146	0.38008	11	8.33520	P007	3.34034	54	37.70137	P049	1.56975	91	12.49799
P168	0.33185	13	6.97166	P017	1.98742	39	27.00305	P149	0.53106	26	8.76334
P177	0.29459	92	2.32694	P117	0.39629	45	4.97852	P068	0.32678	61	3.41821
P078	0.28429	23	4.93555	P172	0.31192	197	1.34679	P135	0.26241	21	4.71965
P144	0.19816	21	3.56404	P138	0.24268	29	3.81570	P145	0.20036	23	3.47847
P046	0.15000	53	1.71233	P148	0.20036	60	2.11797	P123	0.15000	51	1.75234
P063	0.15000	73	1.39405	P133	0.15000	184	0.68618	P344	0.15000	1	4.21348
P219	0.15000	16	2.96443	P341	0.15000	52	1.73210	P345	0.15000	5	3.78788
Total			85.37545	T167			271.80634	Total			171.13341

Table VI.
AR values and *NR* for the papers citing P013-P015 in dataset-2

Notes: *AR*, ArticleRank; *NR*, number of references; $C = AR/(NR + \overline{NR})$

plot is shown in Figure 6a, with the Kendall test showing that there is not a statistically significant correlation ($p > 0.05$) between the two sets of rankings. The scatter plot in Figure 6b shows the analogous analysis for the 19 highly cited papers in Table III that provide the basis for dataset-2, and there is again no statistically significant correlation between the two sets of rankings. Overall, it would hence appear that ArticleRank

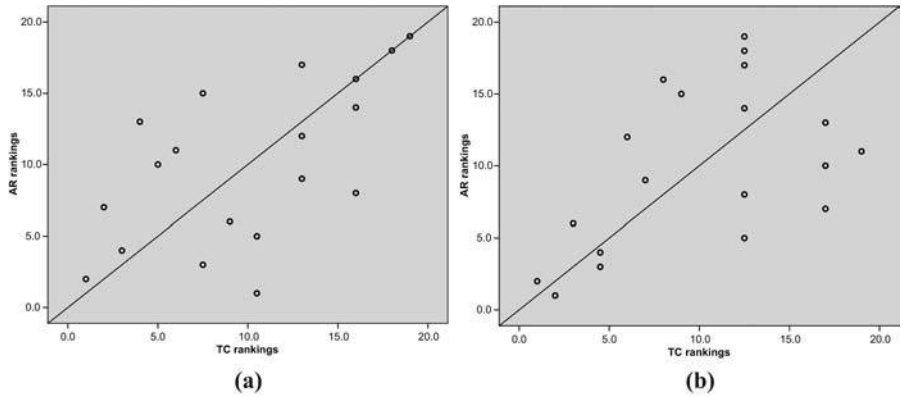

Figure 6.
Scatter plots of R_{AR} and R_{TC}

Note: The rankings of **(a)** the top 19 papers in Table IV and **(b)** the 19 papers in Table III

affects the ranking of the highly-cited papers more than it does the less-cited papers: this is intuitively reasonable since – other things being equal – papers that have been cited only a very few times are unlikely to have been cited by prestigious papers.

Conclusions

In this paper, we have described a modification of the PageRank algorithm that can be used as an alternative to the number of citations for the analysis of citation data. The algorithm, called ArticleRank, provides a simple way of discriminating between papers with equal numbers of citations, boosting the position of papers that are cited by papers that have considerable impact in their own right. ArticleRank hence provides an interesting alternative to Times Cited as a way of analysing a citation network. It does, however, require substantial computation if the algorithm is to run for many iterations on large numbers of papers: that said, large-scale processing is clearly possible.

Two areas for future work suggest themselves. First, the present paper has presented the method and discussed the quantitative characteristics of the rankings that result from its use. There should now be a qualitative study in which users are asked to comment on the relative merits of the two types of ranking. Second, both Times Cited and ArticleRank values will increase (or will at least not decrease) the longer that a paper has been published; however, the latter value for a paper can continue to grow even though the paper itself is no longer being cited, this growth reflecting changes in the prestige of its citing papers. It would hence be of interest to study the changes that occur in ArticleRank values over time and to compare these with the comparable obsolescence data for citations.

References

Bar-Ilan, J. (2008), "Informetrics at the start of the 21st century – a review", *Journal of Informetrics*, Vol. 2 No. 1, pp. 1-52.

Björneborn, L. and Ingwersen, P. (2001), "Perspectives of webometrics", *Scientometrics*, Vol. 50 No. 1, pp. 78-9.

Downloaded by Yale University At 10:53 04 February 2015 (PT)

Bollen, J., Rodriguez, M. and van de Sompel, H. (2006), "Journal status", *Scientometrics*, Vol. 69 No. 3, pp. 669-87.

Borgman, C.L. and Furner, J. (2002), "Scholarly communication and bibliometrics", *Annual Review of Information Science and Technology*, Vol. 36, pp. 3-72.

Brin, S. and Page, L. (1998), "The anatomy of a large-scale hypertextual web search engine", *Computer Networks and ISDN Systems*, Vol. 30 Nos 1-7, pp. 107-17.

Cole, J. and Cole, S. (1971), "Measuring the quality of sociological research: problems in the use of the Science Citation Index", *The American Sociologist*, Vol. 6 No. 1, pp. 23-9.

Diestel, R. (2000), *Graph Theory*, Springer-Verlag, New York, NY.

Fiala, D., Rousselot, F. and Ježek, K. (2008), "PageRank for bibliographic networks", *Scientometrics*, Vol. 76 No. 1, pp. 135-58.

Garfield, E. (1979), *Citation Indexing – Its Theory and Application in Science, Technology, and Humanities*, ISI Press, Philadelphia, PA.

Gilbert, G.N. (1977), "Referencing as persuasion", *Social Studies of Science*, Vol. 7 No. 1, pp. 113-22.

Jezek, K., Fiala, D. and Steinberger, J. (2008), "Exploration and evaluation of citation networks", available at: tp://jps.library.utoronto.ca/ocs-2.0.0-1/index.php/Elpub/2008/paper/view/670 (accessed 7 December 2008).

Kendall, M. and Gibbons, J.D. (1990), *Rank Correlation Methods*, 5th ed., Edward Arnold, London.

Liu, X., Bollen, J., Nelson, M.L. and van de Sompel, H. (2005), "Co-authorship networks in the digital library research community", *Information Processing and Management*, Vol. 41 No. 6, pp. 1462-80.

Ma, N., Guan, J. and Zhao, Y. (2008), "Bringing PageRank to the citation analysis", *Information Processing and Management*, Vol. 44 No. 2, pp. 1-11.

Page, L., Brin, S., Motwani, R. and Winograd, T. (1998), "The PageRank citation ranking: bringing order to the web", available at http://dbpubs.stanford.edu:8090/pub/1999-66 (accessed 1 November 2008).

Pinski, G. and Narin, F. (1976), "Citation influence for journal aggregates of scientific publications: theory, with application to the literature of physics", *Information Processing and Management*, Vol. 12 No. 5, pp. 297-312.

Sidiropoulos, A. and Manolopoulos, Y. (2006), "Generalized comparison of graph-based ranking algorithms for publication and authors", *Journal of Systems and Software*, Vol. 79 No. 12, pp. 1679-700.

Siegel, S. and Castellan, N.J. (1988), *Nonparametric Statistics for the Behavioural Sciences*, 2nd ed., McGraw-Hill, New York, NY.

Thelwall, M. (2004), *Link Analysis: An Information Science Approach*, Academic Press, San Diego, CA.

Wilson, R. (1996), *Introduction to Graph Theory*, 4th ed., Longman, Harlow.

Zuckerman, H. (1987), "Citation analysis and the complex problem of intellectual influence", *Scientometrics*, Vol. 12 Nos 5/6, pp. 329-38.

Corresponding author
Peter Willett can be contacted at: p.willett@sheffield.ac.uk

To purchase reprints of this article please e-mail: **reprints@emeraldinsight.com**
Or visit our web site for further details: **www.emeraldinsight.com/reprints**

Downloaded by Yale University At 10:53 04 February 2015 (PT)

This article has been cited by:

1. Onur Küçüktunç, Erik Saule, Kamer Kaya, Ümit V. Çatalyürek. 2014. Diversifying Citation Recommendations. *ACM Transactions on Intelligent Systems and Technology* **5**:10.1145/2699158, 1-21. [CrossRef]
2. Pierluigi Amodio, Luigi Brugnano. 2014. Recent advances in bibliometric indexes and the PaperRank problem. *Journal of Computational and Applied Mathematics* **267**, 182-194. [CrossRef]
3. Star X. Zhao, Fred Y. Ye. 2012. Exploring the directed h-degree in directed weighted networks. *Journal of Informetrics* **6**:4, 619-630. [CrossRef]

Downloaded by Yale University At 10:53 04 February 2015 (PT)