

The journal download immediacy index (DII): experiences using a Chinese full-text database

Jin-kun Wan · Ping-huan Hua · Ronald Rousseau · Xiu-kun Sun

Received: 11 May 2009 / Published online: 10 February 2010
© Akadémiai Kiadó, Budapest, Hungary 2010

Abstract Relationships between the journal download immediacy index (DII) and some citation indicators are studied. The Chinese full-text database CNKI is used for data collection. Results suggest that the DII can be considered as an independent indicator, but that it also has predictive value for other indicators, such as a journal's *h*-index. In case a journal cannot yet have an impact factor—because its citation history within the database is too short—the DII can be used for a preliminary evaluation. The article provides results related to the CNKI database as a whole and additionally, some detailed information about agricultural and forestry journals.

Keywords Journal evaluation indicators · Download immediacy index · Journal immediacy index · Full-text databases · Agricultural and forestry journals

Background

Nowadays libraries and information providers offer electronic full-text access to most journal articles. The articles themselves are made available by publishers or information

J. Wan (✉) · P. Hua
Library of Tsinghua University, 100084 Beijing, China
e-mail: wanjk@lib.tsinghua.edu.cn

P. Hua
e-mail: gfhs@cnki.net

J. Wan · P. Hua · X. Sun
China Scientometrics and Bibliometrics Research Center, Tsinghua University,
PO Box 84-48, 100084 Beijing, China

R. Rousseau
KHBO (Association K.U.Leuven), Industrial Sciences and Technology, 8400 Oostende, Belgium
e-mail: ronald.rousseau@khbo.be

R. Rousseau
Department of Mathematics, K.U.Leuven, Celestijnenlaan 200B, 3001 Louvain (Heverlee), Belgium

intermediaries. Moreover, many articles (in preprint or reprint form) are freely available on the Internet. Librarians, intermediaries and publishers are all interested in the use that is made of the products they provide (Fowler 2007). As simple information can be read on screen the number of accesses made to a database, a journal or a particular article is an indication of its usefulness. Yet when a reader decides to download an article this indicates more than just casual interest. For this reason the number of downloads of an article (and on an aggregate level, of a journal or from a database) is an essential indicator of its scientific importance. Already in 2002, Rousseau (2002) observed that for articles published in e-journals three visibility indicators could be collected: the number of visits to the article's page, the number of downloads and the number of inlinks. As an example he noted that the e-journal *Conservation Ecology* was already counting accesses since (at least) 1999.

Even though users download articles for other purposes than for citing them in their own publications, downloads may be a first step towards citations, and as such may be predictors of later citations. The relation between online "hit counts" and subsequent citations has received a lot of attention in recent years. Perneger (2004) wrote that "the number of early hits is a potentially useful measure of the scientific value of published medical research papers". He further observed that in his study for every 100 additional hits, 4.4 additional citations accrued over 5 years (Perneger 2004). In the context of Open Access, Harnad and Brody (2004) found a significant correlation between the frequency of downloads and citations 2 years later. Brody et al. (2006) analyse how short-term web usage predicts medium-term citation impact and find a significant correlation of approximately 0.4 between the citation and download impact of articles in physics and mathematics. Yet, Bollen and Van de Sompel found a negative correlation between journal impact factors and journal usage impact for users in the California State University system (Bollen and Van de Sompel 2008). Results varied, however, over different disciplines and were highly affected by demographic factors (relative numbers of undergraduates, graduate students and researchers).

In an extensive study Nicholas et al. (2003) analysed user statistics, including downloads of full articles in order to study the impact of so-called *Big Deal* agreements on user online searching behaviour. Recall that the term *Big Deal* agreement, coined by Kenneth Frazier (2001), refers to an agreement between a publisher and a library (or group of libraries) where access is provided to a large collection of online journals for an established price, often at the cost of existing print subscriptions plus an increment. Under the contract annual price increases are capped for a number of years.

Research by Wan et al. (2005) indicates that curves representing time versus number of downloads are different from time-citation curves. Peak values of download curves are visibly higher than those of citation curves and appear earlier. Recently Schloegl and Gorraiz (2008) studied the usage impact factor (UIF) and usage half-life (UHL) for a number of oncology journals based on full-text access in Science Direct. Among other aspects they investigated differences in obsolescence characteristics between citations and downloads, using absolute as well as relative counts.

Research topics

In this paper, the following problems are studied:

- (1) Can the general shape of download curves and citation curves as observed by Wan et al. (2005) be confirmed?

- (2) What are the relations between the download immediacy index (DII) and other indicators?
- (3) Can the DII be used as an independent bibliometric indicator?

Before describing our methodology and results we introduce our main data source: the China Academic Journals Full-text Database (CAJ), a part of the Chinese National Knowledge Infrastructure (CNKI).

The Chinese National Knowledge Infrastructure

The CNKI is an electronic platform launched in June 1999 by the Tsinghua Tongfang Knowledge Network (TTKN) Technology Company, a spin-off company of Tsinghua University, located in Beijing. CNKI contains several databases as shown in Table 1.

Of the academic journals in the database the large majority is in Chinese but 400 are in English, confirming the fact that there are some English language journals in China, but that they generally play a small role within the country's S&T system (Ren and Rousseau 2004). CNKI has more than 5,500 customers (universities, public and corporate libraries, hospitals, etc.). Most are in China but a growing number is outside China. For instance the British Library, Cambridge University and the Staatsbibliothek zu Berlin are subscribers.

CNKI moreover publishes the *Chinese Academic Journals Comprehensive Citation Report* on the basis of the *Chinese Academic Journals Comprehensive Evaluation Database* in which about 6,000 Chinese academic journals are involved. In this report 12 bibliometric indicators are provided, including the total number of citations, the journal impact factor, the number of published articles, the journal immediacy index, the *h*-index, and the DII (under the name of web immediacy index) (Wan and Xue 2007; Wan et al. 2007).

Methodology

The journal DII is defined as the number of downloads of a journal's articles within one publication year, divided by the number of published articles by that journal in that same year (Wan et al. 2007). This definition is obviously modelled after Thomson/Reuters' journal immediacy index. It can be considered as a measure of how intensely the "average article" in an online journal is downloaded. As such it is an indicator of the attractivity of a journal's new articles (probably largely based on the reputation of the journal). We could

Table 1 Contents of the CNKI databases

Type	Database	Number of titles	Coverage	Number of records
Journal	China Academic Journals Full-text Database (CAJ)	7,500	1994 onward	15,908,460
Journal	Century Journals Project (CJP)	3,964	1887 onward	2,965,927
Dissertation	China Doctor/Master Dissertations Full-text Database (CDMD)	400	1999 onward	322,458
Newspaper	China Core Newspapers Database (CCND)	1,000	2000 onward	5,845,314
Proceedings	China Proceedings of Conference Full-text Database (CPCD)	1,018	1999 onward	464,127
Yearbook	China Yearbooks Full-text Database (CYFD)	1,000	1912 onward	2,934,233

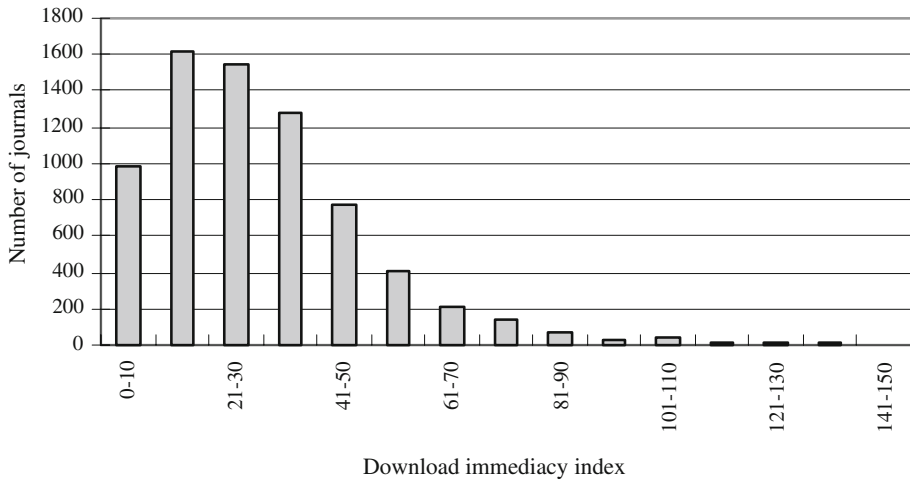


Fig. 1 The distribution of the download immediacy index (DII) of Chinese academic journals

have studied other indicators suitable for the analysis of journals available in electronic form, such as the web h -index (Hirsch 2005; Hua et al. 2009). In this article, however, we focus on the DII because it fully embodies quick response time, one of the main advantages of electronic resources available on the Internet.

The distribution of the DII of 6,000 Chinese academic journals published in 2006 is shown in Fig. 1. The highest value is 396.5, but more than 200 journals have a download immediacy value equal to zero. Although statistically (according to a chi-square test) no distribution fits the data the gamma distribution with shape parameter 2.38 and scale parameter 0.0754 fits best (after removing journals with DII = 0), see Fig. 2.

Based on the download data of the central website of the CNKI, the DII of Chinese academic journals has been calculated. It was found that among the 2.4 million articles included in the CAJ and published in 2006 1,778,174 (or 75%) have been downloaded at least once. These articles were downloaded more than 48 million times in total, or an average of 27 times per downloaded article (during the year of publication, i.e. the year 2006).

Values for other indicators used in this article such as the immediacy index, impact factor, total cites and h -index were obtained from the *Chinese Academic Journals Comprehensive Citation Report*, a yearly publication of scientometric data about the CNKI.

General results

Based on the statistical data of 2006, the citation distribution curve and the download distribution curve are drawn and shown in Fig. 3. The numbers on the y-axis are relative numbers of downloads and citations, expressed as percentages. The peak value of the download distribution curve is visibly higher than that of the citation distribution curve and appears earlier. Concretely, 25.6% of all downloads in the year 2006 are articles published in 2005 (and 22.9% are articles published the same year); 18.7% of all citations in 2006 are given to articles published in 2004. Downloading can be characterized as quick-response behaviour. This is less the case for citing. These observations confirm those made by Wan

Fig. 2 Best-fitting gamma distribution (frequencies on vertical axis)

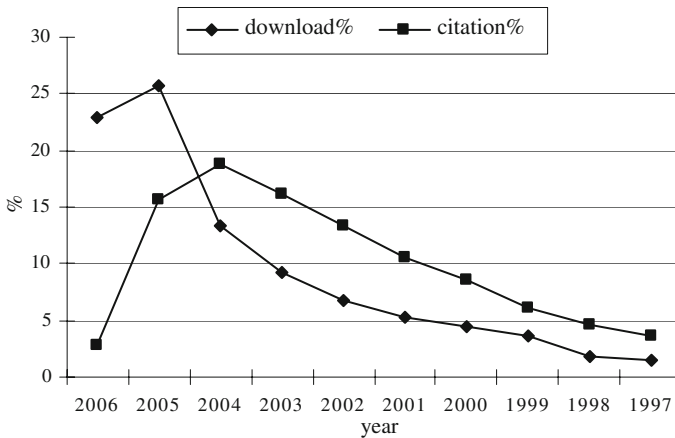
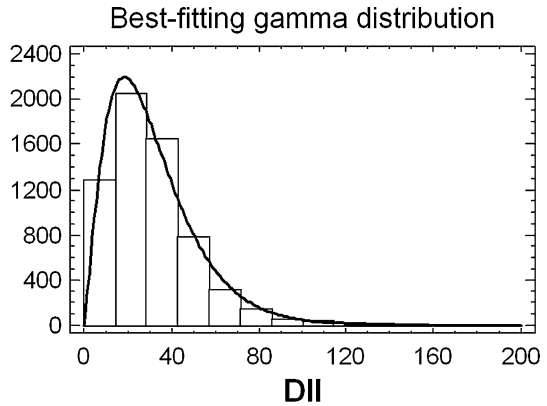


Fig. 3 Comparison between the download frequency distribution and the citation distribution (2006 data)

et al. (2005). We note that for the oncology journals studied by Schloegl and Gorraiz (2008) most of the download curves have a mode during the first year. Probably the databases used (Science Direct vs. CNKI) and fields studied (oncology vs. all fields) explain this difference.

Next we calculated the correlation coefficient between the download curve and different shifted versions of the citation curve. These values are known as normalized cross-covariances (NCCV) with different lag times. They are measures of similarity between two signals (here: two distribution curves). Concretely: the normalized cross-covariance with time lag 0 is the Pearson correlation between the download curve and the citation curve. The normalized cross-covariance with time lag 1 is the Pearson correlation coefficient of the download curve (with last point removed) and the citation curve with the first point removed; the normalized cross-covariance with time lag 2 is the Pearson correlation coefficient of the download curve (with the last two points removed) and the citation curve with the first two point removed, etc. Such NCCV have been used in informetrics before to study the age distribution of Chinese scientists over different years (Jin et al. 2003). Results are shown in Table 2. The NCCV with lag time equal to three is the largest (in bold in

Table 2 Normalized cross-covariances

Lag	Value
0	0.305
1	0.900
2	0.943
3	0.946
4	0.928

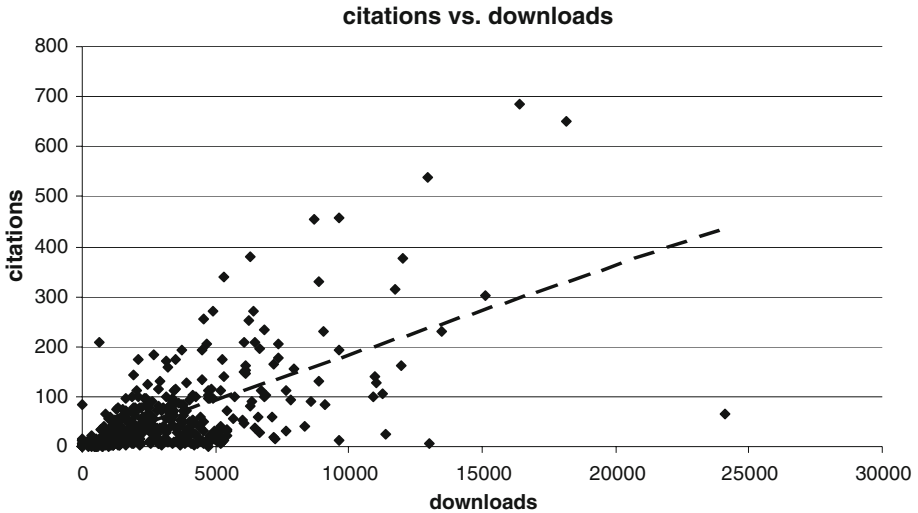


Fig. 4 The relationship between number of downloads in 2005 and number of citations received in 2006 for the papers published in agricultural and forestry journals in 2004

Table 2), but a shift of 2 years gives almost the same result. This result points to a strong relation between downloading and citing, be it shifted in time.

Results related to the field of agriculture and forestry

In this section the relationship between the number of downloads and the number of citations for the papers published in agricultural and forestry journals (460 in total) in the year 2004 are discussed. According to the characteristics of the curves shown in Fig. 3 the number of downloads in year 2005 and the number of citations in 2006 are used, because they may be the peak values. Indeed, we found a positive Pearson correlation, equal to 0.635 between citations (in 2006) and downloads (in 2005), see (Fig. 4). The Spearman rank correlation is 0.59. Although statistically acceptable (standard *t*-test, $H_0: \rho = 0$; *p*-value $\ll 0.001$) Fig. 4 clearly shows that this relation is not linear at all: some journals have a much higher number of downloads than predicted by the trend line, while other ones have a much higher number of citations than predicted. This observation again illustrates the difference between download frequency and citation frequency.

Figure 5 clearly shows that the relationship between the DII and total cites is not very good. Generally speaking, the higher the total cites, the higher the download immediacy (the Pearson correlation coefficient is 0.72). But some technical journals have a high DII

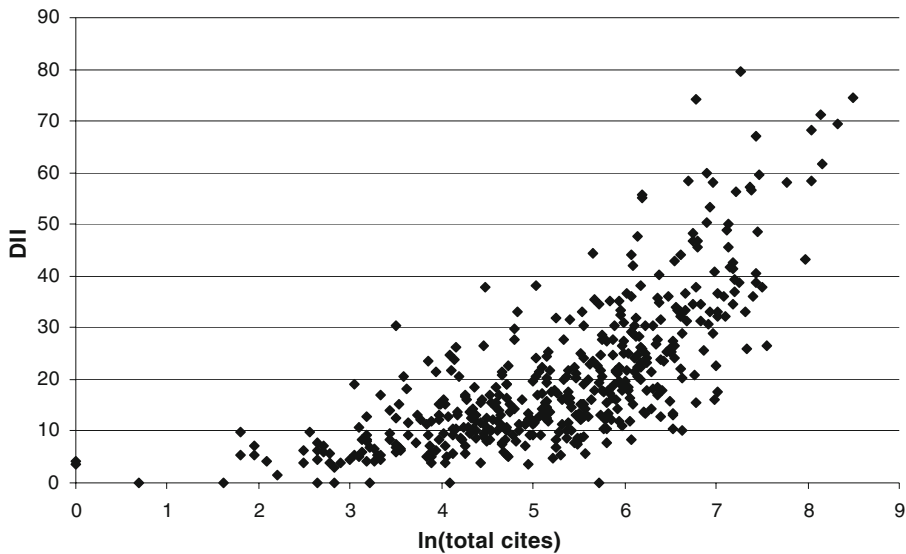


Fig. 5 The relationship between total cites and download immediacy index for agricultural and forestry journals (semi-logarithmic scales)

and low total cites. Hence, the DII is perhaps best seen as an independent indicator to evaluate the value of a journal to its readers. As will be discussed further, this value can be scientific, but also practical (=of immediate use).

Table 5 (see appendix) provides a comparison between the DII rank and other indicators' ranks for the top 50 journals (according to DII) among the 459 agricultural and forestry journals. These other indicators are: the (citation) immediacy index, the classical impact factor (calculated within the CNKI database), the total number of citation and the h -index. The data in Table 5 indicate that the ranks of the DII are often not consistent with the ranks of other indicators. The DII ranks of, for example, the journals *Sugarcane* and *Sichuan Forestry Exploration and Design* are 47 and 49, but their ranks for the other indicators are quite high.

We have recalculated the ranks among these 50 journals (so all ranks are now between 1 and 50) and calculated Spearman rank correlation coefficients between the new ranked lists. These are shown in Table 3, corresponding p -values are shown in Table 4. The smallest correlation occurs between the DII and the traditional journal immediacy index. This is the only rank correlation that is statistically not significant (in bold in Table 4). The highest correlation occurs between the journal h -index and the total number of citations. This confirms observations made in the Web of Science (WoS) by van Raan (2006). When considering the first row of Table 3, this is the row that contains the correlation coefficient between the DII and the other indicators, it is clear that the highest correlation occurs with the h -index. So the DII is not only a predictor of the impact factor and the total number of citations (be it an imperfect one), but even more of the h -index. We recall that Yue et al. (2004) found that the classical immediacy index and impact factor are highly correlated.

As mentioned above, although there are clear relations between the DII and other indicators, many exceptions occur. In these cases, predicted impact factors or total citation results are not reliable. Yet from a librarian's or a scientometrician's point of view these are the interesting cases, requiring insight in the concrete circumstances. In Fig. 4, for

Table 3 Spearman correlation

	A	B	C	D	E
A	1	0.24	0.41	0.42	0.58
B		1	0.59	0.51	0.60
C			1	0.66	0.75
D				1	0.83
E					1

A Download immediacy index, *B* the immediacy (citation) index, *C* the Garfield-Sher impact factor, *D* the total number of citations received, *E* the *h*-index

Table 4 Corresponding *p*-values

	A	B	C	D	E
A		0.0964	0.0034	0.0022	<0.0001
B			<0.0001	0.0002	<0.0001
C				<0.0001	<0.0001
D					<0.0001
E					

A Download immediacy index, *B* the immediacy (citation) index, *C* the Garfield-Sher impact factor, *D* the total number of citations received, *E* the *h*-index

instance, we notice journals with a high number of downloads, but with no or almost no citations. Upon investigation we found that these journals (e.g. Gansu Agriculture) are not really scientific journals but more of the “practical-news” type. They publish short articles containing popular technical information or new facts that can increase farmers’ profits. Hence farmers like to download their contents because they offer a lot of good practical advice. There are also a few journals in the database with no downloads, but with a small number of citations. These are generally journals of low quality.

Download immediacy index versus citation immediacy index

Now we discuss the difference between the citation immediacy index (CII) and the DII for journals. Obviously, see Fig. 6, the DII is several orders of magnitude larger than the CII. Also here, for agricultural and forestry journals, the correlation coefficient is not high ($R = 0.57$).

Going beyond the field of agriculture and forestry we note that statistical results indicate that the DII is dependent on academic fields (see Fig. 7), e.g. the DII of some popular (such as financial and economic) journals is distinctly higher than that of some special (such as silkworm and beekeeping) journals, because they have more readers. For comparison the values of the journal immediacy index are also shown in Fig. 7 (but remember that the DII is several orders of magnitudes larger). Among these selected fields the immediacy index of psychology journals is the highest. The average DII of this selected group of journals is 103.53, its average immediacy index is 0.126. The Pearson correlation coefficient between the two is 0.8.

Conclusion

Returning to the list of problems we set out to study we first mention that the general shape of download curves as observed earlier by Wan et al. (2005) is confirmed in this new study. The relationship between the DII and other indicators suggests that the DII can be considered as an independent indicator, but that it also has some predictive value for other

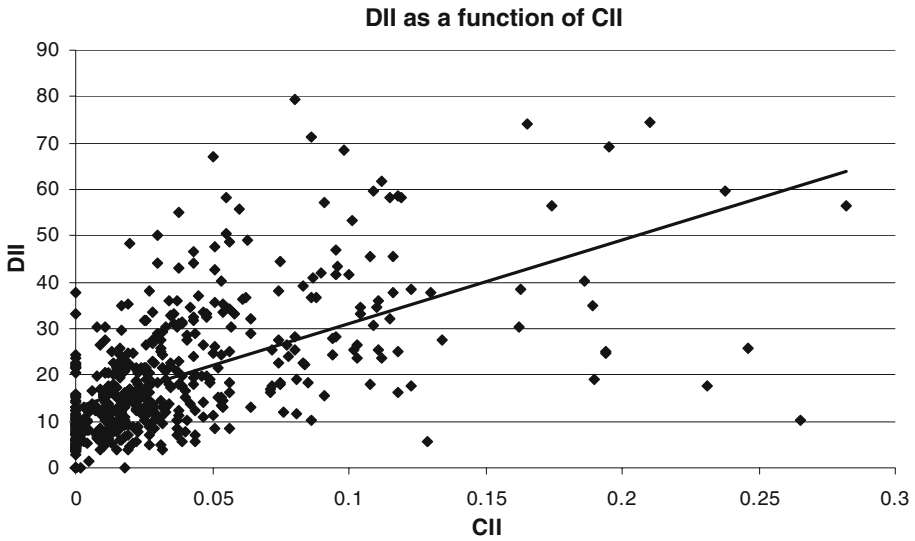


Fig. 6 The relationship between the citation immediacy index (CII) and download immediacy index (DII) for all agricultural and forestry journals (year 2006)

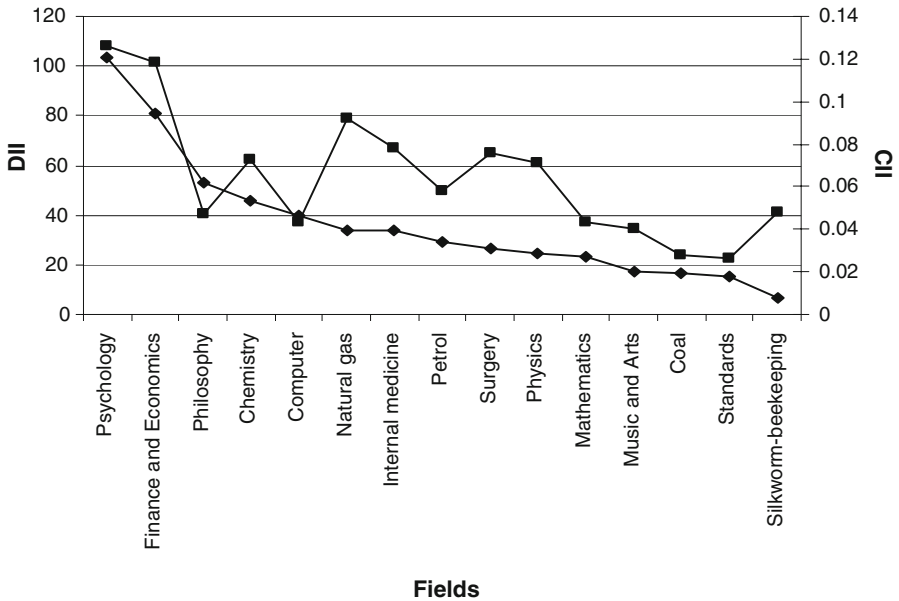


Fig. 7 Download immediacy index (DII), lower line, and citation immediacy index (CII), upper line of some selected fields

indicators, such as the *h*-index. It is especially useful for those journals that cannot yet have an impact factor as their citation history within the database is too short (1 or 2 years). In such cases the DII can be used for a preliminary evaluation. A question not studied in depth

here is whether download frequencies predict citation frequencies. In order to answer this question one should study the same cohort of articles over several years and compare their download data and their citation data. This is not done here. Using NCCV we did, however, compare a download curve and a citation curve in the same year and found out that a shift of two year yielded already a correspondence of 0.94, indicating a strong relation between downloading and citing.

In conclusion we would like to mention that on the same retrieval platform, as in our study, all journals, in spite of their academic fields, are facing the same selection of readers. The download frequency of each journal reflects its overall impact, and the way the journal embodies social and academic values.

Acknowledgements Rousseau's work was supported by the National Natural Science Foundation of China (NSFC Grant No 70673019). The authors thank Raf Guns (Antwerp University), Christian Schögl (University of Graz), Juan Gorraiz (University of Vienna) and two anonymous reviewers for helpful comments on an earlier version.

Appendix

In this appendix we present Table 5 in which the download immediacy index is compared to the CII, the classical impact factor, the total number of citations received and the *h*-index. All indices refer to the year 2006. This means that publications in the years 2004 and 2005 are used to determine the impact factor; publications in the year 2006 are used for the two immediacy indices. For 'total number of citations received' citations are received in 2006 and refer to publications over all years. Finally, the *h*-index is the 2006 *h*-index using publications during the period 2002–2006.

Table 5 Comparison between download immediacy index rank and other indicators' ranks for top 50 agricultural and forestry journals (using all journals to rank)

A	B	C	D	E	Journal
1	67	20	21	22	Acta Phytopathologica Sinica
2	6	7	1	1	Scientia Agricultura Sinica
3	14	19	54	10	Research of Agricultural Modernization
4	58	23	4	7	Acta Horticulturae Sinica
5	7	10	2	4	Acta Agronomica Sinica
6	46	2	5	5	Acta Pedologica Sinica
7	116	43	14	18	Chinese Journal of Soil Science
8	30	15	3	2	Transactions of the Chinese Society of Agricultural Engineering
9	35	32	48	15	World Forestry Research
10	4	6	11	11	Plant Nutrition and Fertilizer Science
11	23	9	6	3	Journal of Soil and Water Conservation
12	98	27	61	46	Chemistry & Industry of Forest Products
13	22	18	8	9	Agro-Environmental Protection
14	28	58	42	40	Chinese Journal of Eco-agriculture
15	53	8	18	14	Soils
16	1	5	17	13	Chinese Journal of Rice Science
17	13	42	23	17	Journal of Fruit Science

Table 5 continued

A	B	C	D	E	Journal
18	89	46	112	50	System Sciences and Comprehensive Studies In Agriculture
19	155	56	111	54	Acta Zoonutrimenta Sinica
20	44	86	44	43	Chinese Journal of Preventive Veterinary Medicine
21	99	48	47	25	Journal of Agricultural Biotechnology
22	196	84	31	35	Bulletin of Soil and Water Conservation
23	86	14	32	38	Journal of Fishery Sciences of China
24	92	120	12	28	Chinese Agricultural Science Bulletin
25	265	41	59	88	Soybean Science
26	111	69	120	33	Review of China Agricultural Science and Technology
27	48	30	60	23	Cotton Science
28	135	50	52	51	Soils and Fertilizers
29	26	77	53	27	World Agriculture
30	37	4	30	8	Acta Prataculturae Sinica
31	73	146	194	183	Forest Engineering
32	197	128	70	37	Progress in Veterinary Medicine
33	136	96	133	115	China Biogas
34	47	13	7	6	Scientia Silvae Sinicae
35	156	194	75	104	Journal of Agricultural Mechanization Research
36	112	76	26	19	Agricultural Research in the Arid Areas
37	55	49	129	63	Chinese Journal of Pesticide Science
38	49	33	29	24	Acta Veterinaria et Zootechnica Sinica
39	45	95	28	57	Chinese Journal of Veterinary Science and Technology
40	57	31	40	48	Chinese Journal of Oil Crop Sciences
41	105	3	13	20	Journal of Maize Sciences
42	12	29	92	47	Journal of Tea Science
43	63	62	25	32	Chinese Journal of Veterinary Science
44	20	36	22	21	Pesticides
45	15	28	15	31	Forest Research
46	76	64	116	90	Jiangsu Journal of Agricultural Sciences
47	213	215	273	192	Sugarcane
48	18	21	56	30	Chinese Journal of Biological Control
49	401	222	334	218	Sichuan Forestry Exploration and Design
50	27	16	10	29	Journal of Fisheries of China

Ranks according to values for: *A* download immediacy index, *B* the citation immediacy index, *C* the traditional Garfield-Sher impact factor (within CNKI), *D* the total number of citations received, *E* the *h*-index

References

- Bollen, J., & Van de Sompel, H. (2008). Usage impact factor: The effects of sample characteristics on usage-based impact metrics. *Journal of the American Association for Information Science and Technology*, 59(1), 136–149.
- Brody, T., Harnad, S., & Carr, L. (2006). Earlier web usage statistics as predictors of later citation impact. *Journal of the American Association for Information Science and Technology*, 57(8), 1060–1072.

- Fowler, D. C. (Ed.). (2007). *Usage statistics of E-serials*. Binghamton: Haworth Information Press.
- Frazier, K. (2001). The librarian's dilemma: Contemplating the costs of the "Big Deal". *D-Lib Magazine*, Vol. 7(3). <http://www.dlib.org/dlib/march01/frazier/03frazier.html>.
- Harnad, S., & Brody, T. (2004). Prior evidence that downloads predict citations. *BMJ Rapid Responses*. Retrieved September 6, 2004 from <http://bmj.bmjournals.com/cgi/eletters/329/7465/546#73000>.
- Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences*, 102(46), 16569–16572.
- Hua, P. H., Rousseau, R., Sun, X. K., & Wan, J. K. (2009). A download $h^{(2)}$ -index as a meaningful usage indicator of academic journals. In B. Larsen & J. Leta (Eds.), *Proceedings of ISSI 2009* (pp. 587–596). Rio de Janeiro: BIREME/PAHO/WHO and Federal University of Rio de Janeiro.
- Jin, B. H., Li, L., & Rousseau, R. (2003). Production and productivity of Chinese scientists as a function of their age: the period 1995–1999. In G. Jiang, R. Rousseau, & Y. Wu (Eds.), *Proceedings of the 9th international conference on scientometrics and informetrics* (pp. 112–120). Dalian: Dalian University of Technology Press.
- Nicholas, D., Huntington, P., & Watkinson, A. (2003). Digital journals, Big Deals and online searching behaviour: A pilot study. *Aslib Proceedings*, 55(1/2), 84–109.
- Perneger, T. V. (2004). Relation between online "hit counts" and subsequent citations: prospective study of research papers in the *BMJ*. *BMJ*, 329(7465), 546–547.
- Ren, S. L., & Rousseau, R. (2004). The role of China's English-language scientific journals in scientific communication. *Learned Publishing*, 17, 99–104.
- Rousseau, R. (2002). Journal evaluation: Technical and practical issues. *Library Trends*, 50(3), 418–439.
- Schloegl, C., & Gorraiz, J. (2008). Comparison of citation and usage indicators: The case of oncology journals. In J. Gorraiz & E. Schiebel (Eds.), *Excellence and emergence, book of abstracts, 10th international conference on science and technology indicators* (pp. 98–102). Vienna: Austrian Research Centers GmbH-ARC.
- van Raan, A. F. J. (2006). Comparison of the Hirsch-index with standard bibliometric indicators and with peer judgment for 147 chemistry research groups. *Scientometrics*, 67(3), 491–502.
- Wan, J. K., Hua, P. H., Du, J., & Song, H. M. (2007). Focus on the front of scientific evaluation, putting new indicators into practice: Three new bibliometric indicators adopted by China Academic Journals Comprehensive Citation Report. *Digital Library Forum* (in Chinese), issue 3, pp. 36–41.
- Wan, J. K., Hua, P. H., & Sun, X. K. (2005). Bibliometrics analysis on cited frequency and downloaded frequency of journal papers. *New Technology of Library and Information Service* (in Chinese), 4, 58–62.
- Wan, J. K., & Xue, F. Y. (2007). *Chinese Academic Journals Comprehensive Citation Report*. Beijing: Science Press.
- Yue, W. P., Wilson, C. S., & Rousseau, R. (2004). The immediacy index and the journal impact factor: Two highly correlated derived measures/L'indice d'immédiateté et le facteur d'impact des périodiques: deux mesures dérivées fortement corrélées. *Canadian Journal of Information and Library Science/La Revue canadienne des sciences de l'information et de bibliothéconomie*, 28(1), 33–48.