Journal Download Factor: A Comprehensive Indicator of Dissemination, Impact, Knowledge, and Information Volume–A New Attempt to Improve the Timeliness of Bibliometric Indicators (Postprint)

Authors: YU Liping

Date: 2024-04-03T00:00:00+00:00

Abstract

Purpose/Significance: There is a lack of comprehensive indicators that holistically characterize the dissemination, impact, knowledge, and information volume of academic journals. This paper proposes a download factor indicator to address this gap. Method/Process: First, based on annual variations in download and citation frequencies, and using CNKI citation data for CSSCI journals in library and information science, a panel data model was employed to establish a prediction model for download and citation frequencies, thereby determining the optimal lag period for designing the download factor. The download factor indicator is proposed as the average number of downloads per paper two years after publication divided by 100. Ridge regression was further used to analyze the relationship between the download factor and the impact factor, h-index, and article volume. Results/Conclusion: Download frequencies with lags of one and two years determine 80% of citation frequency; the download factor can effectively measure a journal' s knowledge/information volume, dissemination level, impact, and academic quality; the download factor indicator requires validation across more disciplines and datasets.

Full Text

Preamble

Journal Download Factor: A Comprehensive Indicator of Dissemination, Impact, Knowledge, and Information Volume–A New Attempt to Improve the Timeliness of Bibliometric Indicators

YU Liping[1],[2]
(1. School of Statistics and Mathematics, Zhejiang Gongshang University, Hangzhou 310018;
2. Collaborative Innovation Center of Statistical Data Engineering, Technology & Application, Zhejiang Gongshang University, Hangzhou 310018)

Abstract:
This study addresses the lack of comprehensive indicators characterizing the dissemination, impact, knowledge, and information volume of academic journals by proposing a novel download factor indicator. Based on annual variations in download and citation frequencies, we utilize citation data from CSSCI journals in library and information science from CNKI to establish a predictive model using panel data models. The optimal lag period for designing the download factor is determined, and the indicator is defined as the average number of downloads per paper divided by 100 two years after journal publication . Ridge regression further analyzes the relationship between the download factor and the impact factor, h-index, and publication volume. Results show that download frequencies with lags of 1 and 2 years determine 80% of citation frequency. The download factor effectively measures a journal' s knowledge information volume, dissemination level, influence, and academic quality, though further validation across more disciplines and datasets is required.

Keywords: download factor; level of dissemination; lag period; evaluation indicator
Classification Number: G302
Document Code: A
Article ID: 1002-1248

1 Introduction

Although GARFIELD [1] proposed as early as 1996 that download frequency could replace citation frequency to solve the lag problem in citation analysis evaluation, progress in this area has been very slow. The birth of the Internet has brought revolutionary impact to bibliometrics, giving rise to various online download indicators for academic literature. The most representative basic indicator among them is download frequency, which also includes the Web annual download rate, total download volume, download half-life, and Google Scholar Index. The proposal of these indicators provides new methods and means for measuring academic dissemination and impact, greatly developing traditional bibliometrics and forming an important component of alternative metrics (Altmetrics). However, research on download frequency-related indicators remains insufficient, with many issues requiring further investigation: What is the relationship between download frequency and citation frequency? What distribution patterns does download frequency exhibit? What time span of download frequency is appropriate for constructing evaluation indicators? How can the

information content of new indicators be measured? What are the statistical characteristics of new indicators? How should they be applied in evaluation?

Previous studies have often used graphical methods and correlation analysis to examine the relationship between download frequency and citation frequency, which is insufficient to draw effective conclusions. Most research agrees that download frequency is correlated with citation frequency, but the complexity of this relationship requires in-depth analysis. Although download frequency has better timeliness than citation frequency, there is a lack of discussion on what time span of download frequency should be used for evaluation. Too long a time span reduces timeliness, while too short a span leads to incomplete data and ineffective evaluation. BOTTING et al. [10] found that downloads within the publication year can predict citation counts three years later, with a goodness-of-fit of 0.450. SCHLOGL et al. [11] discovered that in library and information science, the correlation between download counts and citation counts is relatively high. HU [12] found that at the journal level, total download volume and total citation volume are highly linearly correlated, with the correlation coefficient reaching 0.770. ZHAO et al. [13] showed that in computer science and technology, whether at the journal or document level, citation counts and download counts have a strong positive correlation.

Regarding the characteristics and applications of download indicators, HARY [5] argued that download indicators have the same ability as citation indicators to identify major scientific developments. SU [6] pointed out that download behavior originates from keyword searches, and journals with more downloads have more standardized keywords and topics closer to current scholarly concerns. DING et al. [7] found that downloads in recent years account for a high proportion, with columns and reviews having high download rates while research papers have high citation rates. DANIEL [8] proposed that document downloads are influenced by reader interest, document visibility, and maturity. XIE and GONG [9] believed that paper quality, download time windows, and download data sources all affect the relationship between downloads and citations.

In summary, existing research has adequately addressed the correlation between download frequency and citation frequency, but several aspects require further in-depth study: How do download frequencies from different lag periods affect citation frequency? Which years' download frequencies primarily influence a given year' s citation frequency? How does this influence affect the construction of new evaluation indicators? This paper addresses these questions through in-depth econometric analysis, constructing a new evaluation indicator based on download frequency that comprehensively reflects dissemination, impact, knowledge, and information volume.

---

2 Research Design

2.1 Research Data

This study uses CSSCI journals in library and information science as the research object. Based on CNKI' s citation database, we analyze the relationship between download frequency and citation frequency. Considering that there is a certain time lag between paper download frequency and citation frequency, the publication volume data are selected from 2015, while download counts and citation counts span from 2015 to 2021. Some years have missing data, which are excluded from the analysis. There are 20 journals in total.

2.2 Annual Changes in Download and Citation Frequencies

For the entire discipline of library and information science, papers published in 2015 show distinct patterns: download frequency peaks in the first year after publication and then slowly declines, while citation frequency peaks in the second year after publication, with a slight decrease in the third year [Figure 1: see original paper].

2.3 Lag Relationship Between Citation Frequency and Download Frequency

Many studies suggest that citation frequency lags behind download frequency by 1-3 years, but this is mostly empirical estimation. This paper conducts a rigorous analysis based on panel data models. A key reason for using panel data is that while download frequency is an important factor influencing citation frequency, citation frequency has too many influencing factors. If important variables are omitted, ordinary regression estimates will be biased. However, the fixed effects model in panel data uses differencing estimation, which can better estimate the relationship between download frequency and citation frequency even when important variables are omitted.

The basic model is as follows, considering lag periods of 1-5 years:

$$Y = c + X(-1)\beta_1 + X(-2)\beta_2 + X(-3)\beta_3 + X(-4)\beta_4 + X(-5)\beta_5 + \mu_{it}$$

where $X(-1)$, $X(-2)$, $X(-3)$, $X(-4)$, and $X(-5)$ represent the 1-year, 2-year, 3-year, 4-year, and 5-year lag terms of download frequency, respectively, and β_1 to β_5 are their regression coefficients.

The panel model estimation results are shown in Table 1 . First, the 1-5 year lag model is estimated. The Hausman test yields a chi-square value of 27.523 with a p-value of 0.000, rejecting the null hypothesis of random effects. Therefore, the fixed effects model is adopted. In the estimation results, the 4-year and 5-year lag coefficients are negative, which clearly does not match reality, and

the 3-year lag coefficient fails the statistical test. Thus, using 1-5 year lags is inappropriate.

Continuing with 1-4 year lags, the fixed effects model is ultimately adopted. All download frequency terms with different lag periods pass statistical tests and are positive. The model's goodness-of-fit R^2 is high, far exceeding the prediction accuracy of other scholars, indicating that the model is very appropriate. The regression results show that the 2-year lag download frequency has the largest elasticity coefficient (0.787), followed by the 1-year lag (0.629), then the 3-year lag (0.190), and finally the 4-year lag (0.101). If only considering this factor, using a 1-year time span would be best. However, from the panel data regression results, the 2-year lag download frequency has the greatest impact on citation frequency. Considering both factors comprehensively, a 2-year lag period is adopted to construct the download factor indicator.

---

3 Construction and Analysis of the Download Factor

3.2 Construction of the Download Factor

Based on the design principle of the impact factor indicator, the download factor (DF) is constructed. The download factor is defined as the average number of downloads per paper divided by 100, two years after journal publication. The formula is expressed as:

$$DF = \frac{D_t + D_{t-1} + D_{t-2}}{100 P_{t-2}}$$

where t is the statistical year, D_t, D_{t-1}, and D_{t-2} represent the download frequencies of the current year, previous year, and the year before that, respectively, and P_{t-2} is the number of citable documents from the year before last. The denominator is divided by 100 to reduce the magnitude of the download factor value and make it more readable.

The download factor has the following characteristics: First, it uses citable literature volume for calculation, excluding literature unrelated to citation metrics evaluation, such as popular science materials. Second, it focuses on both academic dissemination and academic impact. Third, its evaluation timeline is synchronized with the impact factor, focusing on academic dissemination level within two years after journal publication.

3.3 Connotation Analysis of the Download Factor

In academic journal evaluation, the timeliness of evaluation indicators is crucial. Due to citation patterns, citation frequency has not yet reached its maximum value. If the evaluation is conducted too early, it is clearly unreasonable even if timeliness is good. However, if the citation pattern is fully considered, the

lag period becomes too long, losing evaluation timeliness. The connotation of the download factor is first and foremost academic dissemination, an important manifestation of academic communication. Download frequency also determines paper impact.

The knowledge and information volume of journal papers is the driving force behind download behavior. It determines download volume and is related to academic influence and quality. The typical indicator representing academic influence is the impact factor, while the typical indicator representing knowledge information volume is publication volume. The h-index represents high-level influence and is related to academic quality. Therefore, to analyze the connotation composition of the download factor, the following model is used:

$$\log(DF) = c + \alpha_1 \log(IF) + \alpha_2 \log(H) + \alpha_3 \log(P) + \mu$$

where IF is the impact factor, H is the h-index, P is publication volume, and μ is the random error term. The analysis must be conducted using the same time frame. The data involved in the download factor are from the 3 years before the statistical year. Therefore, the h-index must also be calculated based on citation frequencies from the 3 years before the statistical year.

---

4 Empirical Results

4.1 Calculation Results of the Download Factor

The calculation results of the download factor are shown in Table 2 . Journals ranking at the top include *Library and Information Service*, *Library Science Research*, and *Information Theory and Practice*. Although the download factor is an average download indicator, journals with higher download factors still have larger publication volumes, fully demonstrating that a journal' s knowledge information volume has a significant impact on the download factor. *Chinese Journal of Library Science* has a relatively low publication volume of only 72 papers, but its download factor still ranks seventh, which is quite outstanding.

4.3 Statistical Properties of the Download Factor

The descriptive statistics of the download factor are shown in Figure 3 [Figure 3: see original paper]. The mean is 6.398, the standard deviation is 4.438, and it has good discriminative power. The Jarque-Bera normality test value is 12.072 with a p-value of 0.002, rejecting the null hypothesis of normal distribution. Like many citation indicators such as the impact factor and total citation frequency, it does not follow a normal distribution. The download factor has the highest correlation with the h-index, the main indicator representing journal quality and influence, and also has high correlations with the impact factor and publication

volume, demonstrating good statistical indicator properties and rich connotation as a journal evaluation indicator.

---

5 Conclusions and Discussion

Considering the multicollinearity problem among the impact factor, h-index, and publication volume, traditional regression is not appropriate. Ridge regression is used, which can effectively reduce multicollinearity. When the sum of standardized coefficients is 0.4, the regression results are stable, so the results at this point are taken as the final regression results:

$$\log(DF) = c + 0.494\log(H) + 0.424\log(IF) + 0.259\log(P)$$

The R^2 is 0.940. From the ridge regression results, the factor with the greatest impact on the download factor is the h-index (elasticity coefficient 0.494), followed by the impact factor (0.424), and then publication volume (0.259). The model has high goodness-of-fit.

Research indicates that download frequencies with lags of 1 and 2 years determine 80% of citation frequency. This paper innovatively adopts a panel data model to comprehensively evaluate the impact of download frequency on citation frequency using both current and lagged periods, greatly improving prediction accuracy. The timeline for download factor indicators is synchronized with the impact factor, both being 2 years after journal publication. Empirical research results show that the download factor can effectively measure a journal' s knowledge information volume, influence, and academic quality. Although the download factor appears to be a single indicator on the surface, its connotation is very rich, representing both the journal' s dissemination level and its influence, academic quality, and knowledge information volume. Essentially, it is an evaluation indicator with multiple information dimensions.

However, the download factor indicator requires further validation across more disciplines and datasets. This article is based on research conclusions drawn from 19 CSSCI journals in library and information science. The relationship between download frequency and citation frequency in other disciplines, as well as the construction of download factors, requires further research using the latest data.

References

[1] GARFIELD E. How can impact factors be improved?[J]. BMJ, 1996, 313(7054): 411-413.
[2] LIU X L. Establishment of download half-life of sci-tech periodicals and

its bibliometrics significance[J]. Chinese journal of scientific and technical periodicals, 2012, 23(4): 561-564.
[3] XU X J. Empirical research on half-life period of journals based on downloads[J]. Journal of intelligence, 2014, 33(6): 117-121.
[4] WANG C, LI S N, LI X J. Research on the frequency distribution of journal paper download and its formation mechanics[J]. Information science, 2016, 34(12): 59-63.
[5] SHARMA H P. Download counts - An early indicator for monitoring progress of science[J]. Current science, 2007, 92(10): 1323-1323.
[6] SU X N. Constructing the evaluation system of academic journals of humanities and social sciences[J]. Dongyue tribune, 2008, 29(1): 35-42.
[7] DING Z Q, ZHENG X N, WU X M. Correlation analysis between citation frequency and download frequency of scientific papers[J]. Chinese journal of scientific and technical periodicals, 2010, 21(4): 467-470.
[8] O' LEARY D E. The relationship between citations and number of downloads in Decision Support Systems[J]. Decision support systems, 2008, 45(4): 972-980.
[9] XIE J, GONG K L, CHENG Y, et al. Meta-analysis of the correlation between downloads and citations at paper level[J]. Journal of the China society for scientific and technical information, 2017, 36(12): 1255-1269.
[10] BOTTING N, DIPPER L, HILARI K. The effect of social media promotion on academic article uptake[J]. Journal of the association for information science and technology, 2017, 68(3): 795-800.
[11] SCHLOGL C, GORRAIZ J, GUMPENBERGER C, et al. Comparison of downloads, citations and readership data for two information systems journals[J]. Scientometrics, 2014, 101(2): 1113-1128.
[12] HU M. The law of journal papers web download and correlation of the citation index[J]. Journal of intelligence, 2012, 31(4): 14-18.
[13] ZHAO Y Q, WANG Z M, XIONG W B, et al. Research on the relationship between download and citation of scientific papers: Taking ACM digital library as an example[J]. Chinese journal of scientific and technical periodicals, 2014, 25(6): 818-823.
[14] BRODY T, HARNAD S, CARR L. Earlier Web usage statistics as predictors of later citation impact: Research Articles[J]. Journal of the American society for information science and technology, 2006, 57(8): 1060-1072.
[15] NIU Y X, ZONG Q J, YUAN Q J. A bibliometric study on downloading and citation of open access papers[J]. Journal of library science in China, 2012, 38(4): 119-127.
[16] XIONG Z Q, DUAN Y F. Can downloads predict subsequent citations: A case study on journals of library and information science[J]. Documentation, information & knowledge, 2018(4): 32-42.
[17] COATS A J S. The top papers by download and citations from the International Journal of Cardiology in 2007[J]. International journal of cardiology, 2008, 131(1): e1-e3.
[18] ZHU W, CHEN R, LIU Y. Relationship between citations and the number of downloads of journals: Based on compound H-index[J]. Digital library forum,

2018(10): 25-31.
[19] LU W, QIAN K, TANG X B. Correlation analysis between document citation frequency and download frequency - In the field of library & information science[J]. Information science, 2016, 34(1): 3-8.

Note: Figure translations are in progress. See original paper for figures.