Correlation Analysis Between Download and Citation Frequencies Based on CNKI: A Case Study of the Journal of Southwest Jiaotong University (Postprint)

Authors: Xu Ping

Date: 2023-10-08T00:00:00+00:00

Abstract

Objective: To provide a theoretical basis for enhancing the academic impact of the Journal of Southwest Jiaotong University. Methods: Utilizing the CNKI Citation Database, with 818 articles published in the Journal of Southwest Jiaotong University from 2014 to 2018 as the statistical data source, data correlation analysis was performed between article download frequency and citation frequency, the linear correlation coefficient was calculated and analyzed, and the feasibility of predicting citation frequency using download frequency was discussed. Results: The top 50 papers by citation frequency were mainly concentrated in 2014-2015 (4-5 years post-publication), with 2014 ranking first (27 papers, 54%); followed by papers published in 2015 (15 papers, 30%); no papers published in 2017 and 2018 (1-2 years post-publication) entered the top 50. The top 50 papers by download frequency were concentrated in 2014-2016 (3-5 years post-publication), each accounting for 32%. Conclusion: When download frequency ranges from 500 to 1000, the goodness of fit with citation frequency is high; when exceeding 1000, the distribution becomes relatively dispersed. When citation frequency is approximately 20, the goodness of fit with download frequency is high; when exceeding 20, the distribution becomes relatively dispersed with poor goodness of fit. Predicting citation frequency using download frequency is feasible within a specific range.

Full Text

Correlation Analysis of Paper Download Frequency and Citation Frequency Based on CNKI: A Case Study of the *Journal of Southwest Jiaotong University*

Abstract: Objective: To provide a theoretical basis for improving the academic influence of the *Journal of Southwest Jiaotong University*. Methods: Using the China National Knowledge Infrastructure (CNKI) citation database, 818 papers published in the *Journal of Southwest Jiaotong University* from 2014 to 2018 were used as the data source. Correlation analysis was performed between download frequency and citation frequency, the linear correlation coefficient was calculated, and the feasibility of using download frequency to predict citation frequency was discussed. Results: The top 50 most-cited papers were concentrated in 2014–2015 (4–5 years after publication), with 2014 ranking first (27 papers, accounting for 54%); followed by 2015 (15 papers, accounting for 30%); no papers from 2017 or 2018 (1–2 years after publication) entered the TOP50. The top 50 most-downloaded papers were concentrated in 2014–2016 (3–5 years after publication), each accounting for approximately 32%. Conclusion: When download frequency is between 500–1000 times, the fit with citation frequency is good; when above 1000 times, the distribution is relatively scattered. When citation frequency is around 20 times, the fit with download frequency is good; when greater than 20 times, the distribution is relatively scattered with poor fit. It is feasible to use download frequency to predict citation frequency within a specific range.

Keywords: download frequency; citation frequency; goodness of fit; correlation

With the rapid development and popularization of internet technology, digitalization has become an important pathway for scientific paper dissemination. Metrics such as citation frequency, download frequency, journal impact factor, and CI index play significant roles in evaluating academic influence. As a concentrated reflection of literature value, download frequency and citation frequency have received widespread attention in academia. Some studies have proposed using download frequency as an alternative to citation frequency to address the time lag problem in citation evaluation [1-4], suggesting that download frequency could become a new indicator for considering paper dissemination and impact.

Previous research has found [5-12] that download frequency has a strong correlation with citation frequency (after two years), and that citation frequency can be predicted through corresponding download frequency. Some studies have used early download frequencies of journals to estimate later (two-year) citation frequencies through regression equations, finding that journal download frequency and citation frequency are highly positively correlated but not in a simple linear relationship. However, other studies have directly denied the high consistency between citation frequency and download frequency, or even denied

any correlation between them. For example, Andrew et al. analyzed the *International Journal of Cardiology* and found no obvious relationship between citation frequency and download frequency, thereby questioning the use of citation frequency as a decisive factor in evaluating paper influence [13]. Du Xiujie et al. used simple linear correlation coefficients to demonstrate that academic paper citation frequency is not simply proportional to download frequency [14].

If the two have a strong correlation, can citation frequency be directly predicted using download frequency? Further investigation into their relationship is necessary.

1.1 Research Subject

Data Source: Using the advanced search function of the CNKI "Chinese Citation Database" (http://ref.cnki.net.knsref/index.aspx) (search date: March 1, 2019), "Journal Name" was selected as the search field and "西南交通大学学报" was entered as the search term. The publication period was set from 2014 to 2018 for EI source journals. The download and citation status of papers published in the journal from 2014 to 2018 was retrieved, sorted in descending order by download frequency and citation frequency respectively, and relevant information including article title, publication year, download frequency, and citation frequency was imported into Excel for data analysis.

1.2 Research Methods

Article download frequency and citation frequency are two fundamental metrics in bibliometric evaluation systems. When studying the relationship between two random variables, the simple linear correlation coefficient from mathematical statistics is commonly used, with the specific formula given in [15,16]. In this formula, r represents the simple linear correlation coefficient, n represents the sample size, x represents citation frequency, and y represents download frequency.

Based on quantitative analysis, this study explores the correlation between download frequency and citation frequency, uses the correlation coefficient to determine the strength of the relationship, and employs curve estimation to preliminarily determine the functional relationship curve equation between download frequency and citation frequency, primarily selecting the optimal curve model to fit the data.

2.1 Year Distribution of Top 50 Downloaded and Top 50 Cited Papers

Download frequency can reflect the diffusion rate of online papers and serves as a new indicator of journal dissemination efficiency in the network environment. Paper download volume immediately reflects literature usage and, to a certain

extent, reflects the value of used but ultimately uncited literature, demonstrating stronger timeliness than citation frequency.

The publication years of the top 50 papers by citation frequency and download frequency in the *Journal of Southwest Jiaotong University* from 2014 to 2018 on CNKI were statistically analyzed, with results shown in Table 1 . The data indicate that the top 50 most-cited papers were concentrated in 2014-2016, with 2014 having the most (27 papers, accounting for 54%), followed by 2015 (15 papers, accounting for 30%). No papers from 2017 or 2018 entered the TOP50. The top 50 most-downloaded papers were concentrated in 2014-2016 (3-5 years after publication), each accounting for approximately 32%.

A statistical analysis of the top 25 most-downloaded papers in the *Journal of Southwest Jiaotong University* from 2014 to 2018 on CNKI was conducted, with results shown in Table 2 . The most-downloaded paper was published in 2015 (4,347 downloads) with 121 citations, primarily cited by journal articles and master' s theses. Its download frequency was more than double that of the second-ranked paper published in 2016.

The correlation between citation frequency and download frequency for the top 25 most-downloaded papers is shown in Figure 1 [Figure 1: see original paper], with a correlation coefficient $R = 0.431$ and the relationship equation $y = 0.028x + 6.3127$. As shown in Figure 1, download frequencies were mainly concentrated in the 500-1000 range, with corresponding citation frequencies primarily below 60 times. When download frequency was between 500-1000 times, the fit with citation frequency was good; when above 1000 times, the distribution became relatively scattered.

The specific citation distribution is shown in Figure 2 [Figure 2: see original paper]. The total citation frequency for the top 25 most-downloaded papers was 854, comprising 412 citations (48.24%) from journal articles, 64 (7.49%) from doctoral dissertations, 351 (41.10%) from master' s theses, and 23 (2.69%) from conference papers.

2.3 Analysis of Download Frequency and Specific Citation Distribution for Top 25 Cited Papers

A statistical analysis of the top 25 most-cited papers in the *Journal of Southwest Jiaotong University* from 2014 to 2018 on CNKI was conducted, with results shown in Table 3 . The most-cited paper was published in 2014 (152 citations) with 1,459 downloads, primarily cited by journal articles and master' s theses. Although its citation frequency was only 31 times higher than the second-ranked paper from 2015, its download frequency was approximately 3,000 times lower. The citation frequencies of the top 25 most-cited papers were mainly concentrated around 20 times.

The correlation between citation frequency and download frequency for the top 25 most-cited papers is shown in Figure 3 [Figure 3: see original paper], with

a correlation coefficient $R = 0.4583$ and the relationship equation $y = 18.166x + 193.1$. The specific citation distribution is shown in Figure 4 [Figure 4: see original paper]. Citation frequencies were mainly concentrated around 20 times, with corresponding download frequencies primarily around 500 times. When citation frequency was around 20 times, the fit with download frequency was good; when greater than 20 times, the distribution became relatively scattered with poor fit.

The total citation frequency for the top 25 most-cited papers was 933, comprising 476 citations (51.01%) from journal articles, 56 (6.00%) from doctoral dissertations, 386 (41.10%) from master' s theses, and 25 (2.68%) from conference papers.

3.1 Inconsistent Publication Years for Highly Downloaded and Highly Cited Papers

Citation frequency is an important indicator for evaluating academic quality and influence. Download frequency directly reflects literature usage by readers and indicates the degree of attention a paper receives, though not all downloads result in citations. According to this study' s results, download frequency peaks 3–5 years after publication, while citation frequency peaks 4–5 years after publication. This time lag between high downloads and high citations explains the phenomenon of mismatched publication years for highly downloaded and highly cited papers.

3.2 Good Correlation Between Download Frequency and Citation Frequency

Higher download frequency indicates greater reader attention and increases the probability of being cited. Download frequency directly reflects paper usage by readers and can be considered an early indicator of academic value. Literature citation frequency is highly correlated with its quality; cited papers mean research results have been developed or evaluated, with higher citation frequency producing more significant effects.

3.3 Limitations of Using Download Frequency to Predict Subsequent Citation Frequency

People generally believe that paper download frequency is positively correlated with citation frequency—that is, the more a paper is downloaded, the higher its citation frequency. However, whether to cite a downloaded paper depends on its intrinsic quality. While download frequency and citation frequency share some correlation, it is not a complete linear relationship. Within a certain download frequency range (500–1000 times in this study), download frequency and citation frequency show linear correlation, but beyond this range, the correlation is weak.

Therefore, download frequency cannot be completely used to predict subsequent citation frequency.

References

[1] Wen Xiaoping, Qu Lichun, Ma Qiuming, et al. Analysis of highly cited papers in the *Journal of Northwest A&F University (Natural Science Edition)* from 2001–2012[J]. *Journal of Northwest A&F University (Natural Science Edition)*, 2014, 42(11): 225-230.

[2] Wen Xiaoping. Statistics and analysis of disciplinary distribution of highly cited papers in 21 agricultural university journals selected as comprehensive agricultural science Chinese core journals[J]. *Journal of Library and Information Science in Agriculture*, 2016, 28(1): 51-56.

[3] Zhang Xiaoli, Le Jianxin. Analysis and implications of highly cited paper characteristics[J]. *Journal of Southeast University (Natural Science Edition)*, 2012, 23(6): 1008-1012.

[4] Yang Hong. The relationship between citation frequency and download times of academic journals[J]. *Journal of Anhui Agricultural Sciences*, 2013, 41(4): 1820-1821.

[5] Lu Wei, Qian Kun, Tang Xiangbin. Research on the correlation between literature download frequency and citation frequency: Taking the field of library and information science as an example[J]. *Information Science*, 2016, 34(1): 1-6.

[6] Chen Guangren, Liu Yuanmin. Emphasizing citation rate of scientific papers to improve China' s scientific and technological influence[J]. *Science and Technology Review (Beijing)*, 2008, 26(5): 96-97.

[7] Ding Zuoqi, Zheng Xiaonan. Contradictory analysis of journal impact factor, paper citation times and academic quality evaluation[J]. *Chinese Journal of Scientific and Technical Periodicals*, 2009, 20(2): 286-288.

[8] Li Yajun. Statistical analysis of highly cited papers in Chinese scientific and technological core journals[J]. *Journal of Hebei Polytechnic University (Social Science Edition)*, 2010, 10(4): 96-99.

[9] Guo Qiang, Zhao Jin, Liu Siyuan, Zhang Fang, Liu Xinxin. Research on statistical properties of download times of scientific papers[J]. *Information Science*, 2009, 27(5): 690-694.

[10] Samad Jahandideh. Prediction of future citations of a research paper from number of its internet downloads[J]. *Medical Hypotheses*, 2007, 69(2): 458-459.

[11] Guo Qiang, Zhao Jin, Liu Xinxin. Research on estimating later citation frequency using journal download times[J]. *Library Theory and Practice*, 2010(11): 45-49.

[12] Zhang Xiaoqiang. Correlation between journal download frequency and citation frequency and impact factor: A quantitative analysis of CNKI CSCD

and CHSSCD journals as samples[J]. *Information Studies: Theory & Application*, 2011, 34(8): 36-40.

[13] Andrew J S. The top papers by download and citations from the *International Journal of Cardiology* in 2007[J]. *International Journal of Cardiology*, 2008(1): 1-3.

[14] Du Xiujie, Zhao Daliang, Ge Zhaoqing, Miao Ling. Correlation analysis between download frequency and citation frequency of academic papers[J]. *Acta Editologica*, 2009, 21(6): 508-510.

[15] Ding Zuoqi, Zheng Xiaonan, Wu Xiaoming. Correlation analysis between citation frequency and download frequency of scientific papers[J]. *Chinese Journal of Scientific and Technical Periodicals*, 2010, 21(4): 509-512.

[16] Hauke Jan. Comparison of values of Pearson' s and Spearman' s correlation coefficients on the same sets of data[J]. *Quaestiones Geographicae*, 2011, 30(2): 87-93.

(Author affiliation: Editorial Department of *Journal of Southwest Jiaotong University*)

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