Citation Patterns of Chinese Academic Journal Articles: A Case Study of Library and Information Science Journals, 2006-2008 (Postprint)

Authors: Xiong Ze Quan, Yufeng Duan

Date: 2023-07-26T00:00:00+00:00

Abstract

[Purpose/Significance] This study explores the citation patterns of Chinese academic journal papers and the influence of time window selection on citation patterns, establishing an analytical framework for citation pattern analysis. [Method/Process] Taking journal papers in the library and information science field published from 2006 to 2008 as the research object, this paper employs two-step clustering to conduct cluster analysis on the absolute citation counts and relative citation counts of individual papers within a 7-year window, examining the correlation between main characteristic factors of papers and citation patterns. [Results/Conclusion] From the perspective of absolute citation counts, all journal papers exhibit the classic citation pattern of rising then falling; from the perspective of relative download counts, journal papers display six citation patterns, three of which can be summarized as classic citation patterns, while the other three are "Sleeping Beauty-like", positively skewed, and marathon patterns. Under the relative citation count perspective, first-year citations and total citations show moderate to strong correlation, with the correlation strengthening as average citation counts increase; the results under the absolute citation count perspective show the opposite trend. The findings indicate that the level of correlation between initial citation count and total citation count of journal papers depends primarily on the kurtosis of the citation curve rather than the magnitude of total citation count.

Full Text

Preamble

ChinaXiv Cooperative Journal, Volume 63, Issue 8, April 2019

Research on Citation Patterns of Chinese Academic Journal Papers: A Case Study of Journal Articles in Library and Information Science

from 2006–2008

Xiong Zequan[1],[2], Duan Yufeng[1]
[1] Faculty of Economics and Management, East China Normal University, Shanghai 200241
[2] Library, East China Normal University, Shanghai 200241

Abstract

[**Purpose/Significance**] This study explores the citation patterns of Chinese academic journal papers and examines how the choice of time window influences these patterns, establishing an analytical framework for citation pattern analysis. [**Method/Process**] Using journal articles published in library and information science between 2006 and 2008 as the research sample, we employed two-step clustering to analyze both absolute and relative citation counts over a seven-year period, investigating correlations between key paper characteristics and citation patterns. [**Result/Conclusion**] From the absolute citation perspective, all papers exhibited a classic citation pattern of initial rise followed by decline. From the relative citation perspective, six distinct patterns emerged: three could be classified as classic patterns, while the remaining three were "quasi-sleeping beauty," positive-bias, and marathon types. Under the relative citation view, first-year citations showed moderate to strong correlations with total citations, with higher average citation counts yielding stronger correlations—the opposite of the absolute citation perspective. These results indicate that the correlation between initial and total citations depends primarily on the curvature of the citation curve rather than the magnitude of total citations.

1. Introduction

Citation counts have long played a pivotal role in evaluating academic impact. This metric focuses on static total citation numbers without considering their dynamic evolution, representing an outcome-oriented rather than process-oriented evaluation approach. However, papers with identical total citations may follow fundamentally different citation trajectories. Two extreme examples are "flash-in-the-pan" papers [1–2] and "sleeping beauty" papers [2–3]. These types differ not only in academic value but also in their subsequent citation patterns over time. Therefore, researchers must examine both the outcomes and the underlying processes to properly assess scholarly impact.

The citation process can be described through citation curves—graphs showing how citation counts change over time. These dynamic representations are variously termed citation curves [4], citation patterns [5], citation lifecycles [6], citation trajectories [7], or citation histories [8]. Previous research on citation patterns has generally followed two approaches:

(1) Curve-fitting methods. For instance, A. Avramescu proposed a general

formula: $C(t) = C_0[\exp(-\alpha t) - \exp(-m\alpha t)]$, $m > 1$, where C_0 represents amplitude, α is a time constant, and m is the initial increment. However, Avramescu acknowledged this formula cannot capture all curve types, such as bimodal or irregular patterns [10]. Li Jiang et al. analyzed citation curves of 341 Nobel laureates up to 2011, constructing a framework with two main categories: regular curves (including classic and exponential growth) and irregular curves (including sleeping beauty, bimodal, and wave patterns) [9]. The authors noted this framework applies only to established scholars and papers with sufficient citation history.

(2) Clustering methods. E. S. Aversa used K-means clustering on 400 highly-cited papers published in 1972, identifying two patterns over eight years: "delayed growth—slow decline" and "immediate growth—rapid decline," with the former accumulating more total citations [5]. This supported D. D. S. Price's conclusion that less-cited papers age faster [11]. V. Cano and N. C. Lind studied 10 highly-cited and 10 moderately/low-cited papers over 25 years, also finding two patterns: Pattern A (mixed citation levels) showed a significant decline in growth rate by year six with a broken-line cumulative curve, while Pattern B (exclusively highly-cited papers) maintained stable growth with a single-line cumulative curve [6]. S. E. Baumgartner and L. Leydesdorff applied group-based trajectory modeling (GBTM) to papers from six major journals and 24 virology journals over 15 years, identifying "transient knowledge claims" (peaking early then aging) and "sticky knowledge claims" (sustained citations over a decade) [12]. However, these clustering methods require pre-specifying cluster numbers.

These studies vary in data selection (e.g., highly-cited papers, Nobel laureates), methods, and time windows, yielding inconsistent conclusions. Moreover, most research focuses on English-language literature; no studies have examined whether Chinese academic journals exhibit different citation patterns due to their more limited readership. This study analyzes citation patterns of Chinese LIS journal articles to address this gap. Using the same dataset as our previous work on download patterns [13] and dynamic correlations between downloads and citations [14], this research forms part of a comprehensive study system and lays groundwork for future citation prediction models.

This work was supported by the ISTIC-ELSEVIER Journal Evaluation Research Center Open Fund Project "A Study on Disciplinary Differences Between CiteScore and JCR Impact Factor."

Authors: Xiong Zequan (ORCID: 0000-0002-4349-371X), Librarian, PhD Candidate; Duan Yufeng (ORCID: 0000-0002-4349-371X), Professor, PhD Supervisor, Corresponding Author, E-mail: yfduan@infor.ecnu.edu.cn.

Received: August 8, 2018; **Revised:** November 16, 2018; **Pages:** 107-115

2. Data and Methods

2.1 Data Sources and Processing

Using China Academic Journals (Online Edition)—China's largest journal database—as our data source, we selected 9,066 papers published between 2006 and 2008 in 11 LIS journals, including *Journal of Academic Libraries*, *Information Science*, and *Information Theory and Practice*, that had received at least one citation by December 31, 2015. Journals were selected based on long publication history, complete database coverage, and minimal lag between print and online availability to ensure reliable download and citation data. This raw dataset was named **DataSet1**.

Since CNKI began tracking download data in 2006 and previous research has shown that 7–8 years sufficiently captures citation patterns [5], we used citation data from January 1, 2006, to December 31, 2015. **DataSet1** included bibliographic information and annual citation counts for each paper. To account for publication month variability (early vs. late in the year), we adjusted first-year citations assuming uniform monthly distribution:

$$C'Y_1 = C + C_1 \times (12 - M)$$
$$C'Y_2 = C_1 \times (12 - M)$$

where M is the publication month, C is citations in year Y, and $C'Y_1$ is the adjusted citation count for the first year post-publication. Similar adjustments were made for years 3–7, creating **DataSet2**. We then normalized these values by calculating each year's percentage of the seven-year total, yielding **DataSet3**.

For example, a paper published in May 2008 in *Information Science* titled "Research Progress and Application of Chinese Word Segmentation Technology in Natural Language Retrieval" had adjusted first-year citations of 7.17 and second-year citations of 11.67 in **DataSet2**. In **DataSet3**, these represented 9.61% and 15.64% of its total citations, respectively (see Table 1 for details).

Table 1 Example of Citation Representation Across Three Datasets

Year	DataSet1	DataSet2 (Absolute)	DataSet3 (Relative)
2008	11.75	—	—
2009	10.00	7.17 (Year 1)	9.61%
2010	12.33	11.67 (Year 2)	15.64%
2011	10.42	12.33 (Year 3)	16.54%
2012	11.25	11.25 (Year 4)	15.08%
2013	10.00	10.00 (Year 5)	13.41%
2014	11.75	10.42 (Year 6)	13.97%
2015	10.00	—	—

2.2 Analysis Methods

(1) Clustering Analysis. We used IBM SPSS Statistics 23's two-step clustering algorithm on both absolute citations (**DataSet2**) and relative citations (**DataSet3**). Two-step clustering handles large samples with mixed variable types, automatically determining optimal cluster numbers through pre-clustering and sub-clustering stages. We specified annual citation counts as continuous variables, used the Bayesian Information Criterion (BIC), disabled outlier noise handling (given prior data cleaning), and enabled cluster membership output.

(2) Correlation Analysis. Spearman correlation coefficients were calculated to examine relationships between citation counts and paper characteristics (title length, author count, keyword count, journal impact factor, and initial citations) across different citation patterns.

3. Results

3.1 Citation Patterns Based on Absolute Citations

Clustering **DataSet2** yielded three citation patterns (Figure 1 [Figure 1: see original paper]), all showing a consistent rise-then-fall trend, though peak timing and magnitude varied:

- **Pattern 1** (4.04% of sample): High-citation papers peaking in year 3 (average 67.52 citations)
- **Pattern 2** (31.66%): Medium-citation papers peaking earlier
- **Pattern 3** (64.31%): Low-citation papers (average 5.15 citations) peaking earliest

Pattern 1 accounted for 22.11% of total citations despite its small share, confirming that highly-cited papers age more slowly [5, 11–12].

Table 2 Sample Size, Total Citations, and Average Citations for Three Absolute Citation Patterns

Pattern	Papers	% of Sample	Avg. Citations	% of Total Citations
1	366	4.04%	67.52	22.11%
2	2,870	31.66%	19.87	51.02%
3	5,830	64.31%	5.15	26.88%

3.2 Citation Patterns Based on Relative Citations

Clustering **DataSet3** produced six patterns (68 papers with citations only in 2015 were excluded as they could not be normalized). The 8,998 valid samples revealed:

- **Pattern A** (9.54%): Rapid rise, sharp decline, then slow recovery

- **Pattern B** (27.76%): Slow rise and fall (classic pattern)
- **Pattern C** (11.99%): Similar to A but with delayed peak
- **Pattern D** (38.41%): Very gradual rise and fall (marathon type)
- **Pattern E** (7.10%): Similar to A/C but with further delayed peak
- **Pattern F** (4.90%): Near-zero citations for 3–4 years, then sudden increase ("quasi-sleeping beauty")

Patterns A, C, and E represent variations of a single-peaked curve shifted along the x-axis. Pattern D shows the slowest aging, while Pattern F exhibits sleeping-beauty characteristics but with insufficient post-awakening citations to meet Van Raan's quantitative criteria [3].

Table 3 Annual Citation Percentages and Means for Six Relative Citation Patterns

Pattern	Year 1	Year 2	Year 3	Year 4	Year 5	Year 6	Year 7
A	17.57%	26.48%	26.58%	13.28%	4.75%	5.59%	1.15%
B	9.61%	15.64%	16.54%	15.08%	13.41%	13.97%	2.89%
C	4.86%	17.95%	44.56%	17.16%	6.16%	4.70%	1.70%
D	1.44%	3.57%	14.96%	10.80%	12.62%	37.66%	2.26%
E	1.48%	3.57%	7.40%	5.30%	4.14%	17.09%	44.76%
F	0.96%	1.70%	2.88%	2.91%	15.16%	40.64%	29.23%

Table 4 Sample Size, Total Citations, and Average Citations for Six Relative Citation Patterns

Pattern	Papers	% of Sample	% of Total Citations	Avg. Citations
A	858	9.54%	2.71%	5.80
B	2,517	27.97%	30.74%	13.65
C	1,087	12.08%	6.54%	6.96
D	3,456	38.41%	57.22%	18.50
E	639	7.10%	4.90%	6.90
F	441	4.90%	1.25%	1.94

3.3 Matrix Analysis of Absolute and Relative Citation Patterns

Cross-tabulating the two classification systems (Table 5) revealed that Pattern 1 (high absolute citations) primarily corresponded to Patterns B and D in relative terms. Among 90 high-citation papers (top 1%), 11 followed Pattern B, 78 followed Pattern D, and only 1 followed Pattern F ("Development of Bibliographic Co-occurrence Mining Systems in Document Databases"). This suggests Patterns B and D are necessary but not sufficient conditions for high citations.

Table 5 Sample Matrix of Absolute vs. Relative Citation Patterns

Relative Absolute	Pattern 1	Pattern 2	Pattern 3	Total
Pattern A	1 (0.01%)	66 (0.73%)	791 (8.79%)	858 (9.54%)
Pattern B	95 (1.06%)	1,216 (13.51%)	1,206 (13.40%)	2,517 (27.97%)
Pattern C	1 (0.01%)	179 (1.99%)	907 (10.08%)	1,087 (12.08%)
Pattern D	268 (2.98%)	1,395 (15.50%)	1,793 (19.93%)	3,456 (38.41%)
Pattern E	0 (0.00%)	11 (0.12%)	628 (6.98%)	639 (7.10%)
Pattern F	1 (0.01%)	3 (0.03%)	437 (4.86%)	441 (4.90%)
Total	**366**	**2,870**	**5,830**	**8,998**

3.4 Correlation Between Citations and Paper Characteristics

Correlation analysis revealed:

1. **Across all patterns**, citations showed no significant correlation with title length, author count, keyword count, or impact factor.
2. **Under the relative citation perspective**, first-year citations correlated moderately to strongly with total citations, with stronger correlations in higher-citation patterns (Table 8).
3. **Under the absolute citation perspective**, this relationship was inverse: higher-citation patterns showed weaker correlations (Table 7).

For 90 high-citation papers, total citations correlated weakly only with author count ($r = 0.304$, $p < 0.01$). First-year citations showed no correlation; only three-year cumulative citations correlated weakly with total citations ($r = 0.371$, $p < 0.001$).

Table 6 Correlation Between High-Citation Papers' Total Citations and Characteristics

Characteristic	Correlation	Significance
Title length	-0.062	0.560
Author count	0.304**	0.004
Keyword count	-0.029	0.788
Impact factor	0.100	0.350
Year 1 citations	0.085	0.427
2-year cumulative	0.243*	0.021
3-year cumulative	0.371**	0.000

Table 7 Correlations Under Absolute Citation Patterns

Characteristic	Pattern 1	Pattern 2	Pattern 3
Title length	-0.057	-0.044*	-0.020
Author count	0.132*	0.039*	0.097**
Keyword count	-0.005	-0.038*	0.102**

Characteristic	Pattern 1	Pattern 2	Pattern 3
Impact factor	0.151**	0.092**	0.061**
Year 1 citations	0.106*	0.082**	0.344**
2-year cumulative	0.290*	0.363**	0.576**
3-year cumulative	0.454**	0.587**	0.730**

Table 8 Correlations Under Relative Citation Patterns

Characteristic	Pattern A	Pattern B	Pattern C	Pattern D	Pattern E	Pattern F
Title length	-0.038	0.000	0.049	-0.054	-0.016	0.014
Author count	-0.102**	0.007	0.055	0.031	0.041	0.067*
Keyword count	0.035	0.101*	0.126**	0.109**	0.411**	0.502**
Impact factor	-0.037**	0.090**	0.064**	0.065**	0.632**	0.805**
Year 1 citations	0.594**	0.726**	0.595**	0.690**	0.397**	0.411**
2-year cumulative	0.833**	0.932**	0.843**	0.859**	0.625**	0.502**
3-year cumulative	0.847**	0.959**	0.953**	0.920**	0.737**	0.614**

3.5 Topic Analysis

Analyzing papers by third-level subject classification codes (Table 9) showed:

- **Information Theory (G201)** had the highest proportion of Pattern 1 papers (15.74%)
- **Library Science (G250)** contributed the most Pattern 1 papers (18.31% of total) but had only 4.48% within the topic
- **Document Work (G255)**, **Philology (G256)**, and **Information Industry Economy (F49)** contained no Pattern 1 papers

Topic paper volume correlated significantly with Pattern 1 count (Spearman r = 0.631), suggesting larger topics are more likely to produce highly-cited papers.

Table 9 Pattern 1 Distribution Across Topics

Topic	Papers	Pattern 1	% Within Topic	% of Pattern 1
Information Theory (G201)	127	20	15.74%	5.46%
Library Science (G250)	1,675	75	4.48%	18.31%
Computer Applications	1,023	44	4.30%	12.02%
Document Processing	1,002	42	4.19%	11.48%
Information Resource Management	820	21	2.56%	5.74%

4. Discussion

4.1 Types of Citation Patterns

This study provides a novel perspective by analyzing both absolute and relative citation patterns. The absolute perspective yielded three patterns distinguished primarily by citation magnitude, masking variation in temporal dynamics. All followed the classic rise-and-fall shape, confirming universal literature aging.

The relative perspective revealed six patterns. Patterns A, C, and E represent single-peaked curves shifted along the time axis—similar to Avramescu's "normal" pattern [10] but rarely associated with high absolute citations. Pattern D (marathon type) showed the slowest aging, dominating among highly-cited papers. Pattern F ("quasi-sleeping beauty") exhibited delayed recognition but insufficient post-awakening citations to qualify as true sleeping beauties [3].

4.2 Impact of Time Window Selection

Time window choice critically affects citation pattern identification. Short windows may miss sleeping beauties [20], while longer windows can transform single-peaked into bimodal curves or exponential into normal patterns. Short windows also disadvantage social sciences, which cite more slowly than natural sciences [21].

Our findings show that under the relative perspective, first-year citations correlate strongly with total citations (r 0.7) in high-citation patterns (B and D), whereas the absolute perspective shows inverse correlations. This demonstrates that the initial-total citation relationship depends on curve curvature, not absolute volume. For high-cited papers, slower-aging patterns (D) show stronger early-total correlations than rapid-rise patterns.

4.3 Analytical Framework for Citation Patterns

Absolute and relative perspectives offer complementary advantages. Absolute citations distinguish high- and low-impact papers but obscure temporal variation within citation levels. Relative citations reveal temporal dynamics but may mask high-impact papers among similarly-shaped moderate-citation curves.

Our matrix analysis framework integrates both perspectives, enabling comprehensive aging analysis while identifying both outcomes and processes. This dual approach reveals that higher absolute citations correlate with flatter relative curves, but the converse is not true. Most importantly, initial-total citation correlations depend on curve shape rather than magnitude, refining understanding of early citation prediction.

5. Limitations and Future Work

This study's focus on LIS journals limits generalizability to other disciplines. Future research should test the framework across fields. Building on our download

pattern [13] and download-citation correlation studies [14], we plan to develop citation prediction models using early download and citation metrics to identify potentially highly-cited papers and scholars, supporting research management and decision-making.

Acknowledgments

We thank CNKI for providing data and Ms. Yang Li of East China Normal University Library for guidance on data analysis.

References

[1] VANDALEN HP, HENKENS K. Signals in science? On the importance of signaling in gaining attention in science[J]. Scientometrics, 2005, 64(2): 209–233.

[2] 李江. 科学中的“睡美人”与“昙花一现”现象评述 [J]. 大学图书馆学报, 2016(3): 38–43.

[3] VAN RAAN A. Sleeping beauties in science[J]. Scientometrics, 2004, 59(3): 467–472.

[4] GARFIELD E. More delayed recognition. 1. Examples from the genetics of color-blindness, the entropy of short-term-memory, phosphoinositides, and polymer[J]. Current contents, 1989, 38: 3–8.

[5] AVERSA ES. Citation patterns of highly cited papers and their relationship to literature aging: A study of the working literature[J]. Scientometrics, 1985, 7(3/6): 383–389.

[6] CANO V, LIND NC. Citation life-cycles of 10 citation classics[J]. Scientometrics, 1991, 22(2): 297–312.

[7] 杜建, 武夷山. 文献引文轨迹: 分类及测度 [J]. 情报理论与实践, 2015(7): 52–58.

[8] COLAVIZZA G, FRANCESCHET M. Clustering citation histories in the physical review[J]. Journal of informetrics, 2016, 10(4): 1037–1051.

[9] 李江, 姜明利, 李癮婷. 引文曲线的分析框架研究——以诺贝尔奖得主的引文曲线为例 [J]. 中国图书馆学报, 2014(2): 41–52.

[10] AVRAMESCU A. Actuality and obsolescence of scientific literature[J]. Journal of the American Society for Information Science, 1979, 30(5): 296–303.

[11] PRICE DDS. A general theory of bibliometric and other cumulative advantage processes[J]. Journal of the American Society for Information Science, 1976, 27(5): 292–306.

[12] BAUMGARTNER SE, LEYDESDORFF L. Group-based trajectory modeling (GBTM) of citations in scholarly literature: Dynamic qualities of “transient” and “sticky knowledge claims”[J]. Journal of the Association for Information Science and Technology, 2014, 65(4): 797–811.

[13] DUAN Y, XIONG Z. Download patterns of journal papers and their influencing factors[J]. Scientometrics, 2017, 112(3): 1761–1775.

[14] 熊泽泉, 段宇锋. 论文早期下载量可否预测后期被引量?——以图书情报领域期刊为例 [J]. 图书情报知识, 2018(4): 32–41.

[15] AKSNES DW, SIVERTSEN G. The effect of highly cited papers on national citation indicators[J]. Scientometrics, 2004, 59(2): 213–224.

[16] AKSNES DW. Characteristics of highly cited papers[J]. Research evaluation, 2003, 12(3): 159–170.

[17] TAHAMTANI I, AFSHAR A, AHMADZADEH K. Factors affecting number of citations: A comprehensive review of the literature[J]. Scientometrics, 2016, 107(3): 1195–1225.

[18] LEVITT JM, THELWALL M. Patterns of annual citation of highly cited articles and the prediction of their citation ranking: A comparison across subjects[J]. Scientometrics, 2008, 77(1): 41–60.

[19] WALTERS GD. The citation lifecycle of articles published in 13 American psychological association journals: A 25-year longitudinal analysis[J]. Journal of the American Society for Information Science and Technology, 2011, 62(8): 1629–1636.

[20] WANG J. Citation time window choice for research impact evaluation[J]. Scientometrics, 2013, 94(3): 851–872.

[21] GLANZEL W, SCHOEPFLIN U. A bibliometric study on aging and reception processes of scientific literature[J]. Journal of informetrics, 1995, 21(1): 37–53.

[22] ADAMS J. Early citation counts correlate with accumulated impact[J]. Scientometrics, 2005, 63(3): 567–581.

[23] LI J, YE FY. A probe into the citation patterns of high-quality and high-impact publications[J]. Malaysian journal of library & information science, 2014, 19(2): 17–33.

[24] LI J, YE FY. Distinguishing sleeping beauties in science[J]. Scientometrics, 2016, 108(2): 821–828.

[25] EGGHE L, RAO I. Citation age data and the obsolescence function: Fits and explanations[J]. Information processing & management, 1992, 28(2): 201–217.

[26] MCCAIN KW, TURNER K. Citation context analysis and aging patterns of journal articles in molecular-genetics[J]. Scientometrics, 1989, 17(1/2): 127–163.

Author Contributions

Xiong Zequan: Data analysis, manuscript writing and revision.
Duan Yufeng: Manuscript revision and final approval.

Citation Patterns of Chinese Academic Journal Papers: A Case Study of Journal Articles (2006–2008) in the Field of Library and Information Science

Xiong Zequan[1],[2], Duan Yufeng[1]
[1] Faculty of Economics and Management, East China Normal University, Shanghai 200241
[2] Library, East China Normal University, Shanghai 200241

Abstract: [Purpose/significance] This paper explores the citation patterns of Chinese academic journal papers and the influence of time window, and establishes an analysis frame of the citation patterns. [Method/process] Taking library and information science for example, a two-step cluster analysis on the absolute citations and the relative citations of Chinese journal papers published between 2006 and 2008 was performed. Correlation between main characteristic factors and citation patterns of papers was analyzed. [Result/conclusion] Three patterns were detected from the perspective of absolute citations, and all of them show a classic citation pattern—rise first and then fall. Six citation patterns are detected from the perspective of relative citations, and three of them can be classified as classic citation pattern, while other three clusters can be labeled as quasi-sleeping beauties, positive-bias pattern and marathons. Moderate to high correlations between initial citations and total citations under the relative perspective, and the correlation strengthen with the average citation counts which reversed under the absolute perspective. The results indicate that the correlation between initial citations and total citations depends on the curvature of citation curve but not the number of total citations.

Keywords: citation; citation patterns; time windows; obsolescence; correlation analysis

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