Postprint: A Study on the Relationship Between Paper-Related Parameters and Citation Frequency

Authors: Xiao Xuebin, Chai Yanju

Date: 2017-10-11T00:00:00+00:00

Abstract

[Objective] To investigate whether certain paper-related parameters affect citation frequency.

[Method] Multiple measures were employed to mitigate interference from non-research factors, and temporal curves of the relationship between research factors and citation frequency were plotted to determine the influence of research factors on citation frequency.

Results The number of authors, page count, number of references, and abstract length were positively correlated with citation frequency; the number of author keywords and their average length were unrelated to citation frequency; while different title lengths had varying effects on citation frequency.

Limitations Due to data sampling constraints, all data were collected from the SCIE database, a prestigious citation database, with WOS categories limited to Engineering and Mechanical; the conclusions may not be fully generalizable to papers in other disciplines.

Conclusion Certain paper-related parameters exert an influence on citation frequency.

Full Text

A Study on the Relationship Between Relevant Parameters of Papers and Citation Frequency

Xiao Xuebin[1], **Chai Yanju**[2]
[1]Wuhan University Library, Wuhan 430079, China
[2]Institute of Geodesy and Geophysics, Chinese Academy of Sciences, Wuhan 430077, China

Abstract

[**Objective**] To investigate whether certain parameters of academic papers influence citation frequency. **Methods** Multiple measures were adopted to minimize interference from non-research factors, and temporal curves depicting the relationship between research factors and citation frequency were plotted to assess their impact. **Results** The number of authors, page count, number of references, and abstract length showed positive correlation with citation frequency, while the number and average length of author keywords showed no correlation. Different title lengths had varying effects on citation frequency. **Limitations** Due to data sampling constraints, all data were drawn from the high-level SCIE database with WOS categories limited to Engineering and Mechanical, so the conclusions may not be universally applicable to papers in other subject areas. **Conclusion** Certain parameters of papers do influence citation frequency.

Keywords: Citation frequency; Influencing factors; Field length; Number of references; Number of authors; Positive correlation

Introduction

In China, citation frequency has become a crucial metric for evaluating the impact of both papers and their authors, garnering increasing attention from evaluation institutions, researchers, and journal editorial offices alike. How to increase citation frequency is a widespread concern among researchers and journal editors. To address this question, we must first identify the factors that influence citation frequency and understand how they exert their effects. Previous studies have examined some of these factors.

When authors write papers, constraints arising from their disciplinary background, knowledge, research conditions, and expertise also limit the journals that accept their work and related journal factors (such as publication cycle, article volume, research field, etc.), as well as the databases that index them. All these factors may influence citation frequency. Existing research suggests that: (1) Citation rates correlate with document type, with reviews and commentaries receiving significantly more citations than applied and experimental studies [1-2]; (2) Citation rates vary by research field (discipline), with hot fields receiving more citations than less popular ones [1-5]; (3) Journal-related factors affect citation frequency—higher journal impact factors correlate with higher citation rates [6], longer publication cycles and higher article volumes correlate with higher citation rates [3], and better print quality correlates with higher citation rates [7]; (4) Editorial regulations on length and editorial work attitude have certain effects on citation rates [3,5,7]; (5) Higher database visibility and broader distribution scope correlate with higher citation rates [5,7,8]; (6) The accessibility and dissemination of papers affect citation rates—the easier they are to obtain and distribute, the higher the citation rate [5,9].

Other paper-specific parameters, such as title length, number and length of keywords, abstract length, paper length, and number of references, may also in-

fluence citation frequency. This study focuses on the effects of these parameters. While some have been investigated previously, findings have been inconsistent. For instance, reference [1] concluded that papers with 4-6 authors had the highest citation rates and average citations per paper, while those with more than 7 authors had lower average citations than the 4-6 author group. In contrast, reference [2] found that the number of authors was proportional to journal impact factor. Reference [9] suggested that title length and paper length had no significant effect on citation frequency, while keyword count had a significant effect. However, reference [10] argued that paper length did affect citation frequency. We have also investigated these issues, and our research approach and methods are described below.

Corresponding author: Xiao Xuebin, ORCID: 0000-0003-1933-5006, E-mail: 00201493@whu.edu.cn

Methods

Measures to Isolate Specific Factors Since many factors simultaneously influence citation frequency, numerous measures were required to minimize the effects of other factors and parameters in order to isolate the impact of specific factors. The measures were as follows:

(1) **Limit the database.** Since databases affect citation frequency [5,8], limiting the database and maintaining consistency in indexing databases can weaken this influence. We selected only the SCIE database for sampling.

(2) **Maintain consistent document quality.** As all sampled data came from SCIE—an internationally recognized database representing high-level research papers—we could ensure sufficient citations for selected papers, avoiding unclear or distorted results due to small citation counts. SCI paper citations are also a focus of academic evaluation.

(3) **Limit the subject area.** Since research field significantly affects citation frequency [1-5], we retrieved all literature from 2010–2012 in SCIE with WOS categories including Engineering and Mechanical on September 25, 2015, obtaining 46,378 records. This ensured that retrieved papers had relatively similar subject areas, minimizing differences in citation frequency due to field variation.

(4) **Limit publication type.** The SCIE database includes four publication types: journals, books, series, and patents. To prevent potential effects from publication type, we filtered out nine non-journal records and removed six incomplete records, resulting in 46,363 journal article records.

(5) **Large sample size.** As is well known, averaging large datasets can minimize random effects, highlight main trends, and reveal universal patterns. Our sample size reached 46,363, whereas the largest sample in references [1-10] was 5,716 [2], with the smallest being only a few dozen and most totaling less than 1,000.

(6) **Moving average method.** Since many factors affect citation frequency and the above measures could not completely eliminate their influence, causing excessive fluctuations in some areas, it was sometimes necessary to use moving averages to smooth abnormal variations and maintain overall trends.

Research and Judgment Methods First, the original records were imported into Excel, and VBA was used to extract data fields such as citation frequency, publication year, page count, number of authors, number of references, and funding support from the original record table, which were saved in a separate data table. Title length (in words), abstract length (in words), number of keywords, and keyword length were also calculated and stored in the same table. Average citation frequencies under various conditions were computed. Due to the wide spans of abstract length and number of references, many specific cases had small sample sizes or even zero counts, so we took averages of adjacent values for analysis.

The data were then sorted to highlight the factors under investigation. For example, when analyzing the relationship between paper length (represented by page count) and citation frequency, we performed annual statistics since citations accumulate over time. Records were first sorted by publication year in ascending order to separate papers from different years, then sorted by page count (the relevant parameter) in ascending order to highlight its effect. With page count as the independent variable and citation frequency as the dependent variable, relationship plots were drawn, yielding three curves of average citation frequency versus page count. If these three curves were essentially consistent, the trend was considered universal; otherwise, errors might have occurred. If the overall trends (rising or falling) of the three curves were consistent, the factor was deemed to influence citation frequency. If all three curves were horizontal, there was no effect. If the curves were consistent but showed different trends at different stages, the independent variable had varying effects across stages. This served as the basis for determining whether parameters affected citation frequency. We employed this method to examine the relationship between average citation frequency and factors including page count, number of authors, title length, abstract length, number and average length of author keywords, number of references, and funding support, with results plotted in Figures 1 [Figure 1: see original paper] through 8 [Figure 8: see original paper].

Correlation Analysis To determine whether genuine relationships existed between average citation frequency and the aforementioned parameters, we used Excel' s Correl function to calculate correlations, including with annual average citation frequencies and overall average citation frequency. To reduce random effects, we specifically analyzed correlations when subsample sizes exceeded 10. The correlation test results are shown in Table 1 .

In correlation analysis, the degree of linear correlation is generally classified

into four levels based on the correlation coefficient r: $0 < |r| \leq 0.3$ (weak), $0.3 < |r| \leq 0.5$ (low), $0.5 < |r| \leq 0.8$ (moderate), and $0.8 < |r| \leq 1$ (strong) [11]. According to this standard, when the number of papers included in the statistics exceeded 10, four parameters showed very high correlation with citation frequency ($r > 0.92$): page count, number of authors, number of references, and title length (when $ 8*words*).*Abstractlengthalsoshowedstrongcorrelation*($r >$ 0.8), *whilenumberofkeywords*, *keywordlength*, *andfulltitlelengthshowedlowcorrelation*($r <$ 0.5).*Whencalculatedusingallpapers*, *onlytitlelength*(8*words*)*andnumberofreferencesshowedveryhighcorrel* words) significantly affect citation frequency. Differences in correlation between the two calculation methods are mainly due to random factors.

Results

Overall, in Figures 1–8, average citation frequency was highest for 2010 and lowest for 2012, consistent with the cumulative nature of citations over time. Moreover, the three curves in each of Figures 1–7 were broadly similar, and the ratio of average citation frequency between funded and non-funded papers in Figure 8 was roughly equivalent across years at 1.86, 1.70, and 1.80, confirming that these figures accurately reflect the relationship between average citation frequency and these parameters. While the trends of the three curves within each figure were similar, curves differed across figures (in slope and shape). Combining Figures 1–7 with Table 1 yields the conclusions shown in Table 2 .

Figure 8 shows that funded papers have higher citation frequency. From 2010–2012, the numbers of funded papers were 6,368, 7,502, and 8,680 respectively, while non-funded papers numbered 8,252, 7,946, and 7,615. These three years represent three scenarios: in 2010, funded papers were significantly fewer than non-funded; in 2011 they were roughly equal; and in 2012 funded papers exceeded non-funded (the proportion of funded papers increased yearly—whether this indicates SCIE papers favor funded papers merits further study). Regardless of scenario, average citation frequency of funded papers was consistently higher than non-funded, indicating that funding support significantly affects citation frequency, consistent with references [12-14].

Discussion

Undeniably, for a paper to be cited, the citer must go through four stages: discovery, access, reading, and citation (academic misconduct aside). Being discovered, accessed, and read are prerequisites for citation—in other words, the easier a paper is to discover, the more likely it is to be cited, and the easier the full text is to access, the greater the citation likelihood [5,9-10]. In the internet era, literature searching via the web occurs through two primary methods: direct use of search engines, which most people adopt due to its low cost, speed, and low barrier to entry [15]; and use of professional literature databases, which is expensive (universities typically spend millions to tens of millions annually on access) and requires learned search skills, thus limiting its use to institutions and individuals with access privileges. Search engines generally employ fuzzy match-

ing, breaking search terms into individual characters or words, matching them against their databases, and ranking results by relevance [16]. From this perspective, the positive correlation between citation frequency and paper/abstract length is easily explained. Researchers typically use several search terms to find needed literature, and longer full texts and abstracts provide more opportunities for these terms to appear in different sentences. The longer the full text and abstract, the higher the probability of being matched, and thus the greater the likelihood of being cited.

One might extend this conclusion to paper titles, assuming longer titles yield higher citation frequency. However, Figure 5 shows otherwise. The 2010–2012 curves all demonstrate: when title length is less than 8 words, citation frequency increases rapidly with title length; when length is between 8–20 words, citation frequency changes slowly; when length exceeds 20 words, citation frequency actually decreases with increasing title length. The consistency across all three curves indicates this pattern is not coincidental but inevitable. What causes this phenomenon? We conducted a specialized investigation.

First, for records with title word counts exceeding 20, we extracted citation frequency, the title itself, and word count into a separate Excel worksheet, obtaining 1,378 records. Sorting by citation frequency from low to high revealed many papers that were commentaries (beginning with "Comments on"), replies (beginning with "Reply to" or "Response to"), discussions (beginning with "Discussion of"), or retractions (beginning with "Closure to") of other papers. These titles contained not only the title of the other paper but also its authors, journal name, volume, issue, and page numbers, resulting in very long titles. There were 104 such papers with an average of 26 words, totaling 50 citations and an average citation frequency of only 0.48; 79 of these had zero citations, accounting for approximately 75.96%. These papers were highly specific and of limited relevance to most other researchers, leading to low citation frequency.

After removing these papers, calculations showed that when title length exceeded 20 words, citation frequency still gradually decreased with increasing length. We hypothesize that this is due to excessive specificity. After a paper is discovered, whether it is read ultimately depends on the relevance between the citer' s research interests and the paper' s content. While longer titles increase the probability of being matched in searches, they also contain more independent concepts, indicating more specific and narrower research scope. As shown in Figure 9 [Figure 9: see original paper], A, B, and C each represent the literature scope involving one independent concept (independent meaning no hierarchical relationship among them), while E (the shaded central area) represents the scope containing all three concepts A, B, and C. The E region is much smaller than any single concept region. Scientific paper titles, typically single sentences, are highly condensed summaries of content, comprehensively or partially reflecting the author' s intent, research theme, or paper highlights. If a title is long, researchers may ignore the paper because some concepts are far from their interests, resulting in no citation. The longer the title and the more

independent concepts it contains, the greater the likelihood of being overlooked. Therefore, excessively long titles reduce citation frequency. Both the probability of being matched and the probability of being noticed simultaneously affect citation frequency—longer titles increase matching probability but decrease attention probability. Figure 5 likely represents the combined effect of these two factors.

Figure 2 shows that average citation frequency increases with the number of authors. Multiple authors on a paper necessarily share common research interests and typically follow each other' s work in subsequent research, multiplying citation likelihood among team members geometrically. Additionally, since each author may have their own research team, those team members may also pay attention to the paper, further increasing its citation frequency.

Figures 3 and 4 show that neither the number nor length of author keywords has a significant effect on average citation frequency. Reference [15] found that up to 90% of university students frequently use search engines for literature retrieval, while only about 37% have used the China Journal Full-text Database. This indicates that search engines are the preferred tool for students seeking references. Even though university students have free database access and many have taken information retrieval courses, they still primarily use search engines; those without professional database access have no alternative. When using search engines, most search terms are actually free terms or even natural language, and since most author keyword counts are small (typically around 5), the probability of matching multiple search terms simultaneously is very low, making the effect of author keywords on citation frequency insignificant.

Figure 7 shows a positive correlation between number of references and citation frequency. Generally, more references yield higher citation frequency. This is because more references require more time for searching, reading, and learning, leading to more comprehensive, accurate, and deeper understanding, more reliable conclusions, higher paper quality, and thus greater citation likelihood [6]. Simultaneously, since the paper is strongly related to its references, it attracts attention from the authors of those references. More references mean more attention and increased citation probability. Additionally, searching reference lists is an important method for finding related research.

Limitations

We selected the SCIE database from WOS as our data source both to minimize certain factors' influence and to obtain more meaningful data, as SCI papers and their citations receive widespread attention from Chinese academia and science/technology departments, and sampling was more convenient than other databases, though still labor-intensive. Our limited data sampling conditions may introduce limitations:

(1) SCIE papers represent high-level research, so citation patterns for lower-level papers remain unverified, making our conclusions primarily applica-

ble to high-level papers.

(2) Since data belong to WOS Engineering and Mechanical subjects rather than all subjects, the conclusions may not be fully applicable to other subject categories.

(3) SCIE is an English abstract database where even non-English original documents are represented in English. Language differences may cause variations in length counts for titles and abstracts. Additionally, occasional incomplete data in SCIE, such as papers titled "Untitle" or missing authors, affect calculations to some degree.

(4) The distribution of these parameters across data segments is uneven. For example, in 2010 there were 1,413 papers with title length of 11, while across all three years only 1,378 papers had titles exceeding 20 words—a distribution span accounting for about 40% of the full range (minimum title length 1, maximum 52) but representing less than 3% of total data. This uneven distribution causes differential susceptibility to random effects, with smaller quantities being more vulnerable, which is why we specifically conducted statistical analysis for cases with more than 10 papers. Nevertheless, this cannot completely eliminate effects from small sample sizes, which may account for anomalies at the extremes of Figures 1-7.

Conclusion

Many factors influence citation frequency. Among the parameters examined, paper length, abstract length, number of references, number of authors, and funding support exert positive correlation effects on citation frequency, while title length also affects citation frequency but with varying trends depending on length. Since our conclusions are based on Engineering and Mechanical subjects in the SCIE database, their generalizability to other papers requires further investigation.

References

[1] Liu Xueli, Xu Gangzhen, Fang Hongling, et al. How to Improve the Impact Factor of a Medical Journal: To Refer from the Classification Citation of Recent Advances in Ophthalmology[J]. Chinese Journal of Scientific and Technical Periodicals, 2008, 19(4): 659-661.

[2] Chen Shuxian. An Analysis of Affecting Factor of Journal' s Influencing Multiplier[J]. Journal of Shandong University of Technology: Social Sciences Edition, 2006, 22(4): 110-112.

[3] Huang Ping, Luo Yanqing, Wang Yan, et al. Analysis of Affecting Factor of Journal of Science and Technology and Improving Measurement[J]. Acta Editologica, 2006, 18(S1): 180-181.

[4] Chen Jiashun. An Analysis of the Non-learned Factors in "Affecting Factors" of Learned Journals[J]. Journal of Hubei Normal University: Philosophy and Social Sciences, 2005, 25(5): 133-135.

[5] Hu Jiansheng. Discussion on the Main Influence Factors of Impact Factor[J]. Journal of Anhui Agricultural Sciences, 2009, 37(13): cover2-cover3.

[6] Sun Shujun, Zhu Quan' e. Quality of Contents Determines the Total Cites of an Article[J]. Acta Editologica, 2010, 22(2): 141-143.

[7] Yu Liping, Pan Yuntao, Wu Yishan. Study on Influences to Academic Journal Impact Factor Based on Quantile Regression[J]. Library and Information Service, 2010, 54(16): 145-148.

[8] Xiao Hong, Yuan Fei, Wu Jianguo. Factors Affecting Citations: A Comparison Between Chinese and English Journals in Ecology[J]. Chinese Journal of Applied Ecology, 2009, 20(5): 1253-1262.

[9] Jian Lin, He Jing, Zhou Jian. Document Factors Impacting on the Citation of an Article: Multi-fields View[J]. Library and Information Service, 2011, 55(20): 32-35.

[10] Hudson J. Be Known by the Company You Keep: Citations—Quality or Chance?[J]. Scientometrics, 2007, 71(2): 231-238.

[11] Cai Zhicheng, He Limin. Application of Correlation Analysis Theory in Library and Information Analysis[J]. Modern Information, 2006, 26(5): 151-152.

[12] Liu Hua, Xu Guoyan. Analysis of Highly-cited Papers of Internet Subject Area Published in Domestic Journals from 2001 to 2010[J]. Sci-Tech Information Development & Economy, 2014, 24(15): 123-125.

[13] Zhao Jing, Jie Yali, Du Zhibo. Analysis of Differences in Citation Between Funded Papers and Non-funded Papers in Some High-level Medical Journals[J]. Acta Universitatis Medicinalis Nanjing: Social Science, 2012(6): 499-501.

[14] Yang Dongyan. On the Common Problems of Electronic Documents Retrieval for University Students[J]. Library Tribune, 2009, 29(5): 147-148.

[15] Search Engine Correlation Algorithm Analysis[EB/OL]. [2015-03-23]. http://wenku.baidu.com/link?url=z2nvKb9aBvywcMWYUs2efiqigbjyeJCkW-pwom9CMToOMar_{a0xqsiSLgH1ID6ySlCDPhKoGcbF7orn16uVV2ZMipjlTlE2TpjngfVT5gam}.

Author Contributions

Xiao Xuebin: Conceptualized the research, performed data sampling and analysis, drafted and revised the paper.
Chai Yanju: Designed correlation verification protocols, revised the paper.

Conflict of Interest Statement

All authors declare no conflict of interest.

Supporting Data

Supporting data are self-archived by the authors, E-mail: xxb@lib.whu.edu.cn. [1] Xiao Xuebin. Mechanical Engineering. xlsm. A Study on the Relationship Between Relevant Parameters of Papers and Citation Frequency. Data download: http://pan.baidu.com/s/1i4P4REt.

Figure Captions

Figure 1 [Figure 1: see original paper] Relationship between page count and citation frequency
Figure 2 [Figure 2: see original paper] Relationship between number of authors and citation frequency
Figure 3 [Figure 3: see original paper] Relationship between number of keywords and citation frequency
Figure 4 [Figure 4: see original paper] Relationship between keyword length and citation frequency
Figure 5 [Figure 5: see original paper] Relationship between title length and citation frequency
Figure 6 [Figure 6: see original paper] Relationship between abstract length and citation frequency
Figure 7 [Figure 7: see original paper] Relationship between number of references and citation frequency
Figure 8 [Figure 8: see original paper] Relationship between funding support and citation frequency
Figure 9 [Figure 9: see original paper] Overlapping scope of multiple independent concepts

Table 1 Correlation coefficients between parameters and average citation frequency

Parameter	r (subsample > 10)	r (all papers)
Page count	>0.92	0.649–0.822
Number of authors	>0.92	0.649–0.822
Number of references	>0.92	0.986
Title length ($ $8 words)	>0.92	0.935
Abstract length	>0.8	0.649–0.822
Number of keywords	<0.5	-
Keyword length	<0.5	-
Full title length	<0.5	-

Table 2 Effects of paper parameters on citation frequency

Parameter	Effect on Citation Frequency
Page count	Strong positive correlation
Number of authors	Strong positive correlation
Number of keywords	No relationship
Keyword length	No relationship
Title length	Uncertain effect
Title length ($ $8 words)	Strong positive correlation
Abstract length	Positive correlation
Number of references	Strong positive correlation
Funding support	Higher citation frequency

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