Quantitative understanding of knowledge application transition in CSS through citation behavior

Anonymous

Keywords: Citation behavior, Knowledge application transition, Citation speed, Science of science, Natural language processing

Extended Abstract

With the advent of large-scale academic datasets, researchers are quantitatively identifying the creative process of scientific knowledge. While the number of citations is a well-known indicator to estimate the impact of papers, citation behavior goes beyond a simple count of the number and has been studied to reveal how other studies utilized knowledge in respect of citer's location, polarity, and impact[1]. Nevertheless, there has been a lack of investigation into the temporal evolution of citation behavior to a paper. For example, word2vec[2] had a profound impact on computational text analysis in the early 2010s but has since been surpassed by more advanced models, leading to criticism that it is primitive by comparison.

This study aims to uncover how a paper's influence on future research changes over time by monitoring shifts in citation behavior, which we refer to as *knowledge application transitions.* To more clearly observe the timing of these transitions, we focus on slow-cited papers[3] that experience a sudden citation increase after several years. Through this approach, we clearly separate the time series of a paper's citation growth to uncover changes in citation behavior that occur when previously uncited papers begin to be cited. Our analysis provides practical implications for understanding how knowledge diffusion progresses within academia.

We started our study by selecting highly-cited papers from a Scopus dataset of 78 million papers, which ranked in the top 1% of the field-corrected 15-year citation counts published between 1970 and 2007. To identify slow-cited papers, we used Citation Delay[4] to measure citation speeds for each paper and selected the top 10% of the score. Since citation customs vary by research field, we narrowed our focus to computational social science(CSS) by using network clustering to classify papers at the publication level[5], resulting in 579 slow-cited papers related to CSS. For each paper, we determined the burst year (t_a) when the citation count spiked[6], and divided the citation period into pre-burst and post-burst intervals. The mean sleeping year of slow-cited papers was 8.1 years.

Afterward, we extracted 152,659 citation sentences for all slow-cited papers by consolidating 300 million paper data from S2ORC[7]. Our study focused on two citation aspects: 1) the citer's section, which indicates where prior knowledge contributes to new research, and 2) the citer's context, which reveals how target papers affect subsequent research. For the classification of citation behavior, we utilized Wang's hedge-cue-based method[8] to classify citation contexts into four categories (criticize, extend, improve, compare) through word matching and examined distinctive citation contexts before and after the citation burst.

As the main findings of our analysis, firstly, citer's sections remain consistent both before and after the citation burst, although slow-cited papers receive more citations after the burst year t_a (Figure 1). Compared to fast-highly-cited papers that reach their citation peak a few years after publication, slow-cited papers show no significant difference in citer's locations. The majority of citations are located in the Introduction (65%), followed by the Literature Review (15%), Results/Discussions (15%), and other sections. This suggests that slow-cited papers contribute to research in the conceptual stage, even after the increase in citations, and that the reason for the increase in citations is not due to a change in the utilization of research.

The second finding reveals that the citation context differs between burst and non-burst periods. Figure 2 presents the specialization coefficients for each context term, calculated by comparing the proportion of citing documents before and after the burst. For negative values, the hedge cue appears before the burst and vice versa. Specifically, we found that while slow-cited papers before the burst were mentioned about the benefits of improvement by existing theories, they were used after the burst with adversative expressions to illustrate the limits of the research. This result suggests that previously accepted research became recognized with its limitations as citation increased, and slow-cited papers were cited more due to the relativization of their findings by the emergence of other studies, which were discussed in the introduction or related work sections of articles.

Figure 3 shows an example of citing documents for a slow-cited paper, Adamic and Adar's 2003 paper on web page networks, before and after the burst. The paper's direct method to reveal network structure has been diversified with the rise of CSS and is now widely known. Before the burst, the citing documents compared the proposed index with other similar ones and deemed it effective for predicting network links. After the burst, the citing documents rather addressed its weaknesses and suggested improvements for further development in the next research.

The study suggests that the knowledge application transition in CSS involves positioning previous research compared to other studies, even if it was once considered an improvement. In other words, some studies perceived as mere extensions of existing models may have further potential when attention is focused on parts where different results are obtained from similar studies. Tracking knowledge application transition can extend the citation-count-based indicators such as novelty and disruptions. Limitations of this study include that the results are based solely on the field of CSS, and more advanced natural language processes over word matching are required. In addition, it is necessary to compare the results of this study with those of papers that experience an immediate increase in citations. Further study will provide valuable insights into modeling the process of disseminating scientific knowledge within academic communities.

References

[1] I. Tahamtan and L. Bornmann, “What do citation counts measure? An updated review of studies on citations in scientific documents published between 2006 and 2018,” *Scientometrics*, vol. 121, no. 3, pp. 1635–1684, 2019.

[2] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26, 2013.

[3] T. Miura, K. Asatani, and I. Sakata, “Revisiting the uniformity and inconsistency of slow-cited papers in science,” *Journal of Informetrics*, vol. 17, no. 1, p. 101378, 2023.

[4] J. Wang, B. Thijs, and W. Glänzel, “Interdisciplinarity and Impact: Distinct Effects of Variety, Balance, and Disparity,” *Plos One*, vol. 10, no. 5, p. e0127298, 2015.

[5] L. Waltman and N. J. Eck, “A new methodology for constructing a publication-level classification system of science,” *J Am Soc Inf Sci Tec*, vol. 63, no. 12, pp. 2378–2392, 2012.

[6] Q. Ke, E. Ferrara, F. Radicchi, and A. Flammini, “Defining and identifying Sleeping Beauties in science,” *Proc National Acad Sci*, vol. 112, no. 24, pp. 7426–7431, 2015.

[7] K. Lo, L. L. Wang, M. Neumann, R. Kinney, and D. Weld, “S2ORC: The Semantic Scholar Open Research Corpus,” *Proc 58th Annu Meet Assoc Comput Linguistics*, pp. 4969–4983, 2020.

[8] W. Wang, P. Villavicencio, and T. Watanabe, “Analysis of reference relationships among research papers, based on citation context,” *International Journal on Artificial Intelligence Tools*, vol. 21, no. 2, p. 1240004, 2012.

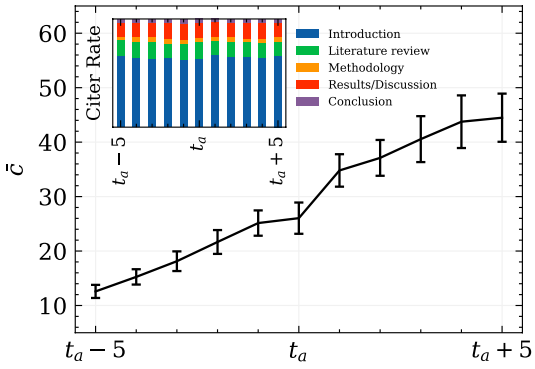

Figure 1. [Outer] Mean annual citations of slow-cited papers. Citation burst occurs at year t_a. [Inner] Located sections of citer sentences over time.

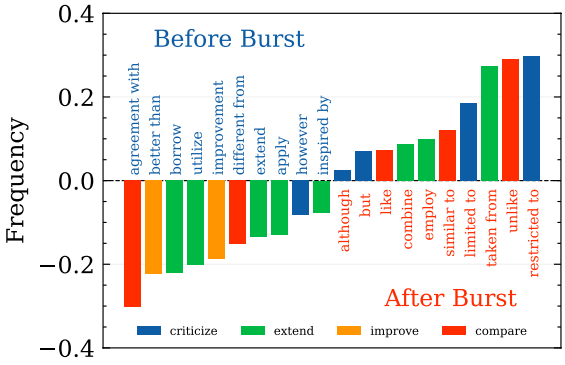

Figure 2. Top 10 frequent hedge cues before and after the citation bursts t_a. Each color represents the group of citer behavior on slow-cited papers.

They found that neighborhood similarity measures, such as the Jaccard BIBREF_14_ , **Adamic-Adar BIBREF_18_**, and the Katz coefficients BIBREF_19_ provided a large factor **improvement** over randomly predicted links.

It shows good results in predicting the friendship according to personal homepage and Wikipedia Collaboration Graph, **but** in the experiment of predicting author collaboration, it shows a poor accuracy prediction **BIBREF_15_**.

Figure 3. Example of citer sentences before the burst(top) and after the burst(bottom) on "Friends and neighbors on the Web (Adamic and Adar 2003)". Bold parts express the reference to the Adamic-Adar paper.