Scientometrics (2024) 129:4495–4512
https://doi.org/10.1007/s11192-024-05080-6

Closer in time and higher correlation: disclosing the relationship between citation similarity and citation interval

Wei Cheng[1] · **Dejun Zheng**[1] · **Shaoxiong Fu**[1] · **Jingfeng Cui**[1]

Received: 6 February 2024 / Accepted: 5 June 2024 / Published online: 20 June 2024

Abstract
Investigating the intricate relationship between citation similarity and the citation interval offers vital insights for refining citation recommendation systems and enhancing citation evaluation models. This is also a new perspective for understanding citation patterns. In this study, we used the Library and Information Science (LIS) field as an example to determine and discuss the correlation between citation similarity and the citation interval. Using the methods of data collection, paper title preprocessing, text vectorization based on simCSE, calculation of citation similarity and the citation interval, and calculation of the index per citing paper, this study found the following LIS domain-based results: (i) there is a significant negative correlation between citation similarity and the citation interval, but the correlation coefficient is low. (ii) The citation intervals of the least relevant series of cited papers exhibit a more pronounced susceptibility to citation similarity than the most relevant series of cited papers. (iii) The citation intervals of the most relevant cited papers are more concentrated within 12 years and more likely to be published within the average citation interval, typically from the newer half of the cited paper list and published later within 5 years of the citation half-life. This study concludes that researchers usually pay more attention to the latest and most cutting-edge and strongly relevant existing research than to weakly relevant existing research. Continuous attention and timely incorporation of knowledge into the research direction will promote a more rapid and specialized diffusion of knowledge. These findings are influenced by the accelerated dissemination of information via Internet, heightened academic competition, and the concentration of research endeavors in specialized disciplines. This study not only contributes to the scholarly discussion of citation analysis but also lays the foundation for future exploration and understanding of citation patterns.

Keywords Citation similarity · Citation interval · Average citation interval · Citation half-life · simCSE

✉ Dejun Zheng
zdejun@njau.edu.cn

[1] College of Information Management, Nanjing Agriculture University, Nanjing, China

Introduction

Scientific citations play a crucial role in knowledge dissemination, validation, evaluation, and discovery (Aksnes et al., 2019; Garfield & Merton, 1979; Jurgens et al., 2018; Synnestvedt et al., 2005). In addition, scientific citations act as connectors of ideas, ensuring the continuity of scholarly discussions and marking points of departure for new investigations. Explaining the pattern of scientific citations helps deepen the understanding of knowledge dissemination, influence, development dynamics, and resource allocation and provides important references and guidance for the further development of scientific research, which is an important research topic in scientometrics (Aistleitner et al., 2019; Chen, 2006; Cui et al., 2023; Slyder et al., 2011).

In the process of determining the patterns of scientific citations, citation similarity, which indicates the degree of relevance of the citing and cited papers, is an important variable used to observe and assess the depth of knowledge dissemination (Chen, 2017; Liu & Chen, 2021; Nassiri et al., 2013; Rodriguez-Prieto et al., 2019). The citation interval, which is the time interval between the publication year of the citing paper and the cited paper, is also an important variable for calculating and assessing the speed of knowledge dissemination (Bornmann et al., 2018; Marx et al., 2014; Thor et al., 2016). However, no study has explored the relationship between citation similarity and the citation interval in-depth, and exploring the relationship between the two can assist in citation recommendations and optimizing the citation evaluation model.

Therefore, the purpose of this study is to determine the relationship between citation similarity and citation intervals. The theoretical foundation for disclosing the relationship between citation similarity and the citation interval is derived from the assumption that greater citation similarity indicates shorter citation intervals. This assumption is rooted in the idea that when researchers encounter a paper highly similar to their own, they are more likely to cite it promptly, resulting in a shorter citation interval. In contrast, papers with less similarity may take longer to be discovered or recognized by researchers, leading to longer citation intervals.

This study reveals the relationship between citation similarity and citation intervals from three progressive analytical perspectives. First, we aim to reveal whether there is a significant correlation between citation similarity and citation interval. Second, we focus on a single citing paper and discuss the correlation between citation similarity and the citation interval at a more detailed level by comparing observations of the most or least relevant cited papers. Third, the average citation interval (the mean citation interval for a paper to cite its total cited papers) and citation half-life (the time it takes for a paper to cite half of its total cited papers) are used as references to investigate differences in citation behavior on the citation interval between single citing papers and explore the relationship between citation similarity and the citation interval at a more specific level by observing multiple pairs of comparison samples. Therefore, we proposed the following research questions:

RQ1: Is there a significant correlation between citation similarity and the citation interval between citing papers and cited papers?

RQ2: Is there a significant correlation between the most or least relevant series of cited papers and the citation interval of citing papers? If this is the case, is there a difference in the degree of correlation between the two.

RQ3: Introducing the average citation interval and citation half-life as references, is there a difference in the citation interval between a citing paper and the most or least relevant series of cited papers? If so, how can this difference be characterized?

In this study, we used English journal papers in the field of Library and Information Science (LIS) as an example for empirical analysis. The preprocessed title representing the paper was used as the text input, whereas the semantic similarity between the cited paper and cited paper was calculated based on simCSE. The publication year interval between the citing and cited papers was calculated. Correlation, regression, and statistical and comparative analyses were performed to determine the relationship between citation similarity and the citation interval. This study theoretically contributes to the body of knowledge as it provides a new perspective on citation analysis by observing the relationship between citation similarity and the citation interval. The practical contribution of this study is to provide empirical data support in the field of LIS for related studies by finding a significant negative correlation between citation similarity and the citation interval.

Background

The measurement of similarity is usually discussed in studies related to recommendation algorithms, such as citation recommendations and research paper recommendations (Ali et al., 2022; Beel et al., 2016; Sharma et al., 2023; Zhang & Zhu, 2022). Furthermore, citation similarity is measured in a variety of ways, and the mainstream methods can be categorized as follows.

First, co-citation network analysis evaluates the frequency with which two papers are cited together, unveiling thematic parallels and intellectual associations (Rodriguez-Prieto et al., 2019; Tantanasiriwong & Haruechaiyasak, 2014). Other studies have used citation networks to calculate citation similarity (Pagani et al., 2015; Pornprasit et al., 2022; West et al., 2016). However, the measurement of citation similarity based on citation networks only considers the direct citation and citation counts between papers and does not consider the semantic information between the citing and cited papers. Second, word vectors are constructed and thus similarity is computed based on traditional statistical learning methods, such as word vectors based on traditional statistical learning methods, which, in turn, calculate text similarity, e.g., TF–IDF (Tata & Patel, 2007), co-occurrence matrix (Rohde et al., 2006), LSA (Niraula et al., 2013), and LDA (Niraula et al., 2013). Although the above method calculates the similarity based on the text content at the word level, it ignores the semantic information of the citing and cited papers. Third, with the development of word-vector embedding models and continuous optimization of their ability to represent text semantics (Ethayarajh, 2019; Jatnika et al., 2019), the semantic similarity analysis using natural language processing techniques has been widely used to assess the similarity between citing and cited papers (Buscaldi et al., 2024; Lu et al., 2023). Semantic similarity calculations can reflect the content relevance between citing and cited papers more realistically. Although semantic similarity calculation is more complex and requires considerable computing time and resources, the wide application of GPUs can cope handle this problem well. Therefore, this study adopted a semantic similarity calculation method.

Text vectorization is the transformation of text into a suitable vector representation, which is the preparation step for citation semantic similarity calculation. From TF–IDF to Doc2Vec and from ELMo to BERT, scholars have explored methods to represent text more accurately and completely, map it into high-dimensional vectors, and improve the quality of the spatial identification of textual semantics. BERT can represent the semantic information of the text better than TF–IDF and Word2Vec; however, the degree of differentiation is small, and most sentence computation similarities are approximately 0.9. To address this

issue, Similarity-based Contrastive Learning for Textual Sentence Embeddings (simCSE) makes similar texts closer in the multidimensional semantic space by contrastive learning and can learn more semantically distinguishable sentence representations compared to BERT, TF–IDF, and Word2Vec. Therefore, this study used simCSE for text vectorization (Gao et al., 2021; Wu et al., 2021).

Citation similarity is widely discussed in innovation calculation and impact measurement research (Bornmann et al., 2019; Su et al., 2021; Zhang et al., 2021). However, the potential relationship between citation similarity and the citation interval has not been deeply analyzed or explored. Therefore, based on the calculation of the two variables, citation similarity and the citation interval, this study used correlation, regression, statistical, and comparative analyses to disclose the relationship between the two variables. This can provide a new quantitative analytical idea for the characterization of field citation knowledge dissemination and inspire research related to citation analysis.

Data and methodology

This study focused on citing and their cited papers to rationally calculate citation similarity and explores the relationship between citation similarity and the citation interval. The overview of the key steps for this study, shown in Fig. 1, comprises five steps: data collection, paper title preprocessing, text vectorization based on simCSE, calculation of citation similarity and the citation interval, and calculation of the index per citing paper.

Data collection

Semantic Scholar was launched in 2015. It was developed by the Allen Institute for Artificial Intelligence, founded by Microsoft co-founder Paul Allen. It supports open access to citation information for papers. Library and information science (LIS) extensively absorbs theoretical knowledge and methods from related disciplines, including numerous subdivided research directions, and it is characterized by knowledge cross-fertilization (Rubin & Rubin, 2020). Therefore, cited papers of a citing paper in this field usually have greater differences, which can be used to distinguish between citation similarities. Furthermore, it facilitates unified processing and comparative analysis of data and excludes the influence of extreme samples (Dixon, 1950; Hwa, 2004). We obtained English journal articles from Semantic Scholar of LIS discipline categories in the Web of Science and excluded papers with fewer than six references and citations where the cited paper was published earlier than 1800. Data were collected on December 8, 2023, and 64,465 papers and 2,966,038 citations were obtained. Figure 2 shows the details of this dataset. Before 1990, the number of papers in the dataset was small. In some of these years, the number of papers was zero; therefore, the average number of citations per citing paper takes the value of zero in these years as well. Meanwhile, the change in the average number of citations per citing paper was affected by a small number of samples and the change is not regular. Furthermore, from 2005 onward, the number of papers began to increase considerably, along with an overall upward trend in the average number of citations per paper. The dataset is of a certain size, and we used this data as a proxy to explore the relationship between citation similarity and the citation interval.

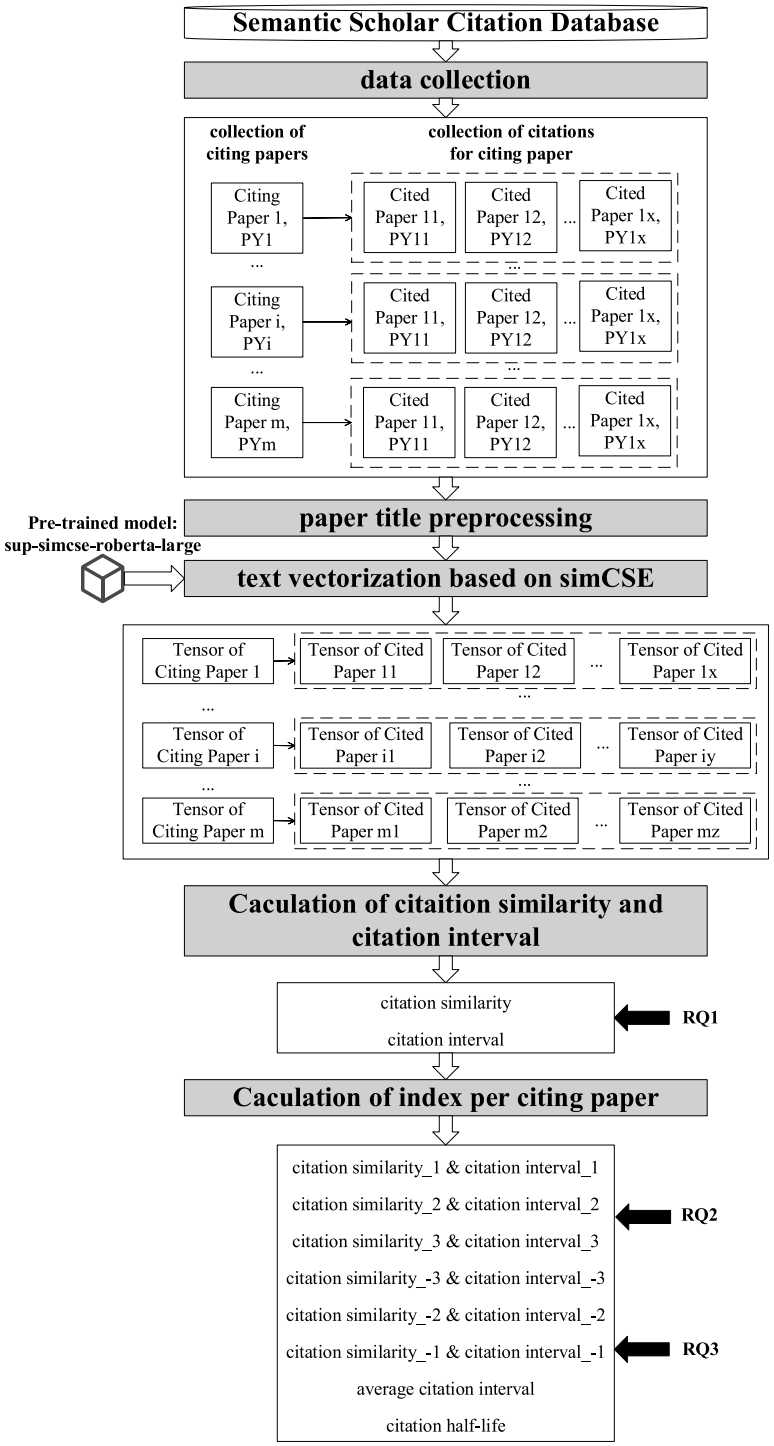


Fig. 1 Overview of the key steps of this study

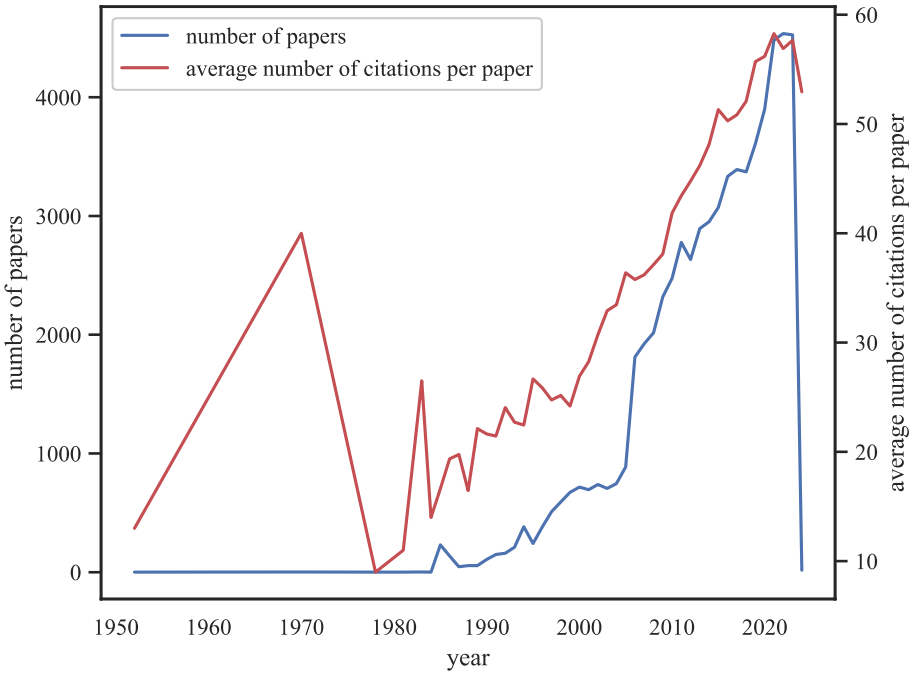


Fig. 2 Details of the dataset

Paper title preprocessing

The title is a brief text that reveals the core content of the paper and, to some extent, reflects the research question, methodology, and topic. Therefore, it can be used as a textual basis for observing the relevance of citing and cited papers. To compare the correlations between citing and cited paper titles more easily, the title text must be preprocessed. First, the titles need to be segmented and all the words converted to lowercase; second, the stopwords in the titles are removed using the stopwords dict of NLTK; finally, NLTK is used to lexically annotate the words, and lexical morphology reduction of the words is achieved by lexical annotation. For example, after preprocessing, "*Information transfer and cognitive mismatch: a Popperian model for studies of public understanding*" will be processed as "*information transfer cognitive mismatch popperian model study public understand.*"

Text vectorization based on simCSE

simCSE is based on the concept of contrastive learning to fully acquire the semantic knowledge of a text. simCSE uses pre-trained word vectors as the initial representations of sentences and constructs contrast samples using data augmentation and negative sample sampling, generating positive and negative sample representations of the input text. It brings similar texts closer to the multidimensional semantic space by minimizing the semantic distance between the positive and negative samples (Gao et al., 2021; Wu et al., 2021).

To obtain better text vectorization results, the pre-training model "sup-simcse-roberta-large" is chosen, compared with "sup-simcse-bert-base-uncased." It uses more data for pre-training and adopts a series of optimized training strategies, which can better capture the statistical features of the language and construct a more comprehensive and accurate semantic vector representation of the text. By inputting preprocessed titles, the simCSE model outputs a 1024-dimensional vector representation stored in tensor format. We iterated through 64,465 citing papers and their corresponding cited papers to vectorize all titles.

Calculation of citation similarity and the citation interval

We calculated citation similarity and the citation interval and used a correlation analysis to answer RQ1. Cosine similarity is one of the most commonly used vector similarity algorithms and was used in our calculations. The cosine similarity based on the tensor between citing and cited papers and the citation interval between citing and cited papers are as follows:

$$\text{citation similarity}\left(\text{Paper}_i, \text{Paper}_{ij}\right) = \frac{\sum_{k=0}^{1023}\left(t_{ik} \times t_{ijk}\right)}{\sqrt{\sum_{k=0}^{1023} {t_{ik}}^2} \times \sqrt{\sum_{k=0}^{1023} {t_{ijk}}^2}},$$

$$\text{citation interval}(\text{Paper}_i, \text{Paper}_{ij}) = \text{PY}_i - \text{PY}_{ij} + 1.$$

Paper_i and Paper_{ij} are the ith ($64{,}465 \geq i \geq 1$) citing paper and the jth cited paper of Paper_i in the dataset, respectively. t_{ik} and t_{ijk} are the kth ($1023 \geq k \geq 0$) element of Tensor of Paper_i and Tensor of Paper_{ij}, respectively. PY_i and PY_{ij} are the publication years of Paper_i and Paper_{ij}, respectively.

Because of the network debut, the citation interval(Paper_i, Paper_{ij}) may be less than 0. For the purposes of statistics and analysis, when the citation interval(Paper_i, Paper_{ij}) was less than 0, it was assigned as 0.

Calculation of index per citing paper

We calculated the index per citing paper and used correlation and comparative analyses to answer RQ2 and RQ3. For any paper in the 64,465 citing papers, if it has r cited paper, it contains r (citation similarity, citation interval) binary records, and the r records are arranged in descending order according to the size of the citation similarity value to form a temporary list L. We defined (citation similarity_1, citation interval_1), (citation similarity_2, citation interval_2), and (citation similarity_3, citation interval_3) as the first, second, and third binaries (indexes based on the three cited papers most relevant to the citing paper) of the positive order of list L, respectively. Furthermore, we defined (citation similarity_-1, citation interval_-1), (citation similarity_-2, citation interval_-2), and (citation similarity_-3, citation interval_-3) as the first, second, and third binaries (indices based on the three cited papers least relevant to the citing paper) in reverse order of list L.

The relationship between citation similarity and the citation interval was analyzed based on the average citation interval and citation half-life. For any paper, if it has r cited paper, it contains r (citation similarity, citation interval) binary records, and the r records are arranged in descending order according to the size of the value of the citation interval to

form a temporary list *L2*, the calculation of the average citation interval and citation half-life is as follows.

$$\text{average citation interval}_i = \frac{\sum_{k=1}^{r} \text{citation interval}_{ik}}{r},$$

$$\text{citation half-life}_i = \begin{cases} \text{citation interval}_{i\frac{r+1}{2}}, & \text{if } r \div 2 \neq 0, \\ \frac{\text{citation interval}_{i\frac{r}{2}} + \text{citation interval}_{i\frac{r}{2}+1}}{2} & \text{if } r \div 2 = 0. \end{cases}$$

i represents ith ($64{,}465 \geq i \geq 1$) citing paper in the dataset, r represents the number of binary records in the temporary list *L2*.

Results

Correlation analysis based on per citation

For the 2,966,038 (citation similarity and citation interval) binary records, the descriptive statistics and Spearman's correlations of the two variables are listed in Table 1. Table 1 shows that in the field of LIS, the citation similarity between the citing paper and cited paper is mostly at the medium–low level; most of the distribution is between 0.2 and 0.6, the citation interval is approximately 11 years, and the distribution varies considerably, with a wide range of values from 0 to 23 years. There was a significant negative correlation between citation similarity and the citation interval, but the correlation coefficient was low at − 0.17. This implies that, to some extent, as the citation similarity between the citing and cited papers increases, the citation interval decreases; that is, the citing paper may be more inclined to cite the more recent and relevant cited paper.

For **RQ1**, we found a significant negative correlation between citation similarity and citation interval between citing papers and cited papers.

Correlation analysis based on per citing paper

For the 64,465 citing paper, Spearman's correlations of citation similarity_1 and citation interval_1, citation similarity_2, citation interval_2, etc., are listed in Table 2. Table 2 shows significant negative correlations between all six pairs of variables; however, the correlation coefficients were low (less than 0.15). The correlation coefficients of the three pairs of variables based on the least relevant cited paper (citation similarity_ − 1 and citation interval_ − 1, et al.) were greater than those based on the most relevant cited paper (citation similarity_1 and citation interval_1, et al.) as a whole. Meanwhile, the higher the

Table 1 Descriptive statistics and Spearman's correlations of variables ($n = 2{,}966{,}038$)

Variables	Mean	SD	Min	Max	1	2
Citation similarity	0.394371206	0.166834259	− 0.227127537	1.0	1	
Citation interval	11.57	11.107	0	201	− .170**	1

**Correlation is significant at the 0.01 level (two-tailed)

Table 2 Spearman's correlations of variables ($n = 64{,}465$)

Variable	Variable	Spearman's correlation
citation similarity_1	citation interval_1	−0.131**
citation similarity_2	citation interval_2	−0.073**
citation similarity_3	citation interval_3	−0.065**
citation similarity_−3	citation interval_−3	−0.149**
citation similarity_−2	citation interval_−2	−0.139**
citation similarity_−1	citation interval_−1	−0.125**

**Correlation is significant at the 0.01 level (2-tailed)

citation similarity, the higher the negative correlation between citation similarity and citation interval for the three pairs of variables based on the most relevant cited paper and the least relevant cited paper.

To show the relationship between the six pairs of variables more intuitively, regression analysis was carried out on each of the six pairs of variables, and on the basis of the scatter plot of the distribution of the values of the six pairs of variables, the linear regression line and nonlinear regression line with marginal distributions were supplemented, as shown in Fig. 3. Figure 3 illustrates that the scatter distributions between individual subplots have more significant differences, as reflected between the three pairs of variables based on the most relevant cited paper and the three pairs of variables based on the least relevant cited paper. After calculation, although the linear and nonlinear regression results for each pair of variables were significant, the Standardized Error of the Estimate was high, and the linear regression line had a large degree of overlap with the nonlinear regression line in each subplot. This is because the scatter distributions within each subplot are more discrete, making it difficult to accurately fit all the scatters with a regression line. Nonetheless, the regression line reveals the interdependence and evolutionary trends of pairs of variables to a certain extent. In the six subplots, the citation interval decreased as citation similarity increased, and the trend of the three subplots on the right side was stronger than that on the left side. This is because, overall, the distribution range of the citation interval of the three subplots on the left side was smaller than that of the three subplots on the right side, and the trend of the changes in the citation interval in the three subplots on the right side was stronger when the citation similarity increased.

For **RQ2**, our research found a significant correlation between the most or least relevant series of cited papers and the citation interval with citing papers. The citation intervals of the least relevant series of cited papers are more strongly influenced by citation similarity than those of the most relevant series of cited papers, and their citation intervals tend to decrease more significantly when citation similarity increases.

Comparative analysis based on per citing paper

In "Correlation analysis based on per citing paper" section, we preliminarily analyzed the differences between the most and least relevant series of cited papers. To analyze their differences in more depth, especially the characteristics of the differences in citation intervals, we comparatively analyze them considering the following three aspects.

First, for the six index citing intervals, interval_1, citing interval_2, citing interval_3, citing interval_−3, citing interval_−2, and citing interval_−1, the number of citing

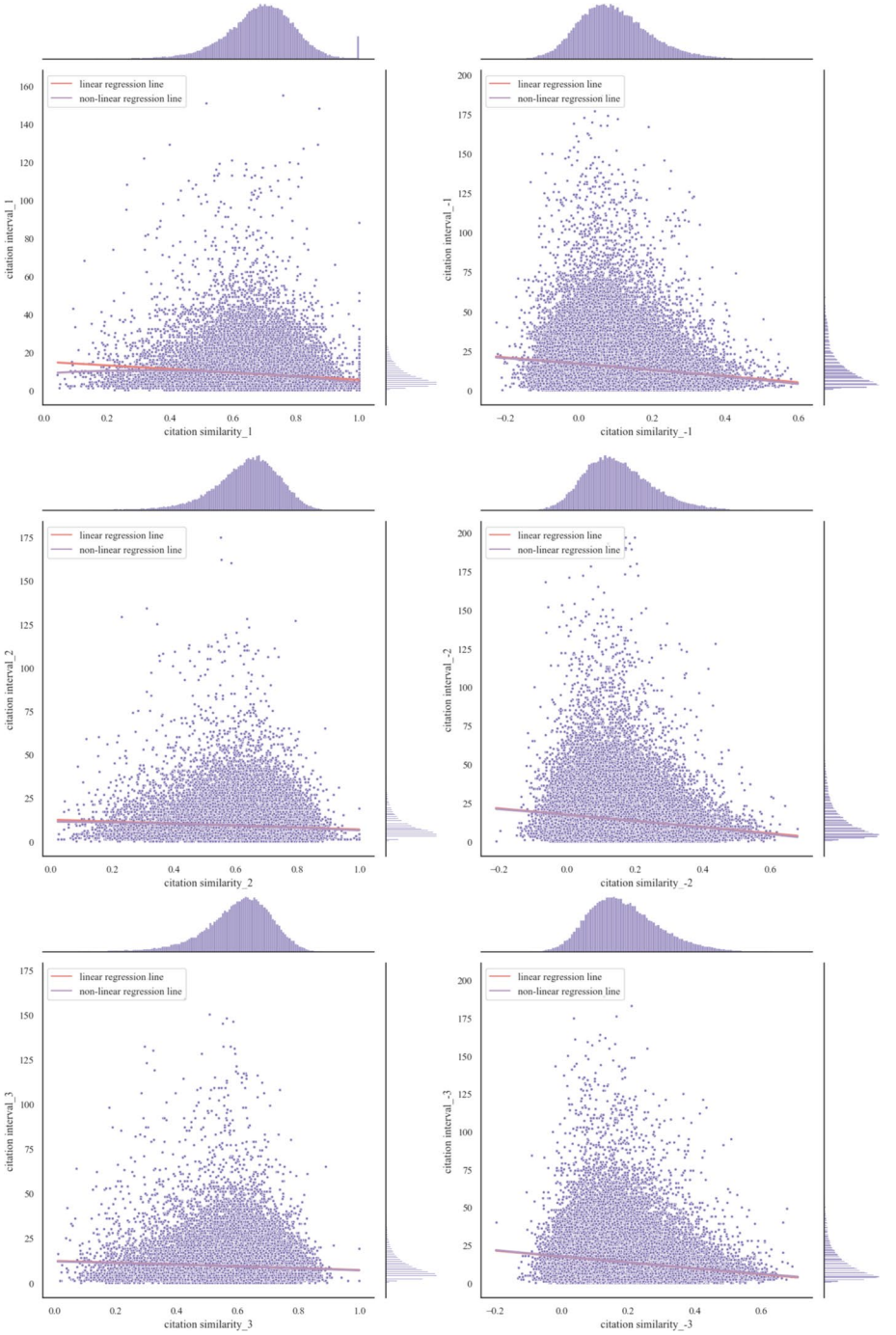


Fig. 3 Citation similarity and the citation interval based on per citing paper

papers at different citation intervals is shown in Fig. 4, and the intervals are limited to 0–50 years for the sake of presentation. Figure 4 illustrates that all six indices correspond to the most citing papers within 3 to 6 years, whereas citation interval_1, citation interval_2, and citation interval_3 are closer to each other in terms of peaks and are obviously significantly ahead of citation interval_ − 3, citation interval_ − 2, and citation interval_ − 1. In addition, the fold change trends of the first three indices converge, as well as of the last three indices; however, there is a clear difference between the fold change trends of the first three and the last three indices. Within approximately 12 years, the first three indices correspond to significantly more citing papers than the last three indices; however, after approximately 12 years, the last three indices correspond to significantly more citing papers than the first three indices. This demonstrates that in the field of LIS, the citation intervals of the most relevant series of cited papers are more centrally distributed within 12 years, whereas the citation intervals of the least relevant series of cited papers are also more distributed within 12 years but are relatively less centrally distributed.

Second, by comparing the average citation interval and citation half-life, the relationship between six indices, such as citation interval_1, and the two variables is shown in Fig. 5a, b. For comparison, the average citation interval and citation half-life are rounded to integers when the values are floating-point numbers. Figure 5a illustrates that in the field of LIS, the vast majority of citing papers have the citation intervals of the three most relevant cited papers within the average citation interval. Although more than half of the citing papers also had citation intervals of the three least relevant cited papers that were less than the average citation interval, they did not have a significant advantage in terms of the number. This suggests that more relevant cited papers are more likely to be published within the average citation interval. Figure 5b illustrates that the majority of citing papers have

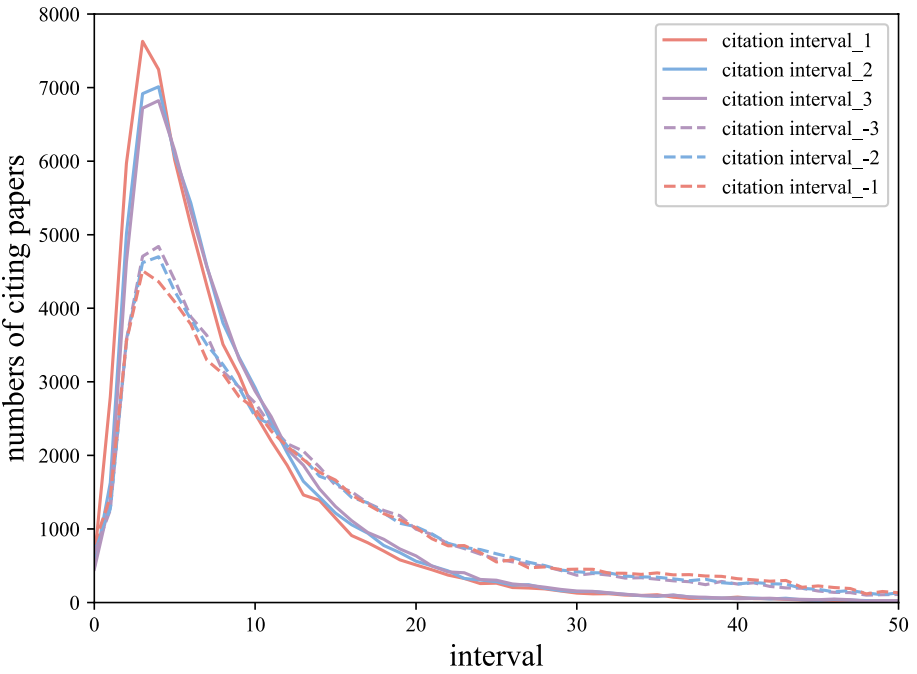
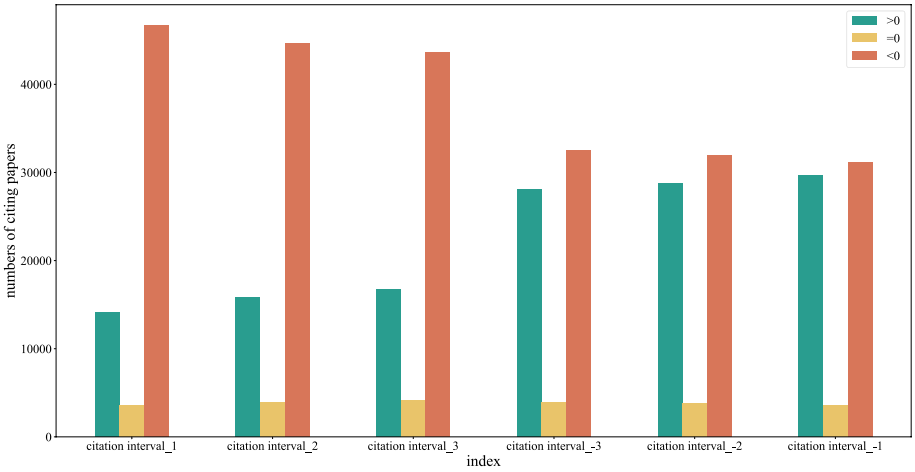
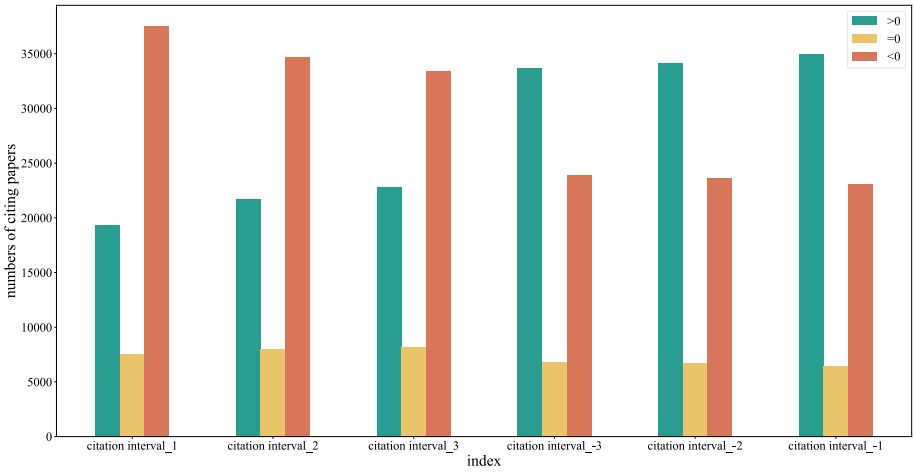


Fig. 4 Number of citing papers at different citation intervals



(a) Frequency of *index* greater than, equal to, or less than the *average citation interval*

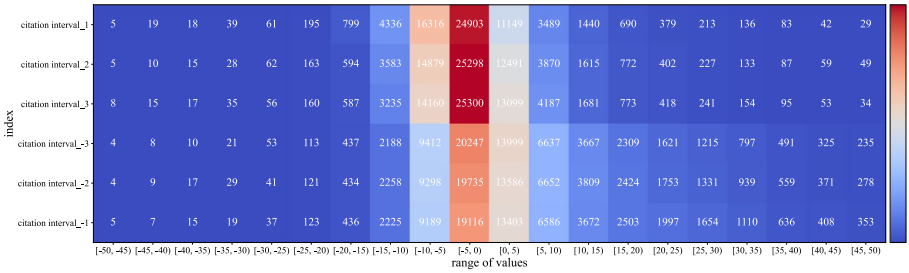


(b) Frequency of *index* greater than, equal to, or less than *citation half-life*

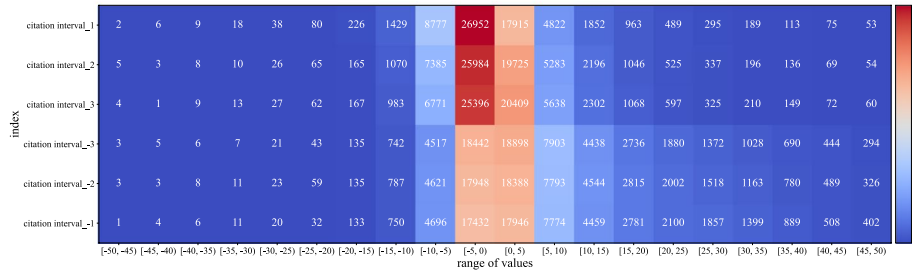
Fig. 5 **a** Frequency of *index* greater than, equal to, or less than the *average citation interval.* **b** Frequency of *index* greater than, equal to, or less than *citation half-life*

the three most relevant cited papers with citation intervals less than the citation half-life, whereas just under half of the citing papers have the three least relevant cited papers with citation intervals less than the citation half-life. This suggests that the three most relevant cited papers are more likely to be the newer half of the cited paper list, whereas the three least relevant cited papers are more likely to be the older half of the cited paper list.

Third, to further compare the differences in the values of the six indices, such as citation interval_1, with the average citation interval and citation half-life, the frequency distribution of the differences in the values of the six indices with the average citation interval and citation half-life are shown in Fig. 6a, b, using 5 years as the time slice. For comparison, the average citation interval and citation half-life are rounded to integers when the values are floating-point numbers. The value interval is defined as



(a) Frequency of range of values (*index* - *average citation interval*) distribution



(b) Frequency of range of values (*index* - *citation half-life*) distribution

Fig. 6 **a** Frequency of range of values (*index–average citation interval*) distribution. **b** Frequency of range of values (*index–citation half-life*) distribution

[− 50, 50). Figure 6a illustrates that the differences in the values of the six indices with respect to the average citation interval are overwhelmingly distributed within the [− 15, 15) interval. The three most or least relevant cited papers were most likely to have been published within 5 or 5 years of an average citation interval. However, the three most relevant cited papers were the second most likely to have been published within 5–10 years of the average citation interval, and the three least relevant cited papers were the second most likely to have been published for more than 5 years of the average citation interval. This suggests that in the field of LIS, the three most relevant cited papers are more likely to be published within less than 10 years of the average citation interval, whereas the three least relevant cited papers are more likely to be published within 5 years of the average citation interval. Figure 6b illustrates that the differences in the values of the six indices with respect to citation half-life were overwhelmingly distributed within the [− 10, 15) interval. The three most relevant cited papers were all the most likely to have been published within 5 or 5 years of less than the citation half-life, and the second most likely to have been published within 5 years of greater than the citation half-life. However, the three least relevant cited papers were all most likely to be published within 5 years of the citation half-life and the second most likely to be published within 5 years of the citation half-life or 5 years. This suggests that in the field of LIS, the three most or least relevant cited papers are all more likely to have been published within 5 years of the citation half-life, but the three most relevant cited papers are more likely to have been published within 5 years later than the citation half-life, whereas the three least relevant cited papers are more likely to have been published within 5 years before the citation half-life.

For **RQ3**, our research found that the more relevant cited papers were more likely to be published within the average citation interval, whereas the three most relevant cited papers were more likely to be in the newer half of the cited paper list.

Discussion

The findings of this study, which reveal a compelling correlation between citation similarity and the citation interval, offer profound insights into the dynamics of scientific progress and the mechanics of knowledge diffusion. This study demonstrates that researchers usually pay more attention to the latest and most cutting-edge and strongly relevant existing research than to weakly relevant existing research. Corresponding to the three research questions posed in the introduction, the findings are categorized into the following three aspects.

First, there is a significant negative correlation between citation similarity and the citation interval; however, the correlation coefficient is low because of the variability of the sample and the complexity of citation behavior. It is widely assumed that new ideas rely on the optimization, extension, and critique of existing research (Fleming, 2001; Kuhn, 1970), therefore, citing more cited papers with a newer publication year suggests that the citing paper incorporates knowledge from newer outputs and is more likely to be innovative. Some studies have found that more innovative and influential citing papers pay more attention to new and recently cited papers and incorporate and absorb new knowledge in their research timely (Liang et al., 2020; Petruzzelli et al., 2018; Sheng et al., 2023), which is consistent with the findings of this study.

Second, the citation intervals of the least relevant series of cited papers exhibited a more pronounced susceptibility to citation similarity than those of the most relevant series of cited papers, with a more marked decrease in their citation intervals observed with increasing citation similarity. With the standardization, accessibility, and maturity of scientific research, many researchers are aware of the latest advances in cutting-edge research and refer to it in time to carry out their own research (Kammari, 2023; Kim et al., 2018; Zhou et al., 2023). Consequently, cited papers more relevant to a citing paper are more likely to be concentrated in recent years, and this concentration lowers the correlation coefficients between citation similarity and citation interval. However, it still reveals, to some extent, that more relevant citations tend to be more recent.

Third, the citation intervals of the most relevant cited papers are more concentrated within 12 years, whereas the least relevant papers are spread over a wider range. More relevant articles are more likely to be published within the average citation interval. The most relevant ones are typically from the newer half of the cited paper list, whereas the least relevant ones are more likely to be from the older half. Both the most and least relevant ones are commonly published within 5 years of the citation half-life; however, the most relevant ones are often published later, whereas the least relevant ones tend to be published earlier. Using the average citation interval and citation half-life as the average rate of knowledge updates for cited papers, the more relevant papers cited on a higher than average rate may be due to more long-term and sustained attention from researchers. This would shorten the knowledge update cycle for more refined research directions in the LIS field and contribute to its rapid development. The findings of this study support the argument that knowledge in the LIS field diffuses faster and becomes more specialized (Ding et al., 2023; Järvelin et al., 2023).

The above findings prompted us to explore various aspects of the real-world reasons for the idea of "closer in time, higher correlation." One explanation is that the Internet has dramatically shortened the time lag between the conception of an idea and its dissemination, fostering an environment in which ground breaking research swiftly becomes part of the collective intellectual discourse (Zhang & Hou, 2023). Consequently, researchers are more inclined to engage with and cite cutting-edge studies that closely align with their ongoing pursuits, thereby reinforcing the contemporaneity of citations. Another contributory factor could be intensifying competition within academic disciplines, driving scholars to build promptly upon the most recent advancements to maintain relevance and secure funding (Yang, 2024). In this context, our findings may reflect a strategic approach to citation practice in which authors deliberately emphasize the currency and immediacy of their work by referencing the most up-to-date literature. Furthermore, these findings may indicate the formation and consolidation of research clusters or communities. As specific areas of inquiry gain momentum, they attract concentrated research efforts, resulting in a proliferation of studies that echo each other in terms of content and methodology (Smith et al., 2021). This clustering effect not only accelerates the pace of knowledge accumulation within these domains but also reinforces the perception of these fields as vibrant and rapidly evolving. In summary, the intricate relationship disclosed between citation similarity and the citation interval underscores the intricate interplay of technological advancements, competitive pressures, and collaborative dynamics in shaping the evolution of scientific knowledge.

Conclusion

This study considered 64,465 papers and 2,966,038 citations in LIS as experimental data and vectorized the titles of citing and cited papers based on simCSE to calculate semantic similarity. Based on the calculation of citation similarity and citation intervals, this study employed correlation, statistical, regression, and comparative analyses to elucidate the relationships between the two variables. This study provides compelling evidence of a significant, albeit nuanced, link between citation similarity and the citation interval, thus illuminating the dynamics of scientific progress and knowledge dissemination. We demonstrate that researchers prioritize citing recent, highly relevant works over less relevant ones, reinforcing the idea that innovative research builds on and extends contemporary scholarship. This also signifies that novel research capitalizes on new insights, thereby validating the importance of contemporaneous knowledge in fostering innovation. Meanwhile, the most relevant citations cluster within a 12-year window, underscoring a quicker knowledge update cycle. This concentration of relevant citations within the most recent half of the cited works supports the rapid development and specialization of the field. Furthermore, these findings are influenced by the Internet's accelerated dissemination of information, heightened academic competition, and the concentration of research endeavors in specialized disciplines.

This study contributes to the body of knowledge by enriching the research on citation patterns as it revealed the association between citation similarity and citation intervals at a deep level. Although we could not assert a causal relationship between the two variables, our findings can provide a new research perspective on citation motivation or behavior and data and argument support at the correlation level. In addition, it contributes to understanding the evolution of scientific knowledge. This underscores the interconnectedness of research endeavors and the immediacy with which new ideas can permeate through a

discipline. This reinforces the concept of knowledge waves or research fronts, where concentrated bursts of activity around specific topics propel the field forward in leaps.

This study has some limitations. Although these findings offer valuable insights into citation patterns and knowledge diffusion dynamics within LIS, the generalizability of these results to other academic disciplines remains unclear. Different fields may exhibit distinct citation behaviors, influenced by unique research cultures, publication cycles, and knowledge dissemination mechanisms. Reliance on a single dataset may also introduce dataset-specific biases or peculiarities that can influence the results. Future studies with broader datasets encompassing multiple disciplines would allow for more robust tests of the hypotheses and could reveal additional complexities or nuances in the relationships under investigation. In addition, we did not consider the effects of other relevant variables such as citation count, citation strength, and citation location. A more comprehensive analysis should be conducted to better understand the relationship between citation similarity and citation intervals.

Declarations

Conflict of interest No conflicts of interest exist in the submission of this manuscript, and the manuscript has been approved for publication by all authors. We declare that the work described here is original research that has not been published previously and is not under consideration for publication elsewhere, in whole or in part. All the authors listed have approved the enclosed manuscript.

References

Aistleitner, M., Kapeller, J., & Steinerberger, S. (2019). Citation patterns in economics and beyond. *Science in Context, 32*(4), 361–380.

Aksnes, D. W., Langfeldt, L., & Wouters, P. (2019). Citations, citation indicators, and research quality: An overview of basic concepts and theories. *Sage Open*, 9(1). https://doi.org/10.1177/2158244019829575

Ali, Z., Qi, G., Kefalas, P., Khusro, S., Khan, I., & Muhammad, K. (2022). SPR-SMN: Scientific paper recommendation employing SPECTER with memory network. *Scientometrics, 127*(11), 6763–6785.

Beel, J., Gipp, B., Langer, S., & Breitinger, C. (2016). Paper recommender systems: A literature survey. *International Journal on Digital Libraries, 17*, 305–338.

Bornmann, L., Haunschild, R., & Leydesdorff, L. (2018). Reference publication year spectroscopy (RPYS) of Eugene Garfield's publications. *Scientometrics, 114*, 439–448.

Bornmann, L., Tekles, A., Zhang, H. H., & Fred, Y. Y. (2019). Do we measure novelty when we analyze unusual combinations of cited references? A validation study of bibliometric novelty indicators based on F1000Prime data. *Journal of Informetrics, 13*(4), 100979.

Buscaldi, D., Dessí, D., Motta, E., Murgia, M., Osborne, F., & Recupero, D. R. (2024). Citation prediction by leveraging transformers and natural language processing heuristics. *Information Processing and Management, 61*(1), 103583.

Chen, C. (2006). CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society for Information, Science and Technology, 57*(3), 359–377.

Chen, L. (2017). Do patent citations indicate knowledge linkage? The evidence from text similarities between patents and their citations. *Journal of Informetrics, 11*(1), 63–79.

Cui, Y., Wang, Y., Liu, X., Wang, X., & Zhang, X. (2023). Multidimensional scholarly citations: Characterizing and understanding scholars' citation behaviors. *Journal of the Association for Information Science and Technology, 74*(1), 115–127.

Ding, J., Liu, C., & Yuan, Y. (2023). The characteristics of knowledge diffusion of library and information science—From the perspective of citation. *Library Hi Tech, 41*(4), 1099–1118.

Dixon, W. J. (1950). Analysis of extreme values. *The Annals of Mathematical Statistics, 21*(4), 488–506.

Ethayarajh, K. (2019). How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. arXiv preprint arXiv:1909.00512

Fleming, L. (2001). Recombinant uncertainty in technological search. *Management Science, 47*(1), 117–132.
Gao, T., Yao, X., & Chen, D. (2021). SimCSE: Simple contrastive learning of sentence embeddings. arXiv preprint arXiv:2104.08821
Garfield, E., & Merton, R. K. (1979). *Citation indexing: Its theory and application in science, technology, and humanities* (Vol. 8). Wiley.
Hwa, R. (2004). Sample selection for statistical parsing. *Computational Linguistics, 30*(3), 253–276.
Jatnika, D., Bijaksana, M. A., & Suryani, A. A. (2019). Word2Vec model analysis for semantic similarities in English words. *Procedia Computer Science, 157*, 160–167.
Järvelin, K., Chang, Y. W., & Vakkari, P. (2023). Characteristics of LIS research articles affecting their citation impact. *Journal of Librarianship and Information Science*. https://doi.org/10.1177/09610006231196344
Jurgens, D., Kumar, S., Hoover, R., McFarland, D., & Jurafsky, D. (2018). Measuring the evolution of a scientific field through citation frames. *Transactions of the Association for Computational Linguistics, 6*, 391–406.
Kammari, M. (2023). Time-stamp based network evolution model for citation networks. *Scientometrics, 128*(6), 3723–3741.
Kim, M., Baek, I., & Song, M. (2018). Topic diffusion analysis of a weighted citation network in biomedical literature. *Journal of the Association for Information Science and Technology, 69*(2), 329–342.
Kuhn, T. S. (1970). *The structure of scientific revolutions* (Vol. 111). University of Chicago Press.
Liang, G., Hou, H., Ding, Y., & Hu, Z. (2020). Knowledge recency to the birth of Nobel Prize-winning articles: Gender, career stage, and country. *Journal of Informetrics, 14*(3), 101053.
Liu, Y., & Chen, M. (2021). Applying text similarity algorithm to analyze the triangular citation behavior of scientists. *Applied Soft Computing, 107*, 107362.
Lu, Y., Yuan, M., Liu, J., & Chen, M. (2023). Research on semantic representation and citation recommendation of scientific papers with multiple semantics fusion. *Scientometrics, 128*(2), 1367–1393.
Marx, W., Bornmann, L., Barth, A., & Leydesdorff, L. (2014). Detecting the historical roots of research fields by reference publication year spectroscopy (RPYS). *Journal of the Association for Information Science and Technology, 65*(4), 751–764.
Nassiri, I., Masoudi-Nejad, A., Jalili, M., & Moeini, A. (2013). Normalized similarity index: An adjusted index to prioritize article citations. *Journal of Informetrics, 7*(1), 91–98.
Niraula, N., Banjade, R., Ştefănescu, D., & Rus, V. (2013). Experiments with semantic similarity measures based on LDA and LSA. In *Statistical language and speech processing: First international conference, SLSP 2013: Proceedings 1*, Tarragona, Spain, July 29–31, 2013 (pp. 188–199). Springer.
Pagani, R. N., Kovaleski, J. L., & Resende, L. M. (2015). Methodi Ordinatio: A proposed methodology to select and rank relevant scientific papers encompassing the impact factor, number of citations, and year of publication. *Scientometrics, 105*, 2109–2135.
Petruzzelli, A. M., Ardito, L., & Savino, T. (2018). Maturity of knowledge inputs and innovation value: The moderating effect of firm age and size. *Journal of Business Research, 86*, 190–201.
Pornprasit, C., Liu, X., Kiattipadungkul, P., Kertkeidkachorn, N., Kim, K. S., Noraset, T., ... & Tuarob, S. (2022). Enhancing citation recommendation using citation network embedding. *Scientometrics, 127*(9), 1–32.
Rodriguez-Prieto, O., Araujo, L., & Martinez-Romo, J. (2019). Discovering related scientific literature beyond semantic similarity: A new co-citation approach. *Scientometrics, 120*, 105–127.
Rohde, D. L., Gonnerman, L. M., & Plaut, D. C. (2006). An improved model of semantic similarity based on lexical co-occurrence. *Communications of the ACM, 8*(627–633), 116.
Rubin, R. E., & Rubin, R. G. (2020). *Foundations of library and information science*. American Library Association.
Sharma, R., Gopalani, D., & Meena, Y. (2023). An anatomization of research paper recommender system: Overview, approaches and challenges. *Engineering Applications of Artificial Intelligence, 118*, 105641.
Sheng, L., Lyu, D., Ruan, X., Shen, H., & Cheng, Y. (2023). The association between prior knowledge and the disruption of an article. *Scientometrics, 128*(8), 1–21.
Slyder, J. B., Stein, B. R., Sams, B. S., Walker, D. M., Jacob Beale, B., Feldhaus, J. J., & Copenheaver, C. A. (2011). Citation pattern and lifespan: A comparison of discipline, institution, and individual. *Scientometrics, 89*(3), 955–966.
Smith, T. B., Vacca, R., Krenz, T., & McCarty, C. (2021). Great minds think alike, or do they often differ? Research topic overlap and the formation of scientific teams. *Journal of Informetrics, 15*(1), 101104.
Su, W. H., Chen, K. Y., Lu, L. Y., & Huang, Y. C. (2021). Identification of technology diffusion by citation and main paths analysis: The possibility of measuring open innovation. *Journal of Open Innovation: Technology, Market, and Complexity, 7*(1), 104.

Synnestvedt, M. B., Chen, C., & Holmes, J. H. (2005). CiteSpace II: Visualization and knowledge discovery in bibliographic databases. In *AMIA annual symposium proceedings*, 2005 (Vol. 2005, p. 724). American Medical Informatics Association.

Tantanasiriwong, S., & Haruechaiyasak, C. (2014, May). Cross-domain citation recommendation based on co-citation selection. In *2014 11th International conference on electrical engineering/electronics, computer, telecommunications and information technology (ECTI-CON)*, 2014 (pp. 1–4). IEEE.

Tata, S., & Patel, J. M. (2007). Estimating the selectivity of TF–IDF based cosine similarity predicates. *ACM Sigmod Record, 36*(2), 7–12.

Thor, A., Marx, W., Leydesdorff, L., & Bornmann, L. (2016). Introducing CitedReferencesExplorer (CREx-plorer): A program for reference publication year spectroscopy with cited references standardization. *Journal of Informetrics, 10*(2), 503–515.

West, J. D., Wesley-Smith, I., & Bergstrom, C. T. (2016). A recommendation system based on hierarchical clustering of an article-level citation network. *IEEE Transactions on Big Data, 2*(2), 113–123.

Wu, X., Gao, C., Zang, L., Han, J., Wang, Z., & Hu, S. (2021). ESimCSE: Enhanced sample building method for contrastive learning of unsupervised sentence embedding. arXiv preprint arXiv:2109.04380

Yang, A. J. (2024). Unveiling the impact and dual innovation of funded research. *Journal of Informetrics, 18*(1), 101480.

Zhang, J., & Hou, J. (2023). Knowledge diffusion for individual literature from the perspective of Altmetrics: Models, measurement and features. *Journal of Information Science*. https://doi.org/10.1177/01655515231174387

Zhang, J., & Zhu, L. (2022). Citation recommendation using semantic representation of cited papers' relations and content. *Expert Systems with Applications, 187*, 115826.

Zhang, X., Xie, Q., & Song, M. (2021). Measuring the impact of novelty, bibliometric, and academic-network factors on citation count using a neural network. *Journal of Informetrics, 15*(2), 101140.

Zhou, H., Dong, K., & Xia, Y. (2023). Knowledge inheritance in disciplines: Quantifying the successive and distant reuse of references. *Journal of the Association for Information Science and Technology, 74*(13), 1515–1531.