Received: 11 June 2021 | Revised: 26 October 2021 | Accepted: 29 November 2021

DOI: 10.1002/asi.24610

RESEARCH ARTICLE

JASIST WILEY

Disclosing the relationship between citation structure and future impact of a publication

Shengzhi Huang[1,2] | **Jiajia Qian**[1,2] | **Yong Huang**[1,2] | **Wei Lu**[1,2] | **Yi Bu**[3] | **Jinqing Yang**[1,2] | **Qikai Cheng**[1,2]

[1]School of Information Management, Wuhan University, Wuhan, Hubei, China

[2]Information Retrieval and Knowledge Mining Laboratory, Wuhan University, Wuhan, Hubei, China

[3]Department of Information Management, Peking University, Beijing, China

Correspondence
Qikai Cheng, School of Information Management, Wuhan University, Wuhan, Hubei, China.
Email: chengqikai0806@163.com

Funding information
National Natural Science Foundation of China, Grant/Award Number: 72004168

Abstract

Each section header of an article has its distinct communicative function. Citations from distinct sections may be different regarding citing motivation. In this paper, we grouped section headers with similar functions as a structural function and defined the distribution of citations from structural functions for a paper as its citation structure. We aim to explore the relationship between citation structure and the future impact of a publication and disclose the relative importance among citations from different structural functions. Specifically, we proposed two citation counting methods and a citation life cycle identification method, by which the regression data were built. Subsequently, we employed a ridge regression model to predict the future impact of the paper and analyzed the relative weights of regressors. Based on documents collected from the Association for Computational Linguistics Anthology website, our empirical experiments disclosed that functional structure features improve the prediction accuracy of citation count prediction and that there exist differences among citations from different structural functions. Specifically, at the early stage of citation lifetime, citations from Introduction and Method are particularly important for perceiving future impact of papers, and citations from Result and Conclusion are also vital. However, early accumulation of citations from the Background seems less important.

1 | INTRODUCTION

Citation count is considered to be one of the commonly used metrics for measuring the impact of research outcome (Abrishami & Aliakbary, 2019; Bai et al., 2019; Oppenheim, 1995; Yu et al., 2014). However, citations are not equally important (Ding et al., 2013, 2014; Hu et al., 2013; Zhu et al., 2015), and some studies have confirmed that references listed in the bibliography of a paper generally make different contributions to that paper (Ding et al., 2013; Hou et al., 2011; Thelwall, 2019; Valenzuela et al., 2015; Wan & Liu, 2014). Exploring inherent differences among citations may help researchers better understand citation behavior, figure out an author's citation motivation, go deep into the citing process, and therefore, establish an effective scientific evaluation system.

With the open access of academic papers, researchers can easily obtain a large number of full-text scientific documents (Boyack et al., 2018; Lu et al., 2018; Thelwall, 2019), which makes content-based citation analysis possible. Content-based citation analysis focuses on revealing inherent differences among citations based on citation contexts. Previous research on content-based

Shengzhi Huang and Jiajia Qian contributed equally to this work.

citation analysis generally focused on analyzing citation distribution by section or overlap of citations among sections (Bertin, Atanassova, Gingras, & Larivière, 2016; Ding et al., 2013; Hu et al., 2013; Thelwall, 2019; Voos & Dagaev, 1976). However, the implicit relationship between the accumulation of citations from different citation locations and future impact of papers is still unclear.

Full-text scientific documents provide not only citation context information, but also fine-grained in-text citation counting information (Ding et al., 2014; Zhao & Strotmann, 2016). The sections of papers are arranged to clearly demonstrate the papers' topics, and each has its own distinct communicative function (Lu et al., 2018; Zhang, 2012). Lu et al. (2018) defined a group of section headers with a similar communicative function as a structural function, and all of the structural functions together constitute the functional structure in a specific field. Ding et al. (2013) and Pak et al. (2020) put forward different in-text citation counting methods, and confirmed that in-text citation count reveals the difference in importance among references. In this paper, we define the distribution of citations from structural functions for a paper as its citation structure. We used a real case from our collected data to clearly clarify what is citation structure and our research issue. As shown in Figure 1, two sample papers both acquired 15 citations 5 years after publication, and their citation structures in the fifth year are presented in bar charts. However, the citation structure of α and β are obviously different, and their citations are highly concentrated in Method and Background, respectively. We found that α and β possess different long-term impact despite their equal early impact. In effect, citations from different structural functions of a paper may indicate distinct citing motivations, and the citation structure of a paper may reflect the focus of the paper. For example, paper α cited frequently in Method is likely to be a methodology-oriented paper, and therefore, is cited later by papers using its methodology. Therefore, the early accumulation of citations from different structural functions may have different impact on future impact of a publication. The purpose of this study is to explore the relationship between citation structure and future impact of a publication, and identify the relative importance among in-text citations from different structural functions.

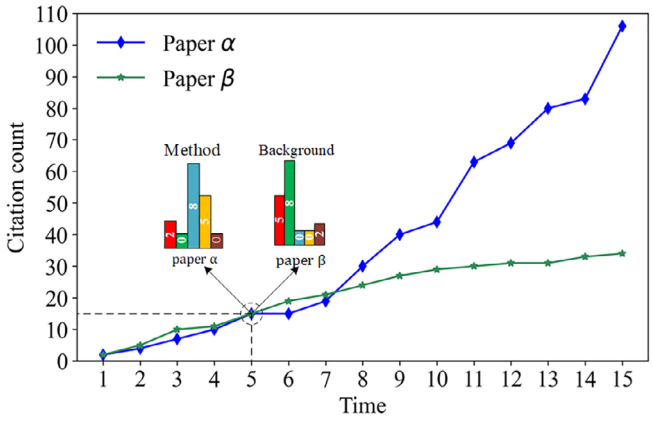

FIGURE 1 A real case for clarifying citation structure and our research issue

In this paper, we collected the full-text documents from the Association for Computational Linguistics (ACL) Anthology website as the data source (hereafter, ACL dataset) and identified the structural functions in the ACL dataset. After that, to count citations from different structural functions (i.e., citation structure), we extended the in-text citation counting methods presented by Pak et al. (2020). The extended method counts citation frequency from each structural function separately, while keeping the total citation frequency equal to that calculated by the original method. Moreover, considering the fact that citation trajectories and citation peaks for papers vary considerably (i.e., the effect of citation life cycle) (J. Wang, 2013; D. Wang et al., 2013), we also proposed a citation life cycle identification method based on corollary presented by D. Wang et al. (2013). Compared with a fixed citation time window for all papers, our method identifies the different stages of citation life cycle of a paper according to its citation trajectory, and enables to control the influence of time factor in our model. Subsequently, we utilized the citation distribution partition algorithm proposed by Huang et al. (2020) to identify the highly cited papers as the research object. We employed citation count to measure the scientific impact of a paper and used a ridge regression model, which to some extent solve the colinearity problem inevitably encountered in our study, to predict the future impact of a paper. Finally, relative weights of regressors in the ridge model are analyzed.

This study has the following contributions. First, it gives researchers some insight into the relationship between citation structure and future impact of a publication by the perspective of citation count prediction, which might contribute to building up an efficient research evaluation system. Specifically, the empirical experiments reveal that: (1) the functional structure feature can obviously improve the prediction accuracy of citation count prediction; and (2) there exist differences among in-text citations from different structural functions. Second, the extended in-text citation counting methods may also be utilized in other research, which analyzes in-text citation count from different structural functions or sections. Third, we derived a citation life cycle identification method, which could figure out the

time consumed for an individual paper to reach a certain ratio of its ultimate impact. This method may also be used in research involving the citation life cycle of papers.

The rest of the paper is organized as follows. In Section 2, we review the recent studies. In Section 3, we clarify the definition of research issues and methods utilized in this research. In Section 4, we give a detailed description of data collection and preprocessing. In Section 5, we provide the experimental setup and analysis of the empirical results. In Section 6, we discuss the contributions and limitations of this study. In Section 7, we summarize our research work.

2 | BACKGROUND

2.1 | Content-based citation analysis

In this study, we refer to citation analysis, which treats all citations with equal weight as traditional citation analysis, and citation analysis, which focuses on revealing inherent differences among citations based on citation contexts as content-based citation analysis.

Traditional citation analysis has been criticized for its weak basic theory on citation, unclear citing motivation, and inadequate citing process (Hjørland, 2013; Yang & Han, 2015). With the increasing accessibility of full-text documents, scholars are focusing on analyzing citation contexts in order to explore the wide variety of functions that citations perform. Many studies have confirmed the value of content-based citation analysis (Boyack et al., 2018; Hooten, 1991; Hou et al., 2011). Content-based citation analysis can be roughly divided into semantic content-based citation analysis and syntactic content-based citation analysis (Ding et al., 2013, 2014).

Semantic content-based citation analysis generally utilizes natural language processing technology and machine learning technology to explore the difference among citations. Teufel et al. (2006) defined the author's reason for citing a paper as the citation function, and proposed a citation function annotation schema. They presented a supervised machine learning framework with linguistically inspired features to classify citation functions. Small (2011) analyzed citation sentiments by contrasting the appearance of sentiment-bearing terms in citation contexts and the cue word sets. Zhu et al. (2015) created a dataset in which 10.3% of the references of 100 papers were manually labeled as influential references. They proposed a variety of features (e.g., context-based features and similarity-based features), and utilized support vector machines (SVMs) to classify references. Aljuaid et al. (2021) identified in-text citation sentiment, and used a series of machine learning algorithms to classify citations into binary classes (i.e., important and non-important). Liu and Chen (2021) proposed triangular citation structure, and identified high citation similarity in the citation contexts of triangular citation by text-similarity algorithm, which indicates a lazy citation motivation.

Scientific papers need to be structured to deliver their message persuasively (Thelwall, 2019), and many researchers focus on syntactic content-based citation analysis. Lu et al. (2018) presented the definition of structural function and functional structure, and proposed three widely applicable functional structure identification algorithms. Hu et al. (2013) visualized the distribution of citations in 350 papers published in the Journal of Informetrics with a four-section structure. They demonstrated that citations are highly concentrated in the first section of a paper. Ding et al. (2013) examined citation location and revealed that highly cited papers were more likely to appear in the introduction and literature review sections of citing papers. Wan and Liu (2014) manually divided the citation strength value of each citation of 40 papers into five levels, and employed regression method with a variety of features (e.g., in-text citation count, located section, etc.) to classify citation strength value. Bertin, Atanassova, Gingras, and Larivière (2016) investigated 45,000 papers published in the Public Library of Science journal families, and found that the introduction and discussion sections contained most of the references. Thelwall (2019) analyzed citation distribution by section and the overlap of citations among sections in 799,055 PubMed Central open access articles. They argued that section headers are not reliable as indicators of citation function in lowly cited articles.

In this study, we employed the definition of structural function and functional structure proposed by Lu et al. (2018). In short, each structural function represents a group of section headers with similar communicative functions, and functional structure is a terminology describing the set composed of all structural functions. The goal of this study is to explore the relationship between citation structure and future impact of a paper. Hence, this study belongs to the study of syntactic content-based citation analysis. However, unlike previous studies, this paper stands on the perspective of citation count prediction and tries to discover the relative importance of citations from different structural functions.

2.2 | Citation count prediction

The accumulation of citations consumes time (Fu & Aliferis, 2010). Various studies focus on predicting

citation count, which may measure the impact of a paper, researcher, and organization in a timely fashion.

For a long time, statistical methods have been utilized to analyze and predict citation count of a paper. For instance, Burrell (2002) used the Poisson process to confirm that the longer a paper is not cited, the less likely it will be cited in the future. Burrell (2003) also employed a similar model to predict the citation count of a paper and showed that the expected number of future citations is a linear function of the current citation frequency. Adams (2005) found a strong correlation between the ranking lists for publications in the life and physical sciences ranked by early citations and those ranked by later citations, and concluded that early citations (1–2 years) are statistically a good proxy for predicting the long-term impact of a paper. Brody et al. (2006) examined the arXiv.org e-print archive, and found a positive correlation (about 0.4) between the number of times an article is downloaded and its later citation count. D. Wang et al. (2013) proposed a differential equation model (WSB) to fit the citation trajectories of papers from different journals and disciplines. The WSB model was observed to effectively estimate ultimate citations frequency of a paper. Based on inherent quality of papers, citation life cycle, early citations, and early citers' impact, Bai et al. (2019) proposed a paper potential index model, which achieves satisfactory predictive performance on 183,336 papers from American Physical Society dataset.

Machine learning algorithms have also been widely utilized in citation count prediction. For example, Fu and Aliferis (2010) utilized content-based and bibliometric features of a paper to predict its citation count by SVMs. Yu et al. (2014) presented paper features, journal features, author features, and citation features of papers, and utilized stepwise multiple regression to predict citations of a paper after 5 years of publication. Onodera and Yoshikane (2015) also proposed various extrinsic factors of a paper, and utilized negative binomial multiple regression to predict citation frequency. Their experiments showed that the proportion of references within 3 and 5 years and number of references are the most important features. Abramo et al. (2019) analyzed the importance of a publication's early citations and the impact factor of the hosting journal in citation count prediction by two linear models. Their experimental results disclosed that a citation time window of 3 years can achieve acceptable accuracy in predicting the long-term impact of publications, and the impact factor (i.e., the impact factor becomes negligible only 2 years after publication). Toubia et al. (2021) proposed three metrics (speed, volume, and circuitousness) to quantify the semantic progression of texts, and confirmed that these metrics have a significant impact on the long-term impact of a paper by a least absolute shrinkage and selection operator (Lasso) regression. In addition, deep learning algorithms have also achieved fruitful results in citation count prediction. Abrishami and Aliakbary (2019) employed early citations (3–5 years) of the paper to predict its long-term impact on the recurrent neural network. The empirical experiments based on 175,432 papers from five journals showed that their model achieves state-of-the-art results. Ruan et al. (2020) employed paper features, journal features, author features, reference features, and early citation features as input variables, and used a feedforward neural network to predict the citations of a paper after 5 years of publication. Their model achieved satisfactory predictive performances on about 50 K papers in the library, information, and documentation field. Akella et al. (2021) used altmetrics (e.g., social media features) to predict early and long-term citations of a paper. After comparing various models, they found that neural networks and ensemble models performed best in term of F1 scores for their tasks.

The citation peaks for papers vary considerably and citation time window length does play a very important role in citation analysis (J. Wang, 2013; D. Wang et al., 2013). In this study, we assumed that the relationship between citation structure and future impact of a paper may evolve over time. Unlike previous studies, which set a unified time window for all papers, we derived a citation life cycle identification method based on corollary presented by D. Wang et al. (2013) to control the influence of time factor in our modeling. Subsequently, we employed the linear model to explore the relationship between citation structure and future impact of a publication.

3 | METHOD

3.1 | Problem definition

In this paper, we aim to reveal the relationship between citation structure and future impact of a publication, and figure out the inherent difference among in-text citations from different structural functions. Specifically, we employed citation count to measure the impact of a paper. We used a linear model to analyze the relationship between the accumulation of citations from different structural functions in the top $r_1\%$ of citation life cycle (i.e., $X^I_{t_{r_1\%}}$) and the impact of a paper in top $r_2\%$ of citation life cycle (i.e., $y_{t_{r_2\%}}$), where $r_2 > r_1$. In the following text, we explain the key concepts and steps for this study step by step, and the overall research framework can be found in Figure 2.

First, we followed the definition of structural functions and functional structure (Lu et al., 2018). To be

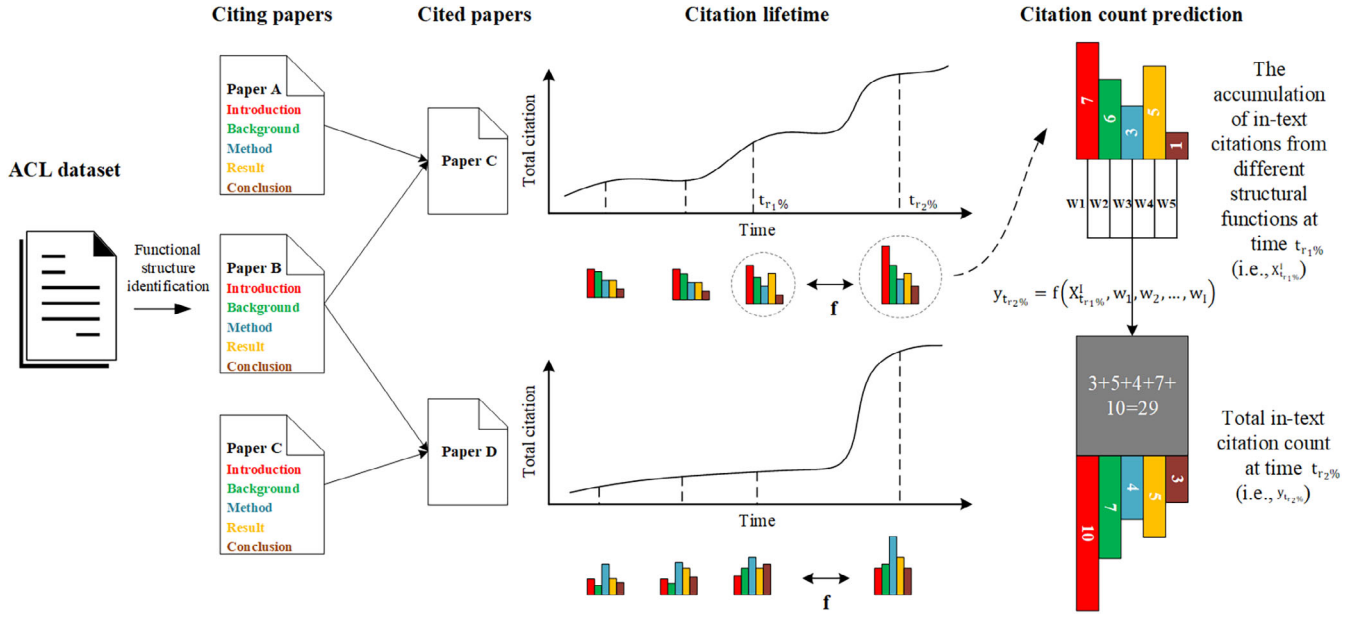

FIGURE 2 Overview of the key steps for this study

specific, we denoted a type of structural function, which defines the roles of sections in conveying the ideas of authors as i, and functional structure, which is comprised of a variety of structural functions as I. That is to say, $i \in I$. As shown in the left of Figure 2, we utilized functional structure identification methods and a text classification algorithm to identify five structural functions in the ACL dataset.

Second, we extended two in-text citation counting methods proposed by Pak et al. (2020), in order to make them suitable for separately counting in-text citations from different structural functions. After the directed citation relationship between the cited paper and the citing paper in the ACL dataset is built, these extended in-text citation counting methods are utilized to count citation frequency for papers.

Third, the accumulation of citations of papers increases unevenly over time and presents different citation trajectories (Avramescu, 1979; Bai et al., 2019; Mingers & Burrell, 2006; D. Wang et al., 2013). To control the influence of citation life cycle in our modeling, we derived a citation life cycle identification method, which can calculate the time, $t_{r\%}$, consumed for a paper to reach $r\%$ of its ultimate number of citations, as shown in the middle of Figure 2. This citation life cycle identification method is adopted to specify time window for papers in citation count prediction.

Finally, we employed in-text citation count of a paper, α, at time, $t_{r_1\%}$, from different structural functions ($X^I_{t_{r_1\%}}$) to predict total in-text citation count of α at $t_{r_2\%}$ (i.e., $y_{t_{r_2\%}}$), as shown in the right of Figure 2. Therefore, the current research issue turns into citation count prediction, which is fundamentally a fitting regression task. Our goal is to find the functional relationship, $y_{t_{r_2\%}} = f\left(X^I_{t_{r_1\%}}, w_1, w_2, \ldots, w_I\right)$.. It is worth noting that we limited the inputs to in-text citation count from different structural functions at $t_{r_1\%}$ (independent variables) and time (control variable), to keep the problem definition simple and general. Previous studies show that a linear model is an effective method for citation count prediction and feature selection analysis (Abramo et al., 2019; Djokoto et al., 2020; Jimenez et al., 2020; Yu et al., 2014). Therefore, we employed a linear model and analyzed the relative weights of regressors (i.e., w_i), which quantitively gauges the effect of independent variables on the dependent variable. Thus, the relative weights of regressors disclose the relative significance among citations from different structural functions. However, there may exist some correlation among citations from different structural functions for a paper, which may result in colinearity in our regression equations. For instance, Introduction citations tend to be cited within Background (Thelwall, 2019). Hence, the ridge regression model and the Lasso model, which are more suitable for problematic data than traditional ordinary least squares regression model, are great choices. Considering Lasso is more suitable for feature selection and obtaining zero weight of regressors, which is inconsistent with our original intention, the ridge regression model was employed in this study.

3.2 | In-text citation counting methods

A citation sentence may be supported by an in-text citation of only one reference or an in-text citation of multiple references. Pak et al. (2020) proposed the definition of independent mention and nonindependent mention, where the former indicates that one reference is mentioned independently, while the latter means one reference is mentioned with other references in a citation sentence. They also proposed the full counting method as well as the fractional counting method. In this study, we extended the two in-text citation counting methods mentioned above, in order to make them suitable for separately counting in-text citation count from different structural functions. The formulas for the two extended in-text counting methods are as follows.

$$C^{\alpha}_{\text{full}}(t) = \sum\nolimits_{\beta \in A_{\alpha}(t)} \sum\nolimits_{i \in I} \sum\nolimits_{j=1}^{n_i^{\alpha\beta}} 1 \tag{1}$$

$$C^{\alpha}_{\text{frac}}(t) = \sum\nolimits_{\beta \in A_{\alpha}(t)} \sum\nolimits_{i \in I} \sum\nolimits_{j=1}^{n_i^{\alpha\beta}} w_{ij}^{\alpha\beta} \tag{2}$$

$C^{\alpha}_{\text{full}}(t)$ and $C^{\alpha}_{\text{frac}}(t)$ denote in-text citation count of a paper, α, at time, t, in full counting method and fractional counting method, respectively. $A_{\alpha}(t)$ indicates a collection of citing papers that have cited α at t. I represents functional structure in the specific domain. $n_i^{\alpha\beta}$ denotes how many times a citing paper, β, mentioned α in a structural function, i. In the full counting, the mention weight, $w_{ij}^{\alpha\beta}$, is equal to 1. In the fractional counting, $w_{ij}^{\alpha\beta}$ is the inverse of the number of references in a citation sentence. In addition, we denote $C^{\alpha}(t,i) = \sum_{\beta \in A_{\alpha}(t)} \sum_{j=1}^{n_i^{\alpha\beta}} w_{ij}^{\alpha\beta}$, which represents in-text citation count from i acquired by α at t. Therefore, we have $C^{\alpha}(t) = \sum_{i \in I} C^{\alpha}(t,i)$.

To facilitate understanding of Equations (1) and (2), we present a simple case. As shown in Figure 3, there are only two citing papers (B, C) and two cited papers (A, D), and arrows indicate directed citation relationships similar to Figure 2, where A was mentioned independently three times in three citation sentences of B, and A and D were simultaneously mentioned one time in one citation sentence of C. Therefore, C^{A}_{full}, C^{D}_{full}, C^{A}_{frac}, and C^{D}_{frac} are 4, 1, 7/2, and 1/2 in turn. Under the assumption of $I = \{i_1, i_2\}$, $C^{\alpha}_{\text{full}}(t,i)$ and $C^{\alpha}_{\text{frac}}(t,i)$ can be separately calculated in each i, and keep the left equal to the sum of the right.

3.3 | Citation life cycle identification method

In this research, we derived a citation life cycle identification method based on the corollary of the citation model proposed by D. Wang et al. (2013). Before introducing our citation life cycle identification method, we briefly reviewed the citation model (hereafter, WSB).

The WSB model describes the citation trajectory of papers from different journals and disciplines in a uniform form, and its mathematical formula is as follows, Equation (3).

$$C^{\alpha}_{\text{WSB}}(t) = m\left(e^{\lambda_{\alpha}\phi\left(\frac{\ln t - \mu_{\alpha}}{\sigma_{\alpha}}\right)} - 1\right) \tag{3}$$

$$\phi(x) = (2\pi)^{-1/2} \int_{-\infty}^{x} e^{-y^2/2} dy \tag{4}$$

$C^{\alpha}_{\text{WSB}}(t)$ represents accumulative citations of a paper, α, at time, t. m indicates the initial attraction of a new paper and can be measured by the average number of references that each paper contains. They employed a

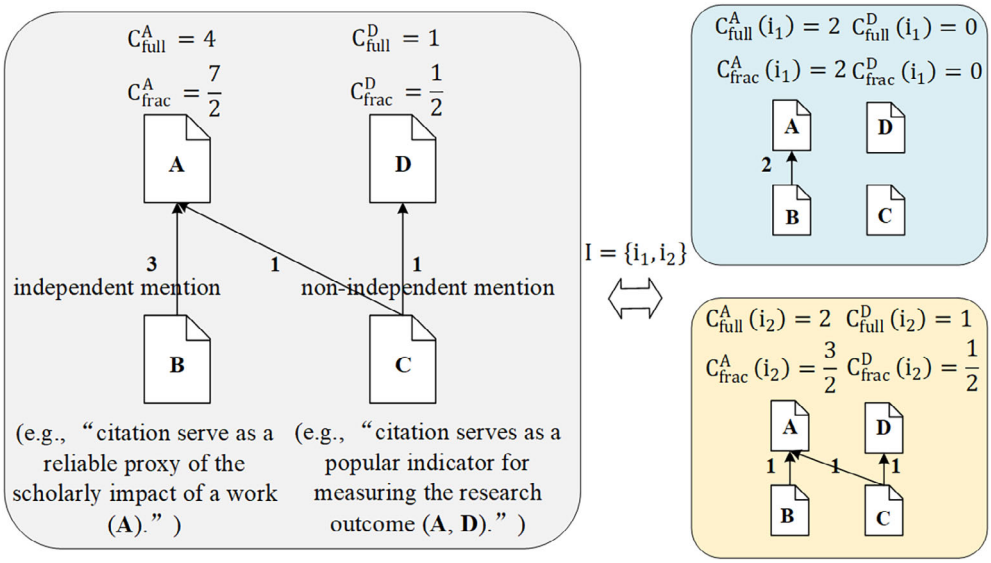

FIGURE 3 Examples for the two in-text citation counting methods

lognormal survival probability to depict the obsolescence of papers, as shown in Equation (4). μ_α denotes immediacy, controlling the time for α to reach its citation peak; and σ_α is longevity, gauging the decay rate of α. λ_α measures the quality of α.

D. Wang et al. (2013) also derived ultimate impact, c_α^∞, and impact time, $t_{50\%}$, of α. More specifically, when $t \to \infty$, $\phi \to 1$. c_α^∞ represents the ultimate citation count that α acquires during its citation lifetime, as shown in Equation (5). $t_{50\%}$ is the time consumed for α to reach geometric mean of c_α^∞, and can be derived from Equation (6).

$$C_{\mathrm{WSB}}^{\alpha}(\infty) = m\left(e^{\lambda_\alpha} - 1\right) \tag{5}$$

$$(m(m + C^{\alpha}(\infty)))^{\frac{1}{2}} = C_{\mathrm{WSB}}^{\alpha}(t_{50\%}) \tag{6}$$

$t_{50\%}$ is determined by the citation trajectory of each individual paper and is not necessarily the same in different papers. The merit of $t_{50\%}$ allows us to determine the middle stage of the citation life cycle of each paper.

To identify different stages of the citation life cycle, we extended Equation (6). Specifically, we simply replaced the exponent of the left side of Equation (6) (i.e., 1/2) with any value between 0 and 1 (i.e., $r\%$), and obtained Equation (7). The solution of Equation (7), $t_{r\%}$, indicates the time necessary for a paper to reach $r\%$ of its ultimate citations. To ensure the solvability and simplify the computing process of Equation (7), we retained the assumption presented by D. Wang et al. (2013) (i.e., $exp(\lambda_\alpha \phi) \gg 1$) and simplified m to 1. Finally, Equation (7) is simplified as Equation (8). The numerical solution of Equation (8) can be easily obtained and is only decided by μ_α and σ_α of a paper. Compared with a fixed time window for all papers, our method aims to eliminate the impact of citation life cycle in our subsequent regression analysis.

$$\left(m\left(m + C_{\mathrm{WSB}}^{\alpha}(\infty)\right)\right)^{r\%} = C_{\mathrm{WSB}}^{\alpha}(t_{r\%}) \tag{7}$$

$$r\% = \phi\left(\frac{\ln t_{r\%} - \mu_\alpha}{\sigma_\alpha}\right) \tag{8}$$

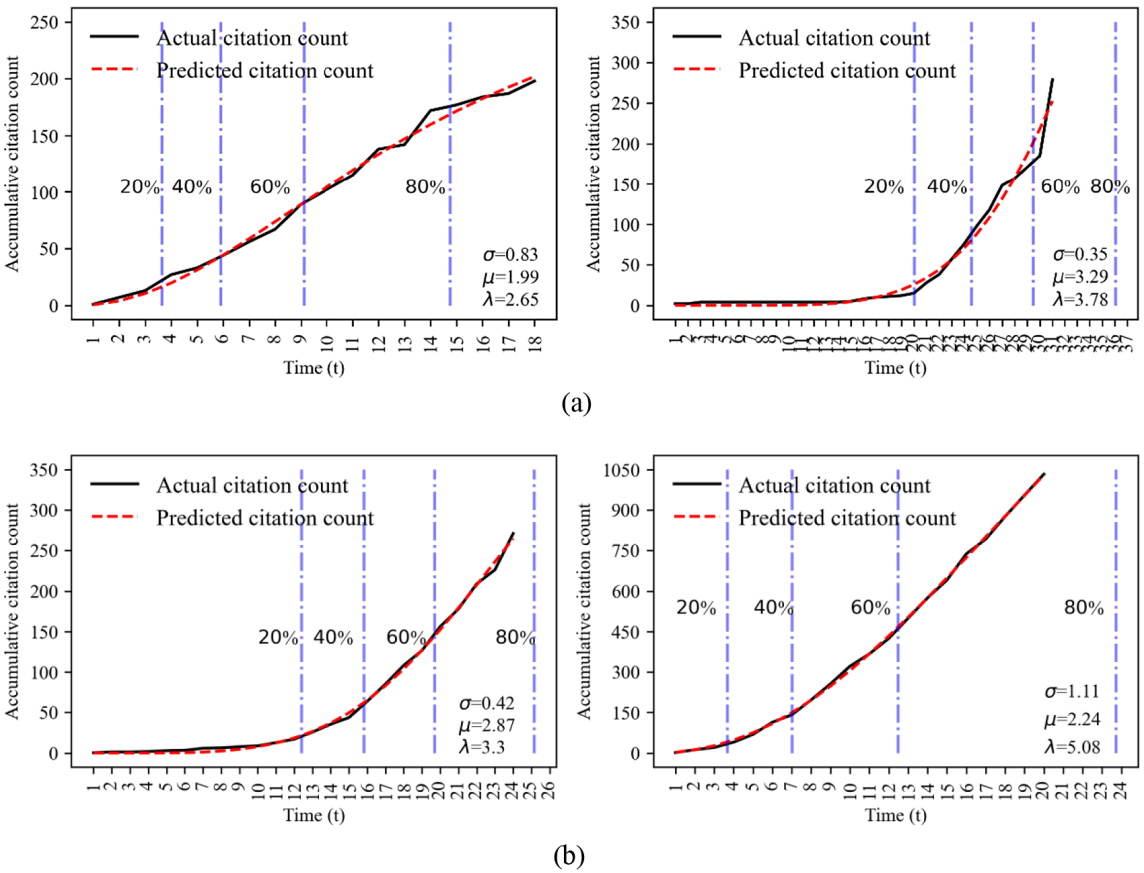

FIGURE 4 Citation life cycle identification results of four sample papers. (a) Two cases in the full counting method. (b) Two cases in the fractional counting method

For easy understanding, we randomly selected four sample papers to clarify our proposed method. As shown in Figure 4, the x-axis starts with the time when a paper is first cited. The black curve is the actual citation trajectory of a paper, and the red curve is the fitting result of WSB. The fitting parameters of WSB are listed in the lower right corner of each sub-figure. In line with our above view, $t_{r\%}$ of each paper is not necessarily the same. For example, in Figure 3A, $t_{60\%}$ of two papers is about 9 and 30, respectively. In addition, because geometric mean is used in Equation (7), the blue vertical lines, which denote $t_{r\%}$ of the paper, are not evenly distributed. Notably, due to the boundary effect of data collection, only the paper on the left sub-figure of Figure 4a reached 80% of its ultimate citations.

3.4 | Prediction model

Considering that there may be some correlation among citations from different structural functions, we employed ridge regression model to fulfill citation count prediction.

$$\hat{y_\alpha} = \sum_{i=1}^{I} w_i x_i^\alpha + bt + c \tag{9}$$

$$\text{loss} = \frac{1}{N}\sum_{\alpha=1}^{N}(y_\alpha - \hat{y_\alpha})^2 + \lambda \sum_{i=1}^{I} w_i^2 \tag{10}$$

As shown in Equations (9) and (10), the independent variable, x_i^α, indicates the accumulation of in-text citations of a paper, α, at time, t_1, from a structural function, i (i.e., $C^\alpha(t_1,i)$). t is the difference between t_1 and a fixed time point, t_0, and is a control variable. The dependent variable, y_α, is the total in-text citation count of α at time, t_2, where $t_2 > t_1$. $\hat{y_\alpha}$ indicates the predicted value. λ is a ridge regression parameter and N denotes the number of samples.

As we mentioned earlier, Equation (8) was utilized to determine t_1 and t_2 in linear regression analysis for each paper. To be more specific, we set t_2 as $t_{r_2\%}$ and t_1 as $t_{r_1\%}$, where $r_2\% > r_1\%$. Subsequently, we obtained the following input–output data pairs, $X_{t_{r_1\%}}(x_1,x_2...,x_I), y_{t_{r_2\%}}$.

In order to evaluate the prediction performance comprehensively, we also simplified the independent variable to conduct control experiments as a baseline. More specifically, x_i^α is simplified as z^α, which represents total in-text citation count acquired at $t_{r_1\%}$ (i.e., $z^\alpha = \sum_{i=1}^{I} x_i^\alpha$). In a few words, the experimental group utilized the functional structure feature, while the control group did not.

4 | DATA

We collected PDF versions of full-text documents from the ACL Anthology website from 1965 to 2020, comprising 59,133 papers. These collected PDF files were parsed by using the Grobid toolkit. In the process of parsing the PDF files, the metadata of some articles was lost, we removed these cases and 52,540 papers remained. After that, we extracted the papers' reference lists and citation contexts. We obtained 1,174,413 references and 1,440,200 pieces of citation context data. Finally, we deduplicated the 52,540 papers indexed in the ACL Anthology website and 1,174,413 references according to the title, authors, and publication year of the articles, and obtained 290,937 unique papers.

4.1 | Functional structure identification

In this research, the functional structure clustering algorithm (Lu et al., 2018) and the text classification algorithm were employed to identify functional structure in the ACL dataset.

Specifically, first, as section headers are not case-sensitive, all section headers extracted from the citation context data were converted to lower case. We removed their punctuation and numbers. Second, the section headers were ranked based on their frequency in descending order. Finally, *Top*100 high frequency section headers were selected and classified to generate a domain-specific functional structure. Finally, structural functions in the ACL dataset are classified into the following five categories: Introduction, Background, Method, Experiment and result (hereafter, Result), and Discussion and conclusion (hereafter, Conclusion). The section headers under each structural function are shown in Table 1. It is worth mentioning that, because we simply extracted the section header that occurred immediately before the citation location, the extracted section headers may be subsection headers.

By clustering *Top*100 high frequency section headers into structural functions, 727,709 pieces of citation context data denoted as S_1 can be classified into their structural functions. The remaining 712,491 pieces of citation context data are recorded as S_2. To classify structural functions on S_2, we built a balanced subset by randomly sampling from S_1 and trained an automatic structural functions classifier by the BERT model (Devlin et al., 2018). Specifically, we randomly sampled 25,000 pieces of citation context data from each structural function in S_1. The subset was split into training set, validation set, and test set at a ratio of 8:1:1. A pretrained BERT model was fine-tuned for classifying automatically structural functions on S_2. A "Tensorflow" framework is utilized to implement neural network training. After training 10 epochs, the accuracy of the BERT model in the training set and the test set is 0.9407 and 0.8298,

respectively. The trained BERT model was applied to S_2. The classification results on S_2 can be found in the Descriptive analysis subsection.

4.2 | Regression data preparation

In citation analysis, highly cited papers have higher research value and citations of lowly cited papers are more random (Thelwall, 2019). Therefore, we focused on analyzing the citation structure in highly cited papers. By employing the dividing method proposed by Huang et al. (2020), we identified 544 and 270 highly cited papers in full counting method and fractional counting method, respectively. The more detailed partitioning results are shown in the Descriptive analysis subsection.

TABLE 1 The functional structure schema generated from the ACL dataset

Functional structure	Section headers
Introduction	Introduction, motivation, introduction and motivation, introduction and related work
Background	Related work, related works, related research, prior work, previous work, background, background and related work
Method	Model, models, method, methods, methodology, Approach, system description, features, baselines
Experiment and result	Experiment, experiments, experimental setup, experiment setup, experimental results, experimental settings, experimental setting, experiments and results, results, data, dataset, datasets, corpus, figure, setup, settings, evaluation, evaluation metrics, implementation details, implementation, training, preprocessing, analysis
Discussion and conclusion	Conclusion, conclusions, conclusion and future work, discussion, results and discussion, conclusions and future work, future work

In our experiments, we repeated the linear regression analysis for different $r_1\%$. Specifically, we set $t_{r_2\%}$ as $t_{80\%}$ and $t_{r_1\%}$ as $t_{20\%}$ to $t_{70\%}$ with intervals of 10%, where $t_{r_1\%}$ and $t_{r_2\%}$ can be obtained by Equation (8). Notably, due to the boundary effect of data collection, some highly cited papers did not reach $t_{80\%}$. Finally, 348 and 173 highly cited papers were selected as cases for this study in full counting method and fractional counting method, respectively. The detailed experimental settings are shown in Table 2.

5 | EXPERIMENTS AND RESULTS

5.1 | Descriptive analysis

Before the regression analysis, we present the distribution characteristics of the ACL dataset. The publications' distribution in the ACL dataset is shown in Figure 5, where the left sub-figure shows the number of papers published annually and the right sub-figure gives a cumulative view. In Figure 6, citation distribution in two in-text citation counting methods is shown in a double-logarithmic coordinate system. It is worth noting that, because the count in the fractional counting method may be less than 1, we plus one before taking its log.

We also present the structural functions' distributions in S_1 (i.e., structural functions identified by clustering *Top*100 high frequency section headers), S_2 (i.e., structural functions identified by BERT model), and $S_1 \cup S_2$, as shown in Table 3. We found that, in S_1, citation context data from Introduction (i_1) and Background (i_2) occupy the majority, and account for 47.66% and 32.93%, respectively. This may be caused by the fact that section headers in i_1 and i_2 are more standardized and unified than section headers in Method (i_3), Result (i_4), and Conclusion (i_5) in a specific field (Hu et al., 2013). In $S_1 \cup S_2$, citations are highly concentrated in i_1 and i_2, which is consistent with the previous results (Ding et al., 2013; Hu et al., 2013).

By using the dividing method proposed by Huang et al. (2020), we classified 290,937 papers into three categories, as shown in Table 4. In addition, citation structures in high, medium, and low cited publications are visualized, respectively, as shown in Figure 7. The x-axis

TABLE 2 Experimental setup

	Full counting		Fractional counting	
Counting methods	Experimental group	Control group	Experimental group	Control group
$r_1\%$	$X_{t_{r_1\%}}, y_{t_{r_2\%}}$	$Z_{t_{r_1\%}}, y_{t_{r_2\%}}$	$X_{t_{r_1\%}}, y_{t_{r_2\%}}$	$Z_{t_{r_1\%}}, y_{t_{r_2\%}}$

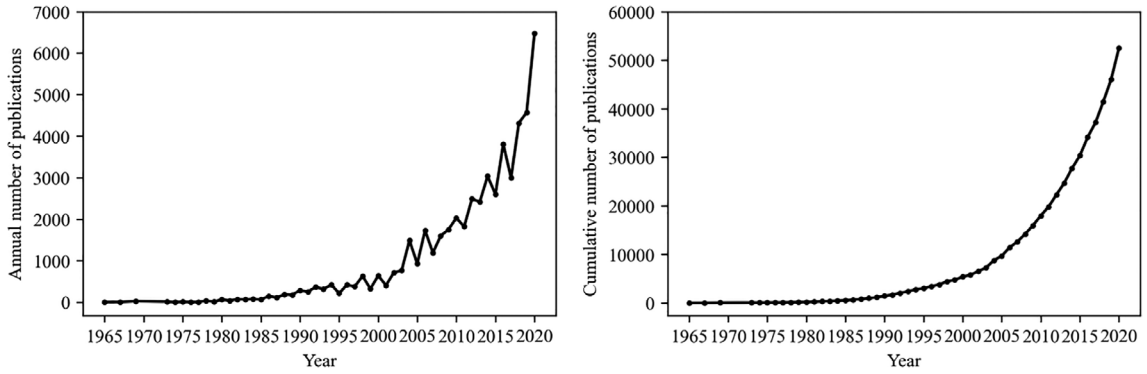

FIGURE 5 Distribution of publications

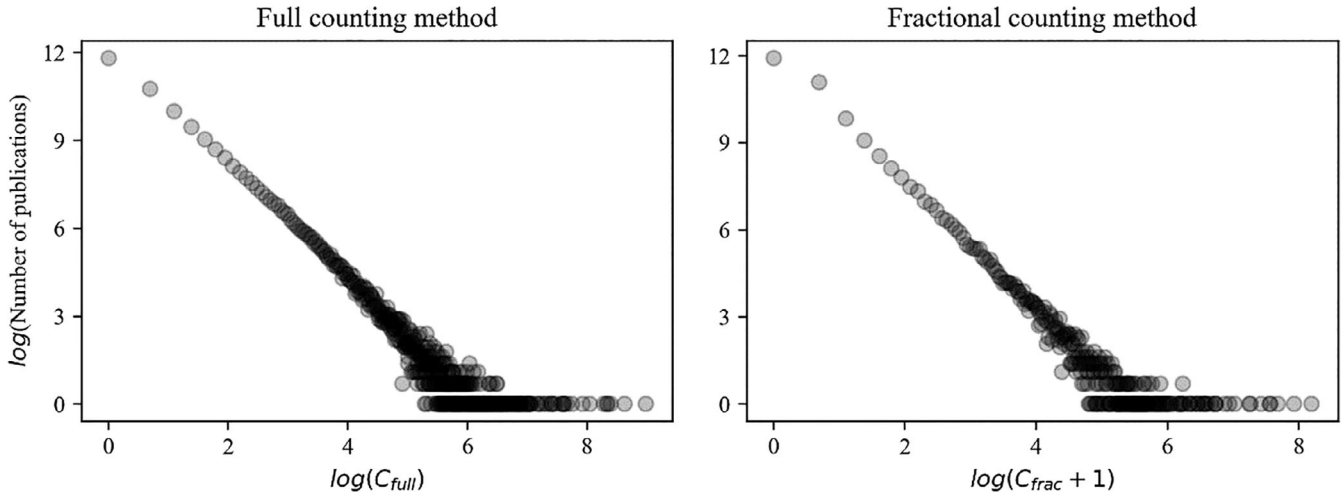

FIGURE 6 Citation distributions in the two in-text citation counting methods

TABLE 3 Distribution of structural functions in S_1, S_2 and $S_1 \cup S_2$

I	Introduction (i_1)	Background (i_2)	Method (i_3)	Result (i_4)	Conclusion (i_5)
S_1	346,829	239,643	24,086	83,119	34,032
S_2	149,176	153,474	152,185	171,891	85,765
$S_1 \cup S_2$	496,005	393,117	176,271	255,010	119,797

denotes the structural functions and the y-axis is the proportion of in-text citation count in each structural function. We found that the proportion of in-text citations in i_3 and i_4 of highly cited papers obviously exceeds that of both the medium and lowly cited papers. In contrast, the proportion of in-text citations in i_1, i_2, and i_5 of medium and lowly cited papers obviously exceeds that of highly cited papers. Thus, highly cited papers tend to be more frequently cited in Method and Result. This rough conclusion motivated us to make further analysis in the following studies.

5.2 | Regression analysis

In this study, regression experiments are implemented through a ridge regression algorithm encapsulated in the MASS library (Venables and Ripley, 2002) of the R programming language, and a grid search algorithm is employed to determine the ridge regression parameter, λ.

We found that the functional structure feature can improve the prediction accuracy of citation count prediction. However, the increase of accuracy decays over time and is obviously significant at the early stage of citation

TABLE 4 The results of dividing citation distribution

Counting method	Level	Interval	Number
Full counting	Lowly cited	$0 \le x < 8$	238,113
	Medium cited	$8 \le x < 197$	29,533
	Highly cited	$197 \le x < \infty$	544
Fractional counting	Lowly cited	$0 \le x < 8$	255,292
	Medium cited	$8 \le x < 122$	12,628
	Highly cited	$122 \le x < \infty$	270

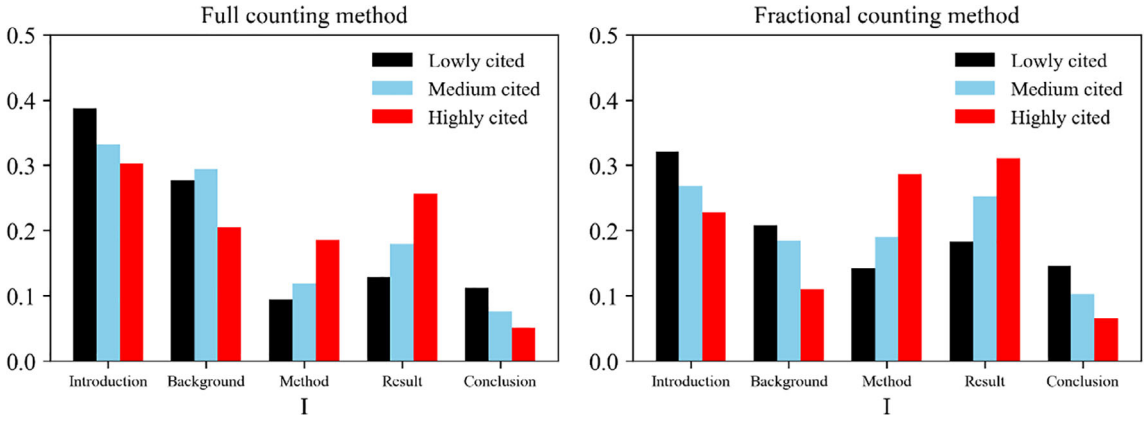

FIGURE 7 Citation structure in high, medium, and low cited publications

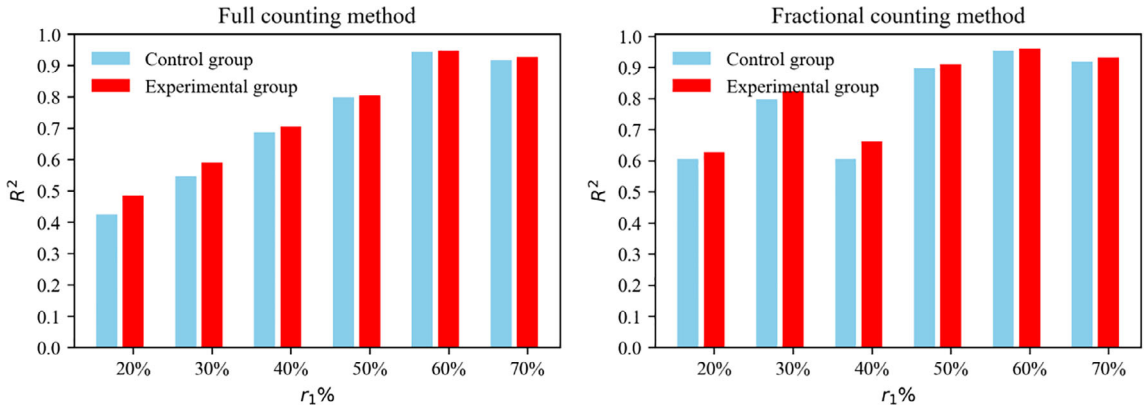

FIGURE 8 R^2 in the ridge regression model

lifetime. As shown in Figure 8, the x-axis represents $r_1\%$, and the y-axis indicates goodness of fit, R^2. The red bar and the blue bar denote R^2 of the experimental group and the control group, respectively. We found that R^2 of the experimental group exceeds that of the control group for different $r_1\%$, which indicates that the separate in-text citation count from different structural functions, $x_i\,(i \in I)$, contains richer citation details than the total in-text citation count, $z\,(\sum_{i=1}^{I} x_i)$. However, the increase of R^2 decreases gradually with the increase of $t_{r_1\%}$. This may be caused by the fact that, with $t_{r_1\%}$ increases, more abundant information about the citation history of the papers is fed into the linear model and the difficulties of citation count prediction decreased. Therefore, both groups can easily achieve excellent performance. In addition, it could also be caused by the formation of specific citation patterns of individual papers at the later stage of citation lifetime. As mentioned by Thelwall (2019), some highly cited articles acquired most of their citations from one section type. As shown in Figure 9, we randomly

chose citation histories of four papers to clarify this view. The paper in the left sub-figure of Figure 9a is frequently cited in i_1 and i_2, while the paper in the right sub-figure of Figure 9a is mainly cited in i_4. The citations of the two papers in Figure 9b are both highly concentrated in i_3 and i_4. Hence, papers tend to be cited frequently in one or more fixed structural functions at the later stage of citation lifetime, which means that the structural function feature may no longer provide variance information for each paper.

By analyzing the regression coefficients in the ridge regression model (i.e., $w_i\,(i \in I)$), we found differences among citations from different structural functions. As shown in Figure 10, when $r_1\% \leq 40\%$, the coefficient weight of Introduction and Method (i.e., w_1 and w_3) is greater than that of Background, Result, and Conclusion (i.e., w_2, w_4, and w_5). This means that citations from i_1 and i_3 of a publication are especially important for perceiving the future impact of the publication. As mentioned by Hu et al. (2013), highly cited and important papers would be so famous that they would tend to come to mind first when an author requires something convincing and persuasive, and tend to be highly concentrated in the first section of a paper. Therefore, the papers cited frequently in i_1 at the early stage are more likely to be high quality, which makes them acquire higher ultimate impact during their lifetime. In addition, i_1 generally contains the most references (Bertin, Atanassova, Gingras, & Larivière, 2016; Boyack et al., 2018; Hu et al., 2013; Tang & Safer, 2008; Voos & Dagaev, 1976). Hu et al. (2013) argued that articles focusing on methodology are more likely to have an excessive number of citations in the method section. We agree with Hu et al. (2013) and hold the view that articles, which have more citations in i_3 at the early stage tend to focus on presenting new methodologies. Methodology-oriented papers may be more likely to be cited by other studies, which utilize new methods to solve different research issues. Therefore, methodology-oriented papers tend to be more highly cited than other types of papers (Boyack et al., 2018). In addition, w_4 and w_5 are both larger than w_2, which means that citations from i_4 and i_5 are also vital. The papers cited frequently in i_4 and i_5 at the early stage may present new results and conclusions, so they may also make an important academic contribution. Finally, the weight of Background (i.e., w_2) is almost

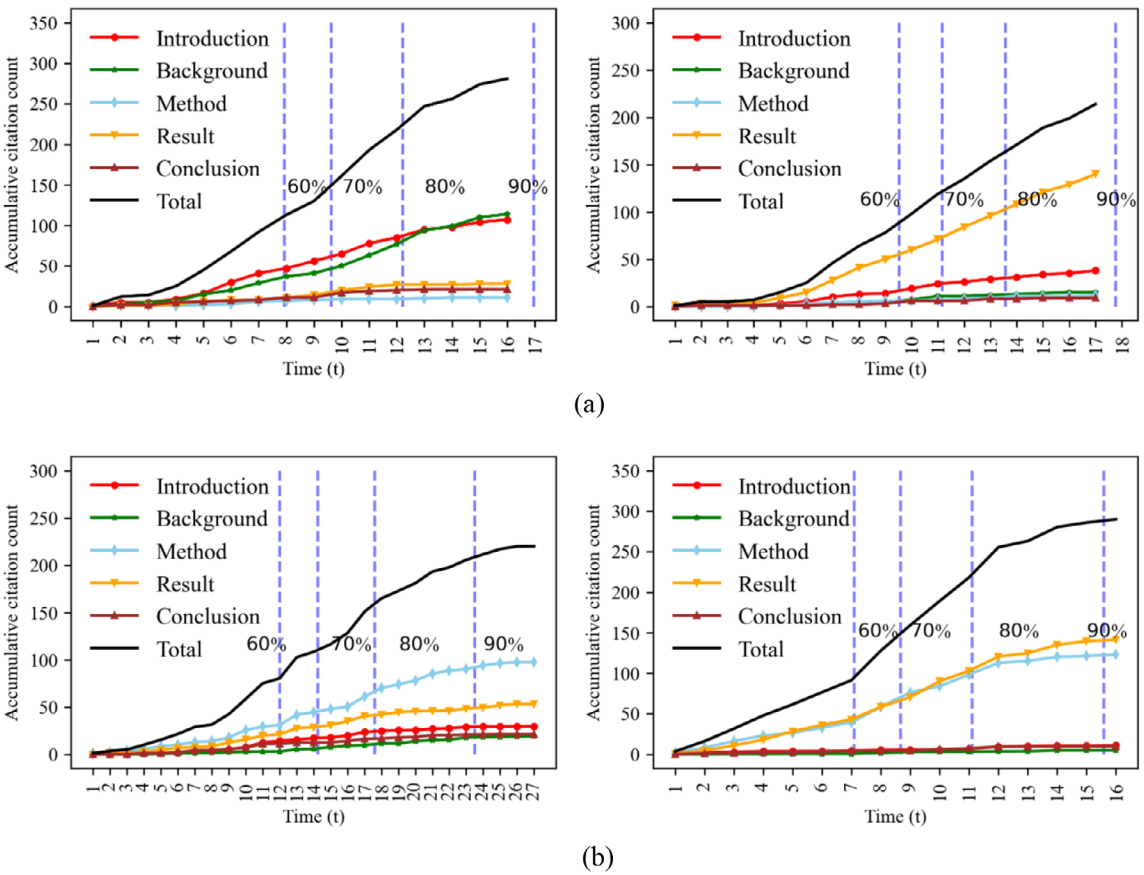

FIGURE 9 Accumulative citation count for four sample papers. (a) The two cases in the full counting method. (b) The two cases in the fractional counting method

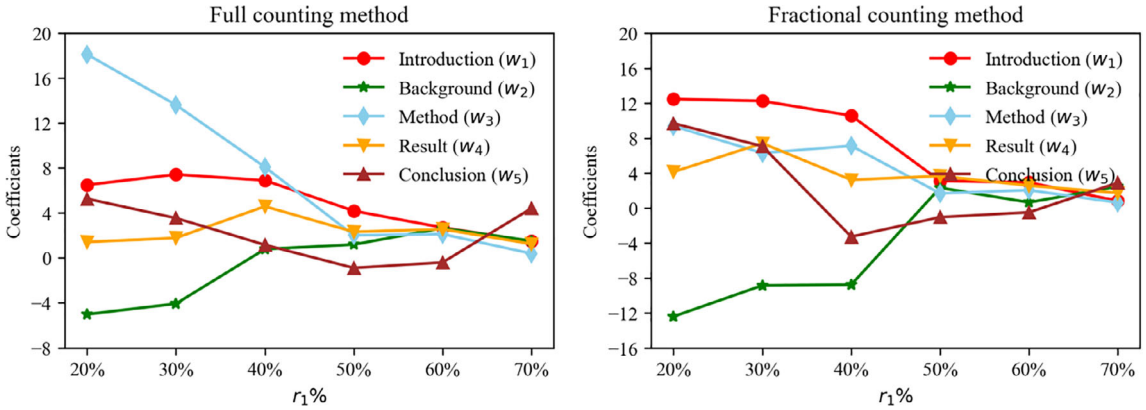

FIGURE 10 Coefficients in the ridge regression model

TABLE 5 Coefficients and λ in the full counting method

r_1%	Introduction	Background	Method	Result	Conclusion	λ
20%	6.5041**	−5.0094	18.1197***	1.4234	5.2685	10.00
30%	7.4316***	−4.0650	13.6210***	1.7809	3.5448	6.00
40%	6.9037***	0.7882	8.0942***	4.5895***	1.1694	9.20
50%	4.1886***	1.2009	2.0201***	2.3233***	−0.8826	6.00
60%	2.7012***	2.6472***	2.1224***	2.5545***	−0.3916	2.80
70%	1.4883***	1.5175***	0.3983*	1.2560***	4.4144***	4.80

*$p < .05$. **$p < .01$. ***$p < .001$.

always the smallest. This is possibly caused by the fact that papers, which are cited in i_2 may be cited only for introducing the research background. Hence, citations in i_2 at the early stage may be more random and perfunctory. Moreover, when $r_1\% \leq 40\%$, we also calculated the Spearman correlation coefficients, $\rho(x_i,y)$, between independent variables and dependent variable. In the full (fractional) counting method, we found that $\rho(x_1,y)$ and $\rho(x_3,y)$ are the largest with an average of 0.48 (0.53) and 0.49 (0.55), respectively, and $\rho(x_2,y)$ is the smallest with an average of 0.32 (0.31), which coincides with the relative weight of regressors.

When $r_1\% \geq 50\%$, the difference among regressors is no longer obvious. On the one hand, this is possibly caused by the fact that $t_{r_1\%}$ is gradually approaching $t_{80\%}$; on the other hand, as mentioned above, this may be due to the formation of specific citation patterns of individual papers. However, in the full (fractional) counting method, we still found that $\rho(x_1,y)$ and $\rho(x_3,y)$ are the largest with an average of 0.63 (0.59) and 0.61 (0.66), respectively, and $\rho(x_2,y)$ is the smallest with an average of 0.44 (0.44).

More detailed coefficients in ridge regression are listed in Tables 5 and 6. Notably, w_2 and w_5 are negative and insignificant in some regression equations. This may be caused by the functional similarity among structural functions identified in this study, which leads to the colinearity in our regression equations. For instances, it is difficult to clearly classify a small part of the section headers in Table 2, such as "introduction and related work." In addition, there is a natural overlap among the functions of sections (Thelwall, 2019).

5.3 | Robustness analysis

To test the robustness of the above results, we respectively changed the strategy of identifying highly cited papers, the values for $t_{r_1\%}$ and $t_{r_2\%}$, and the regression methods, and repeated the above experiments.

Specifically, we simply regarded papers with more than 100 citations as highly cited papers, and selected 893 and 232 highly cited papers in the full counting method and the fractional counting methods, respectively. We kept the rest of the settings unchanged, and repeated the above experiments. As shown in Tables 7 and 8, relative differences among these coefficients are similar to those in Tables 5 and 6.

Subsequently, we still employed the dividing method proposed by Huang et al. (2020), but set $t_{r_2\%}$ as $t_{75\%}$ and $t_{r_1\%}$ as $t_{15\%}$ to $t_{65\%}$ with intervals of 10%. We, respectively, selected 380 and 184 highly cited papers in

TABLE 6 Coefficients and λ in the fractional counting method

r_1%	Introduction	Background	Method	Result	Conclusion	λ
20%	12.4758***	−12.3981*	9.374***	4.1406*	9.6918	6.40
30%	12.2496***	−8.8541**	6.2934***	7.4150***	7.0698	2.80
40%	10.5759***	−8.7628*	7.1224***	3.2219***	−3.2836	3.20
50%	3.1098***	2.3179	1.6826***	3.6663***	−1.0451	4.80
60%	2.9238***	0.6431	2.0419***	2.5673***	−0.4969	1.60
70%	0.8095*	2.4613**	0.6200***	1.7305***	2.8885**	7.20

*$p < .05$. **$p < .01$. ***$p < .001$.

TABLE 7 Coefficients and λ in the full counting method

r_1%	Introduction	Background	Method	Result	Conclusion	λ
20%	5.7214***	−0.9331	15.4215***	2.5559*	6.3210*	16.00
30%	5.9283***	−0.0435	11.2680***	2.7818***	5.0326*	12.40
40%	5.9629***	1.3160	7.7874***	3.8757***	2.4694	10.40
50%	3.7881***	1.6592***	2.0208***	2.4432***	0.4910	6.80
60%	2.6742***	2.0790***	2.2928***	2.3044***	0.4471	2.80
70%	1.4930***	1.5526***	0.4328***	1.3057***	3.8484***	3.60

*$p < .05$. **$p < .01$. ***$p < .001$.

TABLE 8 Coefficients and λ in the fractional counting method

r_1%	Introduction	Background	Method	Result	Conclusion	λ
20%	11.4112***	−9.6295	9.8165***	4.2882*	7.2434	6.40
30%	11.4984***	−7.2793**	6.4236***	7.4113***	5.2286	3.20
40%	9.8634***	−7.8675**	7.0134***	3.2514***	−2.6674	3.20
50%	2.8516***	2.3904*	1.7613***	3.6045***	−0.1907	5.20
60%	2.8217***	0.7301	2.0569***	2.5197***	−0.2772	1.60
70%	0.8618**	2.3380***	0.6264***	1.7391***	2.9259***	7.20

*$p < .05$. **$p < .01$. ***$p < .001$.

TABLE 9 Coefficients and λ in the full counting method

r_1%	Introduction	Background	Method	Result	Conclusion	λ
15%	4.7504	−3.0892	12.4078**	2.3864	5.8634	10.00
25%	8.1534***	−2.9133	16.0295***	2.9776	−1.3691	11.20
35%	7.8016***	−3.0739*	9.1699***	4.5164***	3.4430	4.80
45%	7.6123***	−2.5142*	6.6153***	2.7134**	−7.8753	3.20
55%	3.1930***	1.3405**	1.7710***	3.2196***	−2.6446*	2.80
65%	2.3124***	0.4791	2.3954***	1.5183***	−0.4344	2.00

*$p < .05$. **$p < .01$. ***$p < .001$.

the full counting and the fractional counting methods, and repeated the above experiments. Finally, we also obtained similar results, as shown in Tables 9 and 10.

Moreover, we also repeated the experiments by Lasso. Specifically, we employed the Lasso algorithm encapsulated in the glmnet library (Friedman et al., 2010; Simon et al., 2011) of the R programming language. Finally, we still got similar results, as shown in Tables 11 and 12. It should be noted that the significant test of coefficients generally is not directly carried out in the Lasso.

In the above robustness tests, the relative differences among the coefficients from different structural functions are almost the same. In addition, we also found that R^2 of the experimental group exceeds that of the control group. Hence, our results are stable.

TABLE 10 Coefficients and λ in the fractional counting method

r_1%	Introduction	Background	Method	Result	Conclusion	λ
15%	8.0082*	−7.5768	14.1099***	1.1522	8.825	9.20
25%	14.5494***	−17.2708***	8.0900***	4.2981**	4.5042	2.00
35%	8.8868***	−2.3885	7.2789***	4.6004***	1.1831	4.80
45%	8.2673***	−6.4386*	5.1185***	2.3603***	−2.1959	3.20
55%	3.2798***	0.8600	2.1393***	2.5645***	−0.9008	2.00
65%	2.6530***	0.3064	2.2493***	1.3171***	0.0654	1.20

*$p < .05$. **$p < .01$. ***$p < .001$.

TABLE 11 Coefficients and λ in the full counting method

r_1%	Introduction	Background	Method	Result	Conclusion	λ
20%	6.4917	−5.0108	18.1323	1.417	5.2964	0.00
30%	7.2433	−4.0618	13.6260	1.7791	3.5576	0.00
40%	6.9343	0.6400	8.0809	4.5222	1.2279	3.60
50%	4.4240	0.0000	2.0550	1.8043	0.0000	66.80
60%	2.8107	2.2316	2.1068	2.4207	0.0000	5.20
70%	1.5141	1.3194	0.4244	1.1272	3.7890	75.20

TABLE 12 Coefficients and λ in the fractional counting method

r_1%	Introduction	Background	Method	Result	Conclusion	λ
20%	12.5063	−12.5346	9.3819	4.1236	9.7844	0.00
30%	10.7209	−4.9694	5.9813	7.3762	6.4873	4.00
40%	10.0456	−8.1204	7..0010	3.1673	−2.4057	1.20
50%	3.4875	0.0000	1.5882	3.0965	0.0000	22.40
60%	2.9298	0.3120	2.0328	2.5012	0.0000	2.00
70%	0.8490	1.6967	0.6003	1.4281	1.8214	40.40

6 | DISCUSSION

The functional structure forms the navigation of different kinds of knowledge (Lu et al., 2018). Citations from different structural functions, to some extent, reflect the types of cited papers (Hu et al., 2013; Tahamtan & Bornmann, 2018).

In this paper, we found that, in the field of computational linguistics, citations from Introduction and Method at the early stage of citation lifetime are particularly important for measuring the future impact of papers. Indeed, papers cited in Introduction tend to serve as illustrating the motivation and research problems (Ding et al., 2013; Lu et al., 2018), and are more likely to be pioneers of research issues or state of the art for the research question, and have thus inspired the author(s) to carry out the research work (Fang, 2018). The papers frequently cited in Method tend to be methodology-oriented papers (Hu et al., 2013), and are especially important in computational linguistics, as they provide methodological trajectories (Lu et al., 2018). Moreover, citations from Result and Discussion are also vital. Actually, citations in the Result and Discussion are frequently about comparable results (Hu et al., 2015), and citing authors are more likely to agree with the conclusion from a paper cited in Result and Discussion (Bertin, Atanassova, Sugimoto, & Lariviere, 2016). Hence, the papers frequently cited in Result and Conclusion at the early stage of citation lifetime tend to present valuable findings and make valuable academic contributions. In contrast, citations from Background at the early stage of citation lifetime seem to be less important. This may be due to the fact that papers acquiring citations in Background might only play the role of depicting the research background of the citing papers, and their role may be replaced by other articles on similar topics.

This research also provides several interesting indications for citation count prediction and scientific evaluation. First, this study provides a new exploratory direction for predicting citation count, that is, more features extracted from citation contexts could be investigated further in citation count prediction. For example,

citations from different structural functions may work as distinct input features or output features, which may improve the performance and interpretability of citation count prediction. Actually, based on the above idea, we have proposed a fine-grained citation count prediction task in our future article. Second, our empirical results show that there are differences among citations from different structural functions. Hence, an effective scientific evaluation system should consider the relative importance among citations. For instance, Zhao and Strotmann (2020) presented a location filtered citation counting method to make essential citations stand out. However, unlike simply filtering citations, we argue that the relative weight of regressors in our linear model is well worth considering in weighted citation count.

There are some limitations to this study. First, it focused only on analyzing citation structure in highly cited papers, which can be effectively identified in citation life cycle by our proposed method. Citation structure in medium and lowly cited papers needs to be further explored. Second, there is a functional similarity among structural functions, which leads to the collinearity in our ridge regression analysis. Thus, a sophisticated functional structure identification method needs to be considered in the future. Finally, this paper only analyzed the citation structure by investigating publications in computational linguistics. Whether the differences of citations from different functional structures exist in other fields requires further studies.

7 | CONCLUSIONS

In this study, we employed the ridge regression model to quantitatively reveal the relationship between citation structure and future impact of a publication. Our experimental results show that the functional structure feature can obviously improve the prediction accuracy of citation count prediction, which suggests citation from different structural functions contains richer citation details than the total citation count. Therefore, in citation count prediction, citation may not be treated equally and, other features extracted from citation contexts may be worth further exploration. After analyzing weights of regressors, we also revealed the relative importance among in-text citations from different structural functions. More specifically, in the field of computational linguistics, the early accumulation of citations in Introduction and Method is especially important for measuring the future impact of a paper. The early accumulation of citations in Result and Conclusion is also vital. However, the citations in Background at the early stage of the citation life cycle seem less important. Thus, in citation analysis, researchers may need to consider the inherent difference among citations, and attach great importance to essential citations. In addition, it must be strongly emphasized that the observed results are statistical, not definitive, and individual papers may have a diversity of performance. In future research, citation sentences deserve further analysis from a semantic level by natural language processing technologies by offering a better understanding of the inherent differences among citations from different structural functions. In conclusion, our findings may help researchers to quantitatively perceive differences among citations from different structural functions and figure out the potential reasons for the differences, which contribute to building up an efficient research evaluation system.

ACKNOWLEDGMENT

This work was supported by the Youth Science Foundation of the National Natural Science Foundation of China (grant no. 72004168).

ORCID

Shengzhi Huang https://orcid.org/0000-0002-7035-4627
Yi Bu https://orcid.org/0000-0003-2549-4580

REFERENCES

Abramo, G., D'Angelo, C. A., & Felici, G. (2019). Predicting publication long-term impact through a combination of early citations and journal impact factor. *Journal of Informetrics*, *13*, 32–49.

Abrishami, A., & Aliakbary, S. (2019). Predicting citation counts based on deep neural network learning techniques. *Journal of Informetrics*, *13*, 485–499.

Adams, J. (2005). Early citation counts correlate with accumulated impact. *Scientometrics*, *63*, 567–581.

Akella, A. P., Alhoori, H., Kondamudi, P. R., Freeman, C., & Zhou, H. (2021). Early indicators of scientific impact: Predicting citations with altmetrics. *Journal of Informetrics*, *15*, 101128.

Aljuaid, H., Iftikhar, R., Ahmad, S., Asif, M., & Afzal, M. T. (2021). Important citation identification using sentiment analysis of in-text citations. *Telematics and Informatics*, *56*, 101492.

Avramescu, A. (1979). Actuality and obsolescence of scientific literature. *Journal of the American Society for Information Science*, *30*, 296–303.

Bai, X., Zhang, F., & Lee, I. (2019). Predicting the citations of scholarly paper. *Journal of Informetrics*, *13*, 407–418.

Bertin, M., Atanassova, I., Gingras, Y., & Larivière, V. (2016). The invariant distribution of references in scientific articles. *Journal of the Association for Information Science and Technology*, *67*, 164–177.

Bertin, M., Atanassova, I., Sugimoto, C. R., & Lariviere, V. (2016). The linguistic patterns and rhetorical structure of citation context: An approach using n-grams. *Scientometrics*, *109*, 1417–1434.

Boyack, K. W., van Eck, N. J., Colavizza, G., & Waltman, L. (2018). Characterizing in-text citations in scientific articles: A large-scale analysis. *Journal of Informetrics*, *12*, 59–73.

Brody, T., Harnad, S., & Carr, L. (2006). Earlier web usage statistics as predictors of later citation impact. *Journal of the American Society for Information Science and Technology*, *57*, 1060–1072.

Burrell, Q. L. (2002). Will this paper ever be cited? *Journal of the American Society for Information Science and Technology*, *53*, 232–235.

Burrell, Q. L. (2003). Predicting future citation behavior. *Journal of the American Society for Information Science and Technology*, *54*, 372–378.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K., (2018). *Bert: Pre-training of deep bidirectional transformers for language understanding*. arXiv preprint arXiv:1810.04805.

Ding, Y., Liu, X., Guo, C., & Cronin, B. (2013). The distribution of references across texts: Some implications for citation analysis. *Journal of Informetrics*, *7*, 583–592.

Ding, Y., Zhang, G., Chambers, T., Song, M., Wang, X., & Zhai, C. (2014). Content-based citation analysis: The next generation of citation analysis. *Journal of the Association for Information Science and Technology*, *65*, 1820–1833.

Djokoto, J. G., Agyei-Henaku, K. A. A., Afrane-Arthur, A. A., Badu-Prah, C., Gidiglo, F. K., & Srofenyoh, F. Y. (2020). What drives citations of frontier application publications? *Heliyon*, *6*, e05428.

Fang, H. (2018). A discussion of citations from the perspective of the contribution of the cited paper to the citing paper. *Journal of the Association for Information Science and Technology*, *69*, 1513–1520.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*, 1–22.

Fu, L., & Aliferis, C. (2010). Using content-based and bibliometric features for machine learning models to predict citation counts in the biomedical literature. *Scientometrics*, *85*, 257–270.

Hjørland, B. (2013). Citation analysis: A social and dynamic approach to knowledge organization. *Information Processing & Management*, *49*, 1313–1325.

Hooten, P. A. (1991). Frequency and functional use of cited documents in information science. *Journal of the American Society for Information Science*, *42*, 397–404.

Hou, W.-R., Li, M., & Niu, D.-K. (2011). Counting citations in texts rather than reference lists to improve the accuracy of assessing scientific contribution: Citation frequency of individual articles in other papers more fairly measures their scientific contribution than mere presence in reference lists. *BioEssays*, *33*, 724–727.

Hu, Z., Chen, C., & Liu, Z. (2013). Where are citations located in the body of scientific articles? A study of the distributions of citation locations. *Journal of Informetrics*, *7*, 887–896.

Hu, Z., Chen, C., & Liu, Z. (2015). The recurrence of citations within a scientific article. In A. A. Salah, A. A. A. Salah, C. Sugimoto, U. Al, & Y. Tonta (Eds.), *Proceedings of ISSI 2015 Istanbul, Proceedings of ISSI 2015 Istanbul: 15th International Society of Scientometrics and Informetrics Conference* (pp. 221–229). Bogazici Universitesi.

Huang, Y., Bu, Y., Ding, Y., & Lu, W. (2020). Partitioning highly, medium and lowly cited publications. *Journal of Information Science*, *47*, 609–614.

Jimenez, S., Avila, Y., Dueñas, G., & Gelbukh, A. (2020). Automatic prediction of citability of scientific articles by stylometry of their titles and abstracts. *Scientometrics*, *125*, 3187–3232.

Liu, Y., & Chen, M. (2021). Applying text similarity algorithm to analyze the triangular citation behavior of scientists. *Applied Soft Computing*, *107*, 107362.

Lu, W., Huang, Y., Bu, Y., & Cheng, Q. (2018). Functional structure identification of scientific documents in computer science. *Scientometrics*, *115*, 463–486.

Mingers, J., & Burrell, Q. L. (2006). Modeling citation behavior in management science journals. *Information Processing & Management*, *42*, 1451–1464.

Onodera, N., & Yoshikane, F. (2015). Factors affecting citation rates of research articles. *Journal of the Association for Information Science and Technology*, *66*, 739–764.

Oppenheim, C. (1995). The correlation between citation counts and the 1992 research assessment exercise ratings for British library and information science university departments. *Journal of Documentation*, *51*, 18–27.

Pak, C. M., Wang, W., & Yu, G. (2020). An analysis of in-text citations based on fractional counting. *Journal of Informetrics*, *14*, 101070.

Ruan, X., Zhu, Y., Li, J., & Cheng, Y. (2020). Predicting the citation counts of individual papers via a BP neural network. *Journal of Informetrics*, *14*, 101039.

Simon, N., Friedman, J., Hastie, T., & Tibshirani, R. (2011). Regularization paths for Cox's proportional hazards model via coordinate descent. *Journal of Statistical Software*, *39*, 1–13.

Small, H. (2011). Interpreting maps of science using citation context sentiments: A preliminary investigation. *Scientometrics*, *87*, 373–388.

Tahamtan, I., & Bornmann, L. (2018). Core elements in the process of citing publications: Conceptual overview of the literature. *Journal of Informetrics*, *12*, 203–216.

Tang, R., & Safer, M. A. (2008). Author-rated importance of cited references in biology and psychology publications. *Journal of Documentation*, *64*, 246–272.

Teufel, S., Siddharthan, A., & Tidhar, D. (2006). Automatic classification of citation function. *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*. pp. 103–110.

Thelwall, M. (2019). Should citations be counted separately from each originating section? *Journal of Informetrics*, *13*, 658–678.

Toubia, O., Berger, J., & Eliashberg, J. (2021). How quantifying the shape of stories predicts their success. *Proceedings of the National Academy of Sciences*, *118*, e2011695118.

Valenzuela, M., Ha, V., & Etzioni, O., (2015). *Identifying meaningful citations*. Paper presented at *AAAI Workshop: Scholarly Big Data*. p. 13.

Venables, W. N., & Ripley, B. D. (2002). *Modern Applied Statistics with S* (4th ed.). New York: Springer.

Voos, H., & Dagaev, K. S. (1976). Are all citations equal? Or, did we Op. Cit. your idem? *The Journal of Academic Librarianship*, *1*, 19–21.

Wan, X., & Liu, F. (2014). Are all literature citations equally important? Automatic citation strength estimation and its applications. *Journal of the Association for Information Science and Technology*, *65*, 1929–1938.

Wang, D., Song, C., & Barabási, A.-L. (2013). Quantifying long-term scientific impact. *Science*, *342*, 127–132.

Wang, J. (2013). Citation time window choice for research impact evaluation. *Scientometrics*, *94*, 851–872.

Yang, S., & Han, R. (2015). Breadth and depth of citation distribution. *Information Processing & Management*, *51*, 130–140.

Yu, T., Yu, G., Li, P.-Y., & Wang, L. (2014). Citation impact prediction for scientific papers using stepwise regression analysis. *Scientometrics*, *101*, 1233–1252.

Zhang, L. (2012). Grasping the structure of journal articles: Utilizing the functions of information units. *Journal of the American Society for Information Science and Technology*, *63*, 469–480.

Zhao, D., & Strotmann, A. (2016). Dimensions and uncertainties of author citation rankings: Lessons learned from frequency-weighted in-text citation counting. *Journal of the Association for Information Science and Technology*, *67*, 671–682.

Zhao, D., & Strotmann, A. (2020). Deep and narrow impact: Introducing location filtered citation counting. *Scientometrics*, *122*, 503–517.

Zhu, X., Turney, P., Lemire, D., & Vellino, A. (2015). Measuring academic influence: Not all citations are equal. *Journal of the Association for Information Science and Technology*, *66*, 408–427.

How to cite this article: Huang, S., Qian, J., Huang, Y., Lu, W., Bu, Y., Yang, J., & Cheng, Q. (2021). Disclosing the relationship between citation structure and future impact of a publication. *Journal of the Association for Information Science and Technology*, 1–18. https://doi.org/10.1002/asi.24610