Author Co-citation Analysis Using Semantic and Positional Similarity: An Empirical Study of Effectiveness Postprint

Authors: Ruhao Zhang

Date: 2023-04-01T00:00:00+00:00

Abstract

[Purpose/Significance] Author co-citation analysis is an important method for exploring the intellectual structure of a field. Under complex disciplinary development contexts, its reliance on co-citation frequency for measuring author relatedness is quite controversial. To address this, an improved method for author co-citation analysis based on semantic and positional similarity is proposed. [Method/Process] Based on the introduction of fundamental principles, this study conducts an empirical validation of the effectiveness of the improved method for author co-citation analysis based on semantic and positional similarity, taking the library and information science field as an example. It performs full-text citation mining on the CNKI journal database, extracts citation sentences and citation positions, combines a pre-trained domain word embedding model to calculate deep similarity between co-cited documents and the strength of association between authors, and uses network analysis and factor analysis to compare the effectiveness differences between this method and traditional methods. [Results/Conclusion] The results demonstrate that the improved method for author co-citation analysis based on semantic and positional similarity can more accurately identify the strength of association among co-cited authors, discover more detailed disciplinary intellectual structures, and possesses scalability and applicability.

Full Text

An Improved Author Co-citation Analysis Method Based on Semantic and Location Similarity and Its Empirical Study

Zhang Ruhao

Chengdu Library and Information Center, Chinese Academy of Sciences, Chengdu 610041
Department of Library, Information and Archives Management, School of Economics and Management, University of Chinese Academy of Sciences, Beijing 100049
National Science Library, Chinese Academy of Sciences, Beijing 100190

Abstract: [Purpose/Significance] Author co-citation analysis (ACA) is a vital method for exploring domain knowledge structures. However, under complex disciplinary development trends, its reliance on co-citation frequency for measuring author relevance remains controversial. To address this, we propose an improved ACA method based on semantic and location similarity. [Method/Process] Building upon fundamental principles, we demonstrate the effectiveness of this improved method using the library and information science (LIS) field as an example. We conducted full-text citation mining from the CNKI journal database, extracted citation sentences and their locations, and combined a pre-trained domain word embedding model to calculate deep similarity between co-cited documents and association strength between authors. Network analysis and factor analysis were employed to compare the differences in effectiveness between this method and traditional approaches. [Result/Conclusion] Results prove that the improved ACA method based on semantic and location similarity can more accurately identify the correlation strength among co-cited authors, discover more detailed disciplinary knowledge structures, and possesses scalability and applicability.

Keywords: author co-citation analysis; citation content analysis; co-citation proximity analysis; full-text citation analysis; domain knowledge structure
Classification Number: G250
DOI: 10.13266/j.issn.0252-3116.2020.08.013

1 Introduction

Author co-citation analysis (ACA), first proposed by H.D. White and B.C. Griffith in 1981 [**?**], is one of the primary methods in bibliometric research. Its fundamental assumption posits that when author A and author B are both cited by article C, a co-citation relationship R exists between A and B, with the relationship growing stronger as the number of co-citations increases. Under this methodological assumption, all co-cited authors share certain thematic connections and exhibit distinct thematic clustering distributions within a certain scope. Consequently, ACA is commonly used for identifying disciplinary knowledge structures and discovering scientific communities [**?**].

In recent years, scientific research advancement and technological innovation have accelerated interdisciplinary integration, with new research fields emerging in ways that are difficult to detect and scholars' research topics becoming increasingly diverse. These trends challenge traditional ACA methods, which rely solely on limited bibliographic information and employ simple frequency

statistics based on co-occurrence (binary presence/absence) as the basis for measuring inter-author relationships. This approach is simplistic and controversial [?], ignoring the real and deep connections between co-cited authors, making its accuracy and scientific validity difficult to guarantee.

We propose a novel ACA method based on content semantics and location proximity (semantic and proximity-based author co-citation analysis, SPACA), which measures the relationship strength between authors through the similarity of citation sentence content and the proximity of citation locations to overcome the limitations of traditional methods. By collecting full-text documents from the CNKI journal database in the LIS field, we conduct a comparative empirical study between this method and traditional ACA to explore the characteristics and applicability of this improved approach, thereby promoting the precision innovation of ACA methodology.

2 Literature Review

Contemporary automated text processing technologies have matured, enabling literature full-text to be obtained in semi-structured formats from databases (e.g., PubMed, BioMed Central, CiteSeer, arXiv). These conditions have facilitated the emergence and growing importance of full-text-based citation content analysis, which delves into citing documents to obtain microscopic citation data such as citation intensity, location, function, and semantics, enabling quantitative measurement of relationship strength and impact degree represented by citations.

Citation content analysis research can be divided into two levels [?]: the semantic level focusing on citation content itself, including citation topic analysis [?], citation function analysis [?], and citation scope identification [?]; and the syntactic level focusing on external features of citation content, including citation location analysis [?] and citation intensity analysis [?].

Citation content analysis provides opportunities for improving traditional co-citation theory by moving from bibliographic information to full-text analysis. Some scholars have focused on using citation location proximity in full-text documents to replace traditional co-citation counting. For instance, A. Elkiss, A. Callahan, S. Liu & C. Chen, B. Gipp & J. Beel, and J. An have demonstrated in various ways that the actual similarity between co-cited documents correlates with their proximity in text, proposing methods like Citation Proximity Index (CPI) and section similarity to calculate author relationship strength [?]. M. Eto [?] and Zhao Rongying [?] proved that considering co-citation location in co-citation analysis can improve retrieval effectiveness and clustering results, facilitating deeper research and evaluation. Other scholars have focused on citation sentences themselves, using algorithms like TF-IDF, LDA, and C-Value to extract feature words or topics from citation sentences for content representation and similarity calculation [?].

As research progressed, scholars proposed comprehensive improved co-citation

analysis methods. Liu Shengbo divided co-citation sentences into multiple levels based on location proximity, calculated content similarity within each level to explore the influence of location proximity on co-citation content relevance, and thus more scientifically determined citation location weight values [**?**]. H.J. Kim et al. used chapter locations to construct fine-grained author co-citation matrices and calculated word frequency similarity of citation sentences, achieving favorable results in empirical studies on oncology in the PubMed database [**?**]. However, due to difficulties in obtaining full-text data, domestic research on improving co-citation analysis with content information still primarily utilizes bibliographic information [**?**].

As Y. Ding [**?**] and Zhao Rongying [**?**] noted, current research still lacks deep utilization of full-text. Empirical studies are particularly scarce domestically, with most research focusing on single aspects. Citation topic analysis often remains at the level of fixed term frequency, syntactic structure, and small-scale probabilistic models, rarely delving into semantic depth, resulting in shallow connections between co-cited documents based on superficial textual features. Moreover, few studies approach from the author level. Compared with document-level analysis, author co-citation analysis has greater practical value, but authors are more "multifaceted" than documents, making relationship strength based on citation counts prone to inaccurate inter-author connections and suboptimal clustering results. Therefore, improving author co-citation analysis by fully utilizing full-text information holds practical significance.

3 SPACA Method

Our improved SPACA method fully utilizes both the semantic information of citation sentences and the chapter location information where co-cited documents appear, calculates similarity based on these factors, and uses the maximum similarity value from an author's document set as the inter-author relevance strength to replace the traditional method's single reliance on co-citation frequency.

3.1 Full-Text Citation Mining and Extraction

SPACA implementation requires not only bibliographic information but also citation sentence text and location information of cited documents within citing documents, making text mining and extraction essential.

Using CNKI database's HTML full-text pages as an example, pages were collected locally via URLs. HTML pages contain full-text literature and numerous complex data tags, but their semi-structured characteristics enable data extraction and content analysis. A parser was developed to extract and store required data for SPACA. Citation sentences were located through `<a>` tags with `class_="sup"` or `<citation>` tags with `type_="reference"` (see Figure 1 [Figure 1: see original paper]). The extraction methods for citation location and content are as follows:

(a) **Citation location extraction**: In CNKI's existing HTML full-text format, main headings are encapsulated in `<h3>` tags and subheadings in `<h4>` tags. Based on citation tag locations, the parent `<p>` tag is retrieved and traversed forward to discover these heading tags until a `<h3>` tag is obtained (the `<h4>` tag is optional and replaced by `<h3>` when absent). After obtaining main and subheadings, proximity between two citation sentences can be determined when generating co-citation pairs to assist relevance strength measurement.

(b) **Citation sentence text extraction**: The basic method involves traversing forward and backward from the tag location, using regular expressions for judgment until adjacent tag content forms complete text within paragraph scope. The text is then split by periods, with the tail portion taken as the first half of the citation sentence; the second half is extracted similarly using adjacent tag text until a "。" is encountered, forming complete content text. Additionally, specific citation sentence identification rules were developed for various citation tag formats, positions, and multiple citations within small ranges.

Beyond this data, basic information of cited documents (author, title, etc.) was extracted from reference lists, with non-journal entries filtered out. The data structure for each reference is shown in Figure 2 [Figure 2: see original paper].

3.2 Domain Corpus-Based Word Embedding Model Training

This study involves content similarity calculation between citation sentences, requiring appropriate semantic representation. Considering the need for semantic connotation, conceptual representation, and efficiency/scalability in the learning process, we selected the Word2Vec word embedding model for pre-training based on domain corpora. Word2Vec is a shallow neural network language model that learns semantic knowledge from large text corpora in an unsupervised manner, proposed by T. Mikolov et al. and widely applied in natural language processing [**?**]. It includes CBOW and Skip-gram modes (see Figure 3 [Figure 3: see original paper]), modeling the relationship between a word and its context within a window, mapping to low-dimensional vector space to establish dense vector connections between each word and related words, enabling efficient, high-quality word vector training and optimization [**?**]. Compared with traditional vector space models based on word frequency, Word2Vec's key characteristic is learning contextual relationships between words, ensuring reliability in similarity measurement by maintaining thematic and semantic associations beyond mere word forms [**?**].

Word2Vec model quality depends on sufficient corpus training. These large-scale domain corpora can be derived from full-text documents in the target field, which after customized segmentation and preprocessing based on domain literature keywords, are used for model training to learn weight vectors of domain vocabulary at semantic and contextual levels. Since Skip-gram is overly

sensitive to low-frequency words, training employs the CBOW mode where context words jointly predict the center word and adjust weights collectively. The trained model effect is shown in Figure 4 [Figure 4: see original paper]: when inputting the term "学科化服务" (subject-based services), the model returns semantically/contextually related terms like "学科服务" (subject services), "嵌入式学科服务" (embedded subject services), "学科馆员" (subject librarians), "知识服务" (knowledge services), and "学科知识" (subject knowledge). This means even when writers use different forms of words with the same connotation in two sentences, the model can accurately determine sentence similarity.

3.3 Author Co-Citation Relationship Strength Algorithm

The SPACA author relevance strength algorithm comprehensively considers both the content similarity $Content\_similarity(a_M, b_N)$ and citation location similarity weight $P\_Weight(a_M, b_N)$ for each co-cited document pair a_M (M = 1, 2, ..., n) and b_N (N = 1, 2, ..., n) between co-cited authors A and B (see Equation 1), and takes the maximum similarity generated from the co-cited document set between A and B as their relevance strength (see Equation 2) to represent the maximum possible correlation between authors.

$$Similarity(a_M, b_N) = P\_Weight(a_M, b_N) \cdot Content\_similarity(a_M, b_N) \quad (1)$$

$$Relevance(A, B) = \max\{Similarity(a_M, b_N)\} \quad (2)$$

(1) Content similarity calculation: We utilize pre-trained word embedding vectors and cosine similarity algorithms. The basic principle involves summing the i-dimensional weight vectors of words contained in citation sentences x and y where a_M and b_N appear, forming sentence vectors W_x and W_y (see Equation 3), then calculating the cosine value to quantify content similarity (see Equation 4). Additionally, when content similarity is too low (< 0.2), the value is discarded (set to 0) to eliminate interference from unrelated citation sentence pairs.

$$W_{sentence} = \sum_{i=1}^{n} word_i \quad (where\ sentence = \{word_1, ..., word_n\}) \quad (3)$$

$$Content\_similarity(a_M, b_N) = \cos(W_x, W_y) = \frac{W_x \cdot W_y}{\sqrt{W_x^2}\sqrt{W_y^2}} \quad (4)$$

(2) Citation location similarity weight calculation: The algorithm employs: weight coefficient p when citations occur in the same chapter, multiplied by coefficient q if they further occur in the same subsection (Equation 5), with

weight = 1 if both are different. Location weights then weight the content similarity.

$$P\_Weight(a_M, b_N) = pos(x, y) = \begin{cases} 1, & (P_{chap.}(x) \neq P_{chap.}(y) \& P_{sec.}(x) \neq P_{sec.}(y)) \\ p, & (P_{chap.}(x) = P_{chap.}(y) \& P_{sec.}(x) \neq P_{sec.}(y)) \\ p \cdot q, & (P_{chap.}(x) = P_{chap.}(y) \& P_{sec.}(x) = P_{sec.}(y)) \end{cases} \tag{5}$$

To optimize parameters p and q, we explored the correlation between citation location and content similarity: 177,617 co-citation pairs from experimental full-text data were divided by proximity into article-level, chapter-level, and subsection-level types (TYPE), with content similarity calculated for each pair using the word embedding model to obtain similarity values (SIM). Since SIM values are not normally distributed but exhibit homogeneity of variance, Welch-ANOVA and non-parametric Kruskal-Wallis ANOVA with multiple comparisons were applied. Results showed significance far below 0.05 for distribution differences among groups (see Figure 5 [Figure 5: see original paper]), with non-parametric test $P < 0.05$ (see Figure 6 [Figure 6: see original paper]), indicating SIM value distribution differences across location levels.

Table 1 shows that as proximity increased from article-level to chapter-level and subsection-level, the proportion of low-similarity pairs (< 0.5) decreased from 48.80% to 42.64% and 31.82%, while average similarity cumulatively increased by 8.48% and 21.55%. Each proximity level upgrade increased average similarity by approximately 1.1 times. We conclude that similarity differs across three location levels, with co-citation pairs in smaller logical structures having less unrelated content and higher similarity. Referencing the similarity increase ratio from location proximity, parameters p and q can be approximately set to 1.1.

3.4 SPACA Method Characteristics

Compared with related research, our SPACA method has two main characteristics: First, it uses a domain corpus-based Word2Vec shallow neural network model to establish semantic and contextual associations between domain words for calculating citation sentence similarity, rather than relying on author co-occurrence, fixed terms, or small-scale topic probabilities, ensuring reliability of inter-author connection strength and adaptability in complex domains. Second, it comprehensively utilizes both citation location and thematic full-text information in a weighted fusion to comprehensively characterize inter-author relationship strength, ensuring more diverse considerations and stability in practical application.

4 Empirical Study

To demonstrate SPACA's effectiveness and explore its applicability in complex disciplines, we designed comparative experiments to examine differences be-

tween SPACA and traditional methods, focusing on: (1) accuracy in revealing inter-author connections; (2) identification of domain structures within disciplines.

4.1 Experimental Design

The experimental field was set as domestic LIS for two reasons: First, as a highly interdisciplinary field with complex scholar distribution, LIS presents greater difficulty that more easily reveals methodological differences. Second, unlike basic sciences, LIS terminology lacks controlled vocabularies like MeSH, exhibiting multifaceted and variable characteristics, making it ideal for proving the method's applicability and scalability.

The experimental framework (see Figure 7 [Figure 7: see original paper]) involves: (1) collecting online full-text page URLs using a crawler and downloading HTML pages to a local database; (2) extracting full-text information, cited document details, citation sentences, and location data; (3) constructing author co-citation matrices for both ACA and SPACA using different relevance strength calculations; (4) visually presenting matrix data through network and factor analysis; (5) comparing and discussing results.

4.2 Experimental Methods and Process

4.2.1 Data Source: CNKI Chinese journal database, focusing on LIS research from 2009-2019. The scope was limited to "Library and Information Science" under information technology, with journals restricted to core catalogs (SCI, EI, core, CSSCI, CSCD). Only top 500 most-cited documents annually were selected (142 documents for 2019 with $ $2 citations) due to higher quality and stronger author value representation.

4.2.2 Data Extraction and Matrix Construction: HTML pages were parsed to extract bibliographic information, citation sentences, and location data (see Section 3.1), yielding 23,572 cited document entries (see Figure 9 [Figure 9: see original paper]). For ACA, only basic bibliographic information was provided, while SPACA utilized all information plus Word2Vec model support trained on all corpus text (see Section 3.2). After author name disambiguation, author co-citation matrices were constructed for both methods: ACA used total co-occurrence counts as relevance strength, while SPACA used maximum similarity based on semantic and location proximity. This resulted in 10,684 author nodes, 132,267 ACA pairs, and 118,388 SPACA pairs.

4.2.3 Network Analysis: Visual network analysis helps understand author distribution and disciplinary knowledge structure. Using Gephi with authors as nodes and relevance strength as edge weights, we applied the Louvain community detection algorithm [**?**] for its efficiency and accuracy in large networks. After filtering (excluding modules with $ $1% of nodes or <2 core nodes with >15 citations, and pruning edges with <10 common neighbors using K-brace

algorithm [**?**]), the ACA network contained 1,529 nodes and 3,156 edges (modularity 0.697), while SPACA contained 1,502 nodes and 2,685 edges (modularity 0.793). SPACA showed more concentrated clustering and better cluster distinguishability (see Figures 11 [Figure 11: see original paper] and 12 [Figure 12: see original paper]).

4.2.4 Factor Analysis: To verify network analysis results, factor analysis was conducted on 127 authors with >15 citations. Both co-citation matrices were transformed into dissimilarity matrices using standardized Euclidean distance. After examining eigenvalues and scree plots, factors with eigenvalues $ $1 were extracted, with direct oblimin rotation applied due to low inter-factor correlations. ACA extracted 9 factors explaining 87.992% variance, while SPACA extracted 13 factors explaining 90.767% variance (see Table 5), indicating SPACA's more reasonable distribution.

4.3 Experimental Results

4.3.1 Network Analysis Results: Feature words were extracted from each module using TF-IDF and TextRank algorithms applied to titles and citation sentences of intra-module edges. Modules with similar authors and feature words were assigned identical IDs (see Tables 2 , 3 , and 4).

Key findings: 1. Some keywords grouped in single ACA modules were divided into multiple SPACA modules. For example, ACA's Module F "Mobile Library and Mobile Services" (383 nodes) was split into SPACA's F-1 "Mobile Library Services and Technology" (205 nodes) and F-2 "New Media Behavior and Services Research" (162 nodes), with corresponding author migrations (e.g., Kong Yun, Wang Baocheng) confirming this finer division. 2. Some misidentified nodes in ACA were corrected in SPACA. For instance, authors Ke Ping and Wu Jianzhong, originally in ACA's Module E "Research Library Data and Knowledge Services," were reassigned to Module A "Library Management" in SPACA, which better reflects their actual research focus on library management, evaluation, and transformation. 3. SPACA's finer granularity reveals emerging subfields with topological connections (see Figure 14 [Figure 14: see original paper]).

4.3.2 Factor Analysis Results: Factor analysis corroborated network analysis findings. Both methods identified major domains A-F, with SPACA providing finer subfield identification: - ACA's F4/F5 (library management) were subdivided into SPACA's F7 (reading promotion) - ACA's F1/F3 (smart libraries) were split into SPACA's F4 (smart library theory) and F7 (new technology application) - SPACA additionally identified think tank research (F8) as an emerging subfield within intelligence studies - Mobile library and new media services (ACA's F2) were separated into SPACA's F3 (mobile services) and F9 (new media behavior)

4.4 Discussion

Conclusions: 1. SPACA enables more accurate author relationship representation by incorporating semantic and location information, overcoming ACA's limitations of single metrics and spurious connections from popular topics. 2. SPACA achieves finer subfield identification, revealing emerging or interdisciplinary small domains that traditional methods miss, by establishing strong connections based on deep semantic and contextual relevance while eliminating weak, unrelated connections. 3. SPACA demonstrates applicability and effectiveness for large interdisciplinary fields, with Word2Vec proving effective for social science and interdisciplinary text similarity discrimination.

Limitations: 1. 9% of documents lacked HTML format, potentially omitting important literature. 2. Some citation sentence extraction anomalies occurred due to non-standard formats and informal citation practices. 3. Training corpus volume was insufficient, with some over-segmented terms creating noise. 4. Multi-disciplinary authors pose challenges for current text mining methods.

5 Conclusion

This study proposes the improved SPACA method that calculates co-cited author association strength using Word2Vec and location weighting. Experiments demonstrate SPACA's superior accuracy in identifying author connections and discovering detailed, three-dimensional subfield distributions, providing a reference path for utilizing domestic citation full-text for bibliometric research.

Future research should: (1) conduct deeper mining of citation semantics and location using ontologies and concept graphs; (2) incorporate more meaningful indicators (intensity, motivation, sentiment) and optimize parameters and clustering methods; (3) explore applications of citation content analysis in other informetric methods.

Acknowledgments

Thanks to researchers Yuan Junpeng (National Science Library, CAS), Yang Zhiping and Chen Yunwei (Chengdu Library and Information Center, CAS), Fu Wenqi (Fujian Normal University), and Yu Fan (Wuhan University), three anonymous reviewers, and the *Library and Information Service* editorial office for valuable suggestions.

References

[1] WHITE H D, GRIFFITH B C. Author co-citation: a literature measure of intellectual structure[J]. Journal of the American society for information science, 1981, 32(3): 163-171. [2] BAYER A E, SMART J C, MCLAUGHLIN G W. Mapping intellectual structure of a scientific subfield through author co-citations[J]. Journal of the American society for information science, 1990,

41(6): 444-452. [3] BOYACK K W, SMALL H, KLAVANS R. Improving the accuracy of co-citation clustering using full-text[J]. Journal of the American society for information science and technology, 2013, 64(9): 1759-1767. [4] DING Y, ZHANG G, CHAMBERS T. Content-based citation analysis: the next generation of citation analysis[J]. Journal of the association for information science and technology, 2014, 65(9): 1820-1833. [5] HU Z G. Full-text citation analysis methods and applications[D]. Dalian: Dalian University of Technology, 2014. [6] LIU S B, DING Y, TANG D L. Theories and methods of citation content analysis[J]. Information studies: theory & application, 2015, 38(10): 27-32. [7] LIU S, CHEN C. The differences between latent topics in abstracts and citation contexts of citing papers[J]. Journal of informetrics, 2013, 7(3): 583-592. [8] DING Y, SONG M, HAN J, et al. Entity metrics: measuring the impact of entities[J]. PLoS ONE, 2013, 8(8): 1-14. [9] ZHANG C Z, XU S R, LU C. Methods and empirical study on monitoring interdisciplinary phenomena using citation content[J]. Library and information service, 2016, 60(19): 108-115. [10] TEUFEL S, SIDHARTHAN A, TIDHAR D. An annotation scheme for citation function[C]//Proceedings of the 7th SIGdial workshop on discourse and dialogue. Stroudsburg: Association for Computational Linguistics, 2009: 80-87. [11] NAMBA H, OKUMURA M. Towards multi-paper summarization using reference information[C]//Proceedings of the 16th international conferences on artificial intelligence. San Francisco: Morgan Kaufmann Publishers, 1999: 926-931. [12] ZAFAR L, AHMED U, ISLAM M A. Citation context analysis for improved bibliometrics[C]//Proceedings of the 2013 conference of the North American chapter of the association for computational linguistics. Atlanta: Association for Computational Linguistics, 2013: 596-606. [13] ABU-JBARA A, EZRA J, RADEV D R. Purpose and polarity of citation: towards NLP-based bibliometrics[C]//Proceedings of the 2013 conference of the North American chapter of the association for computational linguistics: human language technologies. Atlanta: Association for Computational Linguistics, 2013: 596-606. [14] LEI S W, CHEN H H, HUANG Y, et al. Research on automatic identification of citation context in academic literature[J]. Library and information service, 2016, 60(17): 78-87. [15] ANGROSH M A, CRANEFIELD S, STANGER N. Context identification of sentences in related works sections using conditional random fields: towards intelligent digital libraries[C]//HUNTER J. Proceedings of the 10th annual joint conference on digital libraries. New York: ACM, 2010: 293-302. [16] ZHU X, TURNEY P, LEMIRE D, et al. Measuring academic influence: not all citations are equal[J]. Journal of the association for information science and technology, 2015, 66(2): 408-427. [17] SOMBATSOMPOP N, KOSITCHAIYONG A, MARKPIN T, et al. Scientific evaluations of citation quality of international research articles in the SCI database: Thailand case study[J]. Scientometrics, 2006, 66(3): 521-535. [18] CHEN C, LIU Z. Where are citations located in the body of scientific articles? A study of the distributions of citation locations[J]. Journal of informetrics, 2013, 7(4): 887-896. [19] LU C, DING Y, ZHANG C. Understanding the impact change of a highly cited article: a content-based citation analysis[J]. Scientometrics, 2013, 94(2): 651-673. [20] DING Y, LIU X, GUO C, et al. The distribution of references across texts:

some implications for citation analysis[J]. Journal of the American society for information science and technology, 2013, 64(3): 627-639. [21] HOU W R, LI M, NIU D K. Counting citations in texts rather than reference lists to improve the accuracy of assessing scientific contribution[J]. Bioessays, 2011, 33(10): 724-727. [22] ELKISS A, SHEN S, FADER A, et al. Blind men and elephants: what do citation summaries tell us about a research article?[J]. Journal of the American society for information science and technology, 2008, 59(1): 51-62. [23] CALLAHAN A, HOCKEMA S, EYSENBACH G. Contextual co-citation: augmenting co-citation analysis and its applications[J]. Journal of the American society for information science and technology, 2010, 61(6): 1130-1143. [24] GIPP B, BEEL J. Citation proximity analysis (CPA)-a new approach for identifying related work based on co-citation analysis[C]//LARSEN B, LETA J. Proceedings of ISSI2009-The 12th international conference on scientometrics and informetrics. Rio de Janeiro: BIREME/PAHO/WHO and Federal University of Rio de Janeiro, 2009: 571-575. [25] GIPP B. Citation proximity analysis-a measure to identify related work[D]. Magdeburg: Otto-von-Guericke University, 2006. [26] LIU S, CHEN C. The effects of co-citation proximity on co-citation analysis[C]//NOYONS E, NGULUBE P, LETA J. Proceedings of ISSI2011-The 13th international conference on scientometrics and informetrics. Durban: Leiden University and University of Zululand, 2011: 474-484. [27] AN J, KIM N, KAN M Y, et al. Exploring characteristics of highly cited authors according to citation location and content[J]. Journal of the association for information science and technology, 2017, 68(8): 1975-1988. [28] ETO M. Evaluations of context-based co-citation searching[J]. Scientometrics, 2013, 94(2): 651-673. [29] ZHAO R Y, GUO F J, ZENG X Q. Empirical study on location-based co-citation analysis[J]. Journal of the China society for scientific and technical information, 2016, 35(5): 492-500. [30] JEONG Y K, SONG M, DING Y. Content-based author co-citation analysis[J]. Journal of informetrics, 2014, 8(1): 197-211. [31] LU K, WOLFRAM D. Measuring author research relatedness: a comparison of word-based, topic-based, and author co-citation approaches[J]. Journal of the American society for information science and technology, 2012, 63(10): 1973-1986. [32] ZHU Q S, LENG F H. Topic identification of highly cited papers based on citation content analysis[J]. Journal of library science in China, 2014, 40(1): 39-49. [33] LIU S B, ZHANG C B, DING Y, et al. Improved co-citation analysis based on citation content and location[J]. Journal of the China society for scientific and technical information, 2013, 32(12): 1248-1256. [34] KIM H J, JEONG Y K, SONG M. Content-and proximity-based author co-citation analysis using citation sentences[J]. Journal of informetrics, 2016, 10(4): 954-966. [35] LI X X, SHAO Z Y. Author co-citation analysis incorporating content information: taking subject service research as an example[J]. Library and information service, 2016, 60(1): 98-104. [36] XIAO X, CHEN Y W, DENG Y. Citation network community division based on node content and topological structure[J]. Library and information knowledge, 2017(1): 89-97. [37] ZHANG Y M, MA X F, CHENG J J. Knowledge flow research integrating citation content and full-text citation analysis[J]. Journal of intelligence, 2015, 34(11): 50-54, 49. [38] DING Y, ZHANG G, CHAMBERS T, et

al. Content-based citation analysis: the next generation of citation analysis[J]. Journal of the association for information science and technology, 2014, 65(9): 1820-1833. [39] ZHAO R Y, ZENG X Q, CHEN B K. Full-text citation analysis: a new development in citation analysis[J]. Library and information service, 2014, 58(9): 129-135. [40] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space[EB/OL]. [2019-12-12]. https://arxiv.org/abs/1301.3781. [41] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality[C]//BURGES C. Advances in neural information processing systems. Lake Tahoe: Neural Information Processing Systems Foundation, 2013: 3136-3144. [42] TANG X B, ZHAI X P. Text knowledge fragment semantic indexing based on ontology and Word2Vec[J]. Information science, 2019, 37(4): 97-102. [43] LAW J, ZHUO H H, HE J, et al. LTSG: Latent topical skip-gram for mutually improving topic model and vector representations[C]//LAI J H. Pattern recognition and computer vision PRCV 2018. Cham: Springer, 2018: 375-387. [44] BLONDEL V D, GUILLAUME J, LAMBIOTTE R, et al. Fast unfolding of communities in large networks[J]. Journal of statistical mechanics: theory and experiment, 2008(10): P10008. [45] UGANDER J, BACKSTROM L, MARLOW C, et al. Structural diversity in social contagion[C]//GRAHAM R L, JOLLA L. Proceedings of the national academy of sciences of the United States of America. Washington: PNAS, 2012: 5962-5966. [46] YUAN B C, FANG S, LIU H Y. Research progress on author co-citation analysis methods[J]. Library and information service, 2009, 53(22): 80-84.

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