Advances in Citation-Based Topic Models: Post-print

Authors: Zou Lixue, Wang Li, Liu Xiwen

Date: 2023-07-26T00:00:00+00:00

Abstract

[Purpose/Significance] As probabilistic topic model algorithms continue to be improved and extended, this study investigates existing citation-based topic models developed domestically and internationally, analyzes and compares the generative processes and algorithms of different models, and explores the applications and extensible research directions of citation-based topic models in scientific text analysis.

[Method/Process] Relevant literature on citation-based topic models from domestic and international sources was retrieved from the Web of Science and CNKI databases. After manual review, representative works were selected, and the citation-based topic models in these studies were compared and analyzed in terms of modeling philosophy, generative process, parameter estimation, and inference algorithms.

[Results/Conclusion] Currently, citation-based topic models developed domestically and internationally primarily include: topic models for research topics and citation distributions, topic models for investigating relationships between cited and citing topics, and citation content-based topic models. The incorporation of citation information into topic models enables the acquisition of more complete topic content and identification of important documents under specific topics, as well as the recognition of relationships and influences between topics in citing and cited documents. Existing models are predominantly extensions based on Probabilistic Latent Semantic Analysis (PLSA) and Latent Dirichlet Allocation (LDA) topic models. Future extensible research directions include citation content-based topic models, model performance optimization and evaluation methods, and applied research on models.

Full Text

Research Advances in Citation-Based Topic Models

Zou Lixue[1],[2], **Wang Li**[1],[2], **Liu Xiwen**[1],[2] [1] National Science Library, Chinese Academy of Sciences, Beijing 100190 [2] Department of Library, Information and Archives Management, School of Economics and Management, University of Chinese Academy of Sciences, Beijing 100190

Abstract

[**Purpose/Significance**] Probabilistic topic model algorithms continue to be improved and extended. This paper examines existing citation-based topic models developed domestically and internationally, analyzes and compares the generative processes and algorithms of different models, and explores the application of citation-based topic models in scientific text analysis and directions for future research. [**Method/Process**] We retrieved relevant literature on citation-based topic models from the Web of Science and CNKI databases. After manual screening, we selected representative papers and conducted a comparative analysis of the modeling concepts, generative processes, parameter estimation, and inference algorithms of the citation-based topic models proposed in these studies. [**Result/Conclusion**] Current citation-based topic models mainly include: models focusing on topic-citation distributions, models studying relationships between cited and citing topics, and citation content-based topic models. Introducing citation information into topic models enables the extraction of more complete topic content and identification of important documents under specific topics, while also revealing relationships and influences between topics in citing and cited documents. Most existing models extend the Probabilistic Latent Semantic Analysis (PLSA) and Latent Dirichlet Allocation (LDA) frameworks. Future research may extend to incorporating citation content into topic models, performance optimization and evaluation methods, and application studies.

Keywords: topic model, citation, topic detection, citation context

---

1. Introduction

In the information age, various types of information, particularly text resources, are growing explosively. This continuous accumulation has led to increasingly large text datasets. Scientific literature, in particular, is expanding exponentially. As the primary carrier of knowledge and the cumulative form of knowledge development, scientific literature contains substantial thematic information that reveals the evolution of disciplines. Extracting latent topics and their evolution from complex, large-scale text data can help researchers and decision-makers identify research themes, quickly grasp the trajectory of scientific development, and track the evolution of topics and knowledge flow in scientific fields.

In recent years, probabilistic topic models [1] have emerged in text mining as statistical models used in machine learning and natural language processing to discover latent topics in document collections, enabling semantic text mining. By introducing the concept of topic space, these models achieve dimensionality reduction of documents in topic space while extracting latent semantics from document collections, providing concise representations for large-scale datasets [2]. By mining deep, latent semantic information, probabilistic topic models can better extract valuable latent topic distributions from scientific literature. This new latent semantic space fills the gap between documents and words, offering researchers a novel method for identifying topics in large text corpora and becoming a highly active research area. With advances in natural language processing, probabilistic topic models have been widely applied to topic identification and evolution analysis.

Among the many algorithms for probabilistic topic models, the two classic approaches are Probabilistic Latent Semantic Analysis (PLSA) [3] and Latent Dirichlet Allocation (LDA) [1]. LDA has been extensively applied and extended due to its solid mathematical foundation and flexibility. However, as research has progressed, scholars have identified limitations in LDA, such as its assumption of document exchangeability (documents have no sequential order) and topic exchangeability (topics have no hierarchical or sequential relationships) [4]. In reality, most corpora—especially scientific literature—are interconnected in various ways rather than being independent, and topics in documents often have sequential and hierarchical relationships. These assumptions fail to model topic relationships and should be reconsidered when analyzing corpora.

Researchers have subsequently improved and extended these models by incorporating temporal factors to model topic evolution [5] and introducing author metadata to build author-topic models [6]. A research paper is more than just a bag of words; it contains additional structural information. Citations, as crucial inheritance elements in scientific literature, contain less noise and can demonstrate the influence of one document on another and the connections between topics. Therefore, scholars have introduced citation relationships into topic models to improve and extend these algorithms. This paper provides an in-depth analysis and comparison of existing citation-based topic models, elaborates on their generative processes and algorithms, identifies existing problems, and explores their applications in scientific text analysis and future research directions.

2. Data Sources and Methods

To comprehensively analyze the latest research progress on citation-based topic models, this study selected the Web of Science Core Collection and CNKI as data sources for retrieving English and Chinese literature respectively. The data collection process involved: (1) Searching English literature using the query: (("topic model" *AND (citation OR citations OR cited OR citing OR reference*)) OR (("Bayesian model" *OR "probabilistic model"*) NEAR (citation OR citations

OR cited OR citing OR reference*)) OR ((model* NEAR topic*) NEAR (citation OR citations OR cited OR citing))). Document types were limited to articles, proceeding papers, reviews, and editorial materials. The search was conducted up to January 28, 2019, yielding 381 English publications, including 229 articles and 149 conference papers. In terms of major contributing countries, U.S. scholars published 129 papers, while Chinese scholars published 93 . (2) Searching Chinese literature using the query: (model* (citation + reference*) (topic + probability + Bayesian)). Chinese document types were limited to journal articles, conference papers, and dissertations, also up to January 28, 2019, resulting in 155 Chinese publications. (3) Manually reviewing the literature from both datasets and filtering out less relevant papers, we finally selected 26 representative studies that introduced citations into topic models for analysis.

3. Analysis of Citation-Based Topic Model Progress

Based on the 26 representative studies, we categorize the proposed citation-based topic models into three research perspectives: models focusing on topic-citation distributions, models studying relationships between cited and citing topics, and citation content-based topic models. We provide detailed analysis and comparison of these citation-based topic models.

3.1 Topic and Citation Distribution Models This category of topic models introduces citations to study the topic distribution of citations, extracting distributions of citation documents and topics, or treating citations as words to obtain joint topic-citation distributions. Representative models include PHITS, PLSA+PHITS, Mixed-membership model, cc-LDA, cp-LDA, CitationLDA++, Citation Author Topic Model (CAT), Citation Topic Model (CT), Citation-Content-LDA, and Citation Network Topic Model (CNTM).

Early work on jointly modeling citations and textual content extended the PLSA model. D. Cohn et al. [7] proposed the PHITS model by incorporating the Hyperlink-Induced Topic Search (HITS) algorithm into PLSA. This model assumes that citations are generated through a process similar to PLSA, but while PLSA models words in documents, PHITS models citations. It introduces a topic space between documents and citations, assuming that citing documents follow a multinomial distribution over topics with specific parameters for cited documents. The model uses the EM algorithm for maximum likelihood estimation of parameters. PHITS demonstrated that incorporating citations improves document classification, reveals the likelihood of a citation being cited under specific topics, and can compute topic probability distributions for citations to identify topic-specific references. However, it cannot extract topic-word distributions.

Subsequently, D. Cohn and T. Hofmann [8] proposed the PLSA+PHITS joint topic model, which utilizes the same factorization for both PLSA and PHITS while sharing the same document-topic mixture distribution, thus introducing a common latent topic space that can simultaneously extract topic-word and topic-

citation distributions. This model employs EM methods for parameter inference and produces more stable topics, achieving better classification performance than PLSA or PHITS alone.

Later, scholars extended LDA using textual and citation data. E. Erosheva et al. [9] proposed the Mixed-membership model (later also called Link-LDA), which treats a document as both a bag of words and a bag of citations. Citations and words are generated through the same process, sharing the same document-topic distribution, with LDA used to generate topic-word and topic-citation distributions separately. Using this model, Erosheva et al. identified topics in 12,036 PNAS life sciences articles, achieving finer-grained classification than the dataset' s own disciplinary categories.

Other researchers have calculated citation topic distributions from the perspective of citing documents. T. Nguyen et al. [10] developed the CitationLDA++ model, which uses LDA to obtain topic-word distributions as prior knowledge for model inference. To compute citation topic distributions, the model obtains citing document sets from the citation network, extracts top-k topics from prior knowledge for each document, and uses Hellinger distance to calculate similarity between citing document topics and top-k topics.

Building on Link-LDA, Y. Li et al. [11] proposed the cc-LDA and cp-LDA models. The cc-LDA model is similar to Link-LDA but differs in processing each citation by additionally extracting citation-word distributions beyond topic-citation distributions. The cp-LDA model introduces citation location information, dividing an article into two parts– "Introduction and Related Work" and "Others" –with citation locations generated from a beta distribution.

Additionally, some scholars have attempted to jointly model citations with other document metadata such as authors. Notable examples include the Citation Author Topic (CAT) model by Y. Tu et al. [12], which jointly models words, authors, and citations, and the Collective Topic Model by Z. Lu et al. [13], which incorporates authors, publication venues, and citation relationships into PLSA to evaluate paper influence based on topics using co-citation relationships.

Other researchers have extracted citation topic distributions. Z. Guo et al. [14] constructed the CT model, which first extracts document-citation distributions and then citation-topic distributions, capturing indirect citation relationships through random walks on directed graphs. Evaluated on 9,998 documents from the Cora dataset, the CT model outperformed PLSI, PHITS, LDA, and PLSA+PHITS in topic clustering.

X. Huang et al. [15] designed the Model for Topic-sensitive Influential Paper Discovery (MTID), which extracts topic distributions from citing documents and models the importance of papers across different topics.

From a hierarchical perspective, H. Zhou et al. [16] proposed the Citation-Content-LDA model with two layers: the first layer uses citations to generate parent topics representing citation clusters, and the second layer extracts child

topics from each parent topic. Since citation relationships are fewer than words, this model reduces computational complexity.

While the above models are parametric, researchers have also developed non-parametric models. K. W. Lim and W. Buntine [17-18] constructed the Citation Network Topic Model (CNTM), extending non-parametric models based on the Poisson Mixed-topic Link Model (PMTLM) [19] and Author-Topic (AT) model [20] by incorporating authors, citations, and textual content.

H. Bai et al. [21] proposed the Neural Relational Topic Model (NRTM), which simultaneously leverages latent correlations between topics and citation networks.

As shown in Table 1 :

3.2 Cited-Citing Topic Relationship Models This category of topic models focuses on the relationship between topics in citing documents and cited documents. By choosing whether to sample topics from cited document topic distributions or by sharing the same topic-citation distribution, these models reveal how cited document topic distributions influence citing document topic distributions. Representative models include Copycat, Citation Influence Model (CIM), Pairwise Link-LDA, Link-PLSA-LDA, Inheritance Topic Model (ITM), Relational Topic Model (RTM), cite-LDA, cite-PLSA-LDA, TERESA, Bernoulli Process Topic Model (BPT), Bi-Citation-LDA, RefTM, and Latent Topical Authority Indexing (LTAI).

L. Dietz et al. [22] proposed the Copycat and CIM models. In Copycat, each topic in a citing document is drawn from a mixture of its citations' topics, and each word in the citing document is associated with a cited document. Thus, the cited document' s topic distribution influences the citing document' s topics, modeling dependencies between cited and citing documents, co-citations, and bibliographic coupling. However, this model forces every word in the citing document to associate with a cited document, which does not always hold in practice, introduces new words into cited document topics, and cannot reveal innovative or emerging topics.

The CIM model overcomes these limitations by allowing citing documents to choose whether to sample topics from cited document topic distributions or from their own distributions via a Bernoulli process. Empirical studies show that CIM outperforms Copycat in predictive performance, though it only handles simple bipartite graphs and cannot process complex citation networks.

M. Kim et al. [23] extended CIM by incorporating PageRank values of cited documents to calculate citation strength, using this strength to set thresholds for building weighted citation networks for topic diffusion analysis.

Subsequently, Z. Guo et al. [24] proposed the Bernoulli Process Topic Model (BPT), which considers that each paper plays two distinct roles: as a document itself and as a cited document. When treated as a cited document, topics are sampled as in LDA; for the document' s own research topics, the distribution

is a mixture of its citations' topic distributions, with the multi-level structure of the citation network captured through a random Bernoulli process. BPT outperforms LDA, Link-LDA, Copycat, and CIM in perplexity.

T. Masada et al. [25] proposed the TERESA model, which similarly combines textual topic information with citation relationships to discover high-quality topics, predict citation strength, and identify inheritance and evolution relationships within topics.

Some models incorporate author metadata alongside citation relationships. J. Kim et al. [32] proposed the LTAI model, which introduces author distributions into each citation relationship, with parameters for calculating citation influence following a Dirichlet distribution.

T. Dai et al. [33] developed the Topic Model with Author Link Community for Citation Recommendation, incorporating author and citation information to obtain author-citation distributions, co-author distributions, and relationships between citing and cited documents. This model outperforms Link-PLSA-LDA and RTM in citation recommendation performance.

As shown in Table 2 :

3.3 Citation Content-Based Topic Models H. Small [34] defined citation context in 1982 as the textual content surrounding reference markers. Researchers have conducted exploratory studies on topic extraction and clustering using citation content. B. Aljaber et al. [35] found that topical terms from citation content can effectively identify research themes for document clustering. L. Bornmann et al. [36] discovered that keywords extracted from citation content are semantically closer to research content than those from titles and abstracts. M. Doslu et al. [37] used citation content to build directed, topic-indexed citation networks and applied the HITS algorithm to rank papers by topic, identifying topic-specific important documents. S. Liu et al. [38] used LDA to identify citation content topics and found that citation content topics cover broader scopes than citation-based topics. Yang Chunyan et al. [39] used a Labeled-LDA combined topic model to extract citation content topics, finding that citation content can eliminate "noise" in full text while covering as many topics as possible. X. Liu et al. [40] employed Labeled-LDA to construct a citation content-based network graph between cited and citing documents, addressing issues of citation reasons and contribution values.

S. Kataria et al. [41] introduced citation content into Link-LDA and Link-PLSA-LDA, proposing the cite-LDA and cite-PLSA-LDA models. These models assume that word and cited document selection in citation content are independent, simultaneously sampling topic-word and topic-citation distributions for words in citation content. In cite-PLSA-LDA, citing documents follow the cite-LDA generative process, while cited documents use PLSA to extract topic-citation distributions. Performance evaluation experiments on 3,312 CiteSeer

documents showed that cite-LDA performs similarly to Link-LDA and Link-PLSA-LDA, while cite-PLSA-LDA outperforms the other three models.

4. Conclusion and Outlook

This paper systematically reviews the development status of citation-based topic models proposed by scholars in recent years, providing detailed analysis and comparison of each model' s concepts and generative processes. This work offers references for method selection in topic identification and evolution analysis in information science, and provides insights for further improvement and refinement of these models.

The research on citation-based topic models has primarily focused on model extension, improvement, and optimization in recent years. Research perspectives include topic-citation distributions and relationships between cited and citing topics. Vocabulary sources for topic identification include both citing and cited documents, and with the development of full-text analysis, citation content-based topic models have emerged. Existing research shows that citation content contains richer semantic information related to topics than citation analysis alone. Unlike citations, the unit of analysis is no longer the document but knowledge elements within documents, offering new interpretations of node attributes and relationships. However, research incorporating citation content into topic modeling remains relatively limited.

Existing models demonstrate that introducing citation information improves topic identification by simultaneously extracting accurate keyword and key document distributions, yielding more complete topic content, improving document classification, revealing relationships and influences between topics in citing and cited documents, and providing important quantitative analysis for topic evolution. However, several issues remain. As our comparative analysis shows, topic-citation distribution models only model document citation features without modeling topic relationships between citing and cited texts, failing to demonstrate thematic inheritance and having overly simple generative processes that cannot explain citation structures and phenomena in corpora. Cited-citing topic relationship models can reveal topic-level associations, but vocabulary for topic identification primarily comes from titles and abstracts of cited documents, with insufficient research using citation content to represent cited document topics.

Clearly, citation-based topic modeling requires further development. Future trends may extend in the following directions.

4.1 Research on and Extension of Citation Content-Based Topic Models Current research on citation content topic models is limited, primarily applying existing mature models to identify citation content topics. As citation content analysis and natural language processing intersect more deeply, semantic and automatic analysis of citation topics using citation content will advance, strengthening citation analysis depth, particularly in semantic under-

standing. Furthermore, with internet technology development and the open access movement, full-text data has become more accessible and parseable, containing richer textual information. For example, XML-formatted full-text data uses structured markup to tag citation content, further promoting citation content analysis. How to extract citation content textual elements from full-text data for text mining and semantic analysis, and how to build appropriate topic models to extract and identify latent topic structures and evolution information, require in-depth exploration.

4.2 Performance Optimization and Evaluation Methods for Citation-Based Topic Models Citation-based topic models require more efficient algorithms for performance optimization. Most current models transform term or citation space into topic space, with existing models primarily extending PLSA and LDA and using EM algorithms, variational inference, and Gibbs sampling for parameter estimation and inference. Additionally, these models, particularly cited-citing topic relationship models, introduce new latent variables that increase runtime. For example, in Link-LDA and Link-PLSA-LDA, the time complexity of single Gibbs sampling iterations is linearly related to the number of links in the corpus, limiting scalability when link counts are large. How to design performance optimization methods for these models and balance complexity reduction with topic quality requires further research. Moreover, current evaluation of citation-based topic models uses perplexity comparison and recall rates, with some models employing AUC, precision, topic coherence, and F1-Score. Evaluation methods could be expanded from multiple perspectives.

4.3 Application Studies of Citation-Based Topic Models Citation-based topic models are currently applied mainly to topic identification, topic evolution, text clustering, link prediction, and citation recommendation. These models essentially provide probabilistic modeling methods for text with link information and can be applied to various text mining tasks. Besides scientific literature, some scholars have applied them to web pages and blog data, demonstrating their extensibility to diverse linked corpora. However, more in-depth research on application effectiveness evaluation is needed.

Author Contributions

Zou Lixue: Designed the research framework, wrote and revised the paper. **Wang Li:** Revised the paper. **Liu Xiwen:** Supervised research direction and revised the paper.

References

[1] BLEI D M, NG A Y, JORDAN M I. Latent dirichlet allocation[J]. Journal of machine learning research, 2003, 3(Jan): 993-1022.

[2] Zhang Jinsong. Research on document retrieval technology based on citation context analysis[D]. Dalian: Dalian Maritime University, 2013.

[3] HOFMANN T. Probabilistic latent semantic analysis[C]//Association for Uncertainty in Artificial Intelligence. Fifteenth conference on uncertainty in artificial intelligence. Stockholm: Morgan Kaufmann, 1999: 289-296.

[4] Fan Yunman, Ma Jianxia. Review of emerging topic detection technology in specific domains using LDA[J]. Modern Library and Information Technology, 2012, 28(12): 58-65.

[5] KAWAMAE N. Trend analysis model: trend consists of temporal words, topics, and timestamps[C]//International conference on web search and data mining. Hong Kong: Association for Computing Machinery, 2011: 317-326.

[6] ROSEN-ZVI M, GRIFFITHS T, STEYVERS M, et al. The author-topic model for authors and documents[C]//Association for Uncertainty in Artificial Intelligence. Proceedings of the 20th conference on uncertainty in artificial intelligence. Banff: Association for Uncertainty in Artificial Intelligence Press, 2012: 487-494.

[7] COHN D, CHANG H. Learning to probabilistically identify authoritative documents[C]//Association for Computing Machinery. Proceedings of the seventeenth international conference on machine learning. San Francisco: Morgan Kaufmann Publishers, 2000: 167-174.

[8] COHN D, HOFMANN T. The missing link: a probabilistic model of document content and hypertext connectivity[C]//Neural Information Processing Systems Foundation. Advances in neural information processing systems 13. Cambridge: NIPS, 2000: 430-436.

[9] EROSHEVA E, FIENBERG S, LAFFERTY J. Mixed-membership models of scientific publications[J]. Proceedings of the National Academy of Sciences of the United States of America, 2004, 101(1): 5220-5227.

[10] NGUYEN T, DO P. Citation LDA plus: an extension of LDA for discovering topics in document networks[C]//Association for Computing Machinery. International symposium on information and communication technology. Danang City: Association for Computing Machinery, 2018: 31-37.

[11] LI Y, HE J, LIU H. Topic analysis and influential paper discovery on scientific publications[C]//14th web information systems and applications conference. Liuzhou: IEEE, 2017: 68-73.

[12] TU Y, JOHRI N, ROTH D, et al. Citation author topic model in expert search[C]//Association for Computational Linguistics. International conference on computational linguistics: posters. Beijing: Association for Computational Linguistics, 2010: 1265-1273.

[13] LU Z, MAMOULIS N, CHEUNG D. A collective topic model for milestone paper discovery[C]//Association for Computing Machinery. Proceedings of the 37th international ACM SIGIR conference on research and development in information retrieval. Boston: Association for Computing Machinery, 2014: 1019-1022.

[14] GUO Z, ZHU S, CHI Y, et al. A latent topic model for linked documents[C]//Association for Computing Machinery. Proceedings of the 32nd international ACM SIGIR conference on research and development in information retrieval. Boston: Association for Computing Machinery, 2009: 720-721.

[15] HUANG X, CHEN C, PENG C, et al. Topic-sensitive influential paper discovery in citation networks[C]//Pacific-Asia conference on knowledge discovery & data mining. Melbourne: Springer, 2018: 16-28.

[16] ZHOU H, HUIMINY, ROLAND H. Topic discovery and evolution in scientific literature based on content and citations[J]. Frontiers of information technology & electronic engineering, 2017, 18(10): 1511-1532.

[17] LIM K W, BUNTINE W. Bibliographic analysis on research publication networks[C]//Asian conference on machine learning. Nha Trang City: Springer, 2014, 39: 142-158.

[18] LIM K W, BUNTINE W. Bibliographic analysis with the citation network topic model[C]//Association for Computing Machinery. Proceedings of the 18th ACM conference on information and knowledge management. Hong Kong: Association for Computing Machinery, 2009: 957-966.

[19] ZHU Y, YAN X, GETOOR L, et al. Scalable text and link analysis with mixed-topic link models[C]//Association for Computing Machinery. Proceedings of the 19th ACM SIGKDD international conference on knowledge discovery and data mining. Chicago: Association for Computing Machinery, 2013: 473-481.

[20] YAN L, NICULESCU-MIZIL A, GRYC W. Topic-link LDA: joint models of topic and author community[C]//Association for Computing Machinery. Proceedings of the 26th annual international conference on machine learning. Montreal: Association for Computing Machinery, 2009: 665-672.

[21] BAI H, CHEN Z, LYU M. Neural relational topic models for scientific article analysis[C]//Association for Computing Machinery. Proceedings of the 27th ACM international conference on information and knowledge management. Torino: Association for Computing Machinery, 2018: 27-36.

[22] DIETZ L, BICKEL S, SCHEFFER T. Unsupervised prediction of citation influences[C]//Association for Computing Machinery. Proceedings of the 24th international conference on Machine learning. Corvallis: Association for Computing Machinery, 2007: 233-240.

[23] KIM M, BAEK I, SONG M. Topic diffusion analysis of a weighted citation network in biomedical literature[J]. Journal of the Association for Information Science and Technology, 2018, 69(2): 329-342.

[24] GUO Z, ZHANG Z M, ZHU S, et al. A two-level topic model towards knowledge discovery from citation networks[J]. IEEE transactions on knowledge & data engineering, 2014, 26(4): 780-794.

[25] MASADA T, TAKASU A. Extraction of topic evolutions from references in scientific articles and its GPU acceleration[C]//Association for Computing Machinery. International conference on information and knowledge management. Maui: Association for Computing Machinery, 2012: 1522-1526.

[26] NALLAPATI R M, AHMED A, XING E P, et al. Joint latent topic models for text and citations[C]//Association for Computing Machinery. Proceedings of the 14th ACM SIGKDD international conference on knowledge discovery and data mining. Las Vegas: Association for Computing Machinery, 2008: 542-550.

[27] CHANG J, BLEI D M. Hierarchical relational models for document networks[J]. Annals of applied statistics, 2010, 4(1): 124-150.

[28] TAN L S L, HUI C A, TIAN Z. Topic-adjusted visibility metric for scientific articles[J]. The annals of applied statistics, 2016, 10(1): 1-31.

[29] HE Q, CHEN B, PEI J, et al. Detecting topic evolution in scientific literature: how can citations help?[C]//Association for Computing Machinery. Proceedings of the 18th ACM conference on information and knowledge management. Hong Kong: Association for Computing Machinery, 2009: 957-966.

[30] SHEN J, SONG Z, LI S, et al. Modeling topic-level academic influence in scientific literatures[C]//Association for the Advancement of Artificial Intelligence. The workshops of the thirtieth AAAI conference on artificial intelligence. Phoenix: Association for the Advancement of Artificial Intelligence, 2016: 1-7.

[31] HUANG L, LIU H, HE J, et al. Finding latest influential research papers through modeling two views of citation links[C]//Asia-Pacific web conference, Web technologies and applications. Suzhou: Springer, 2016: 555-566.

[32] KIM J, KIM D, OH A. Joint modeling of topics, citations, and topical authority in academic corpora[J]. Transactions of the Association for Computational Linguistics, 2017, 5(1): 191-204.

[33] DAI T, ZHU L, CAI X, et al. Explore semantic topics and author communities for citation recommendation in bipartite bibliographic network[J]. Journal of ambient intelligence and humanized computing, 2018, 9(5): 957-975.

[34] SMALL H. Citation context analysis[J]. Progress in communication sciences, 1982, 3(9): 287-310.

[35] ALJABER B, STOKES N, BAILEY J, et al. Document clustering of scientific texts using citation contexts[J]. Information retrieval, 2010, 13(2): 101-131.

[36] BORNMANN L, HAUNSCHILD R, HUG S E. Visualizing the context of citations referencing papers published by Eugene Garfield: a new type of keyword co-occurrence analysis[J]. Scientometrics, 2018, 114(2): 427-437.

[37] DOSLU M, BIGNOLHO H. Context sensitive article ranking with citation context analysis[J]. Scientometrics, 2016, 108(2): 653-672.

[38] LIU S, CHEN C. The differences between latent topics in abstracts and citation contexts of citing papers[J]. Journal of the American Society for Information Science and Technology, 2013, 64(3): 627-639.

[39] Yang Chunyan, Pan Youneng, Zhao Li. Research on document topic extraction based on semantic and citation weighting[J]. Library and Information Service, 2016, 60(9): 131-138.

[40] LIU X, ZHANG J, GUO C. Full-text citation analysis: a new method to enhance scholarly networks[J]. Journal of the American Society for Information Science and Technology, 2013, 64(9): 1852-1863.

[41] KATARIA S, MITRA P, BHATIA S. Utilizing context in generative bayesian models for linked corpus[C]//Association for Computing Machinery. Twenty-fourth AAAI conference on artificial intelligence. Atlanta: Association for Computing Machinery, 2010: 1340-1345.

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