Visualization of the Semantic Knowledge Landscape of Editing and Publishing Domain in China

Shuang Zhang[1], Feifan Liu[1], Li Wang[2,3], Haoxiang Xia*[1,3]

[1] School of Economics and Management, Dalian University of Technology, No. 2 Linggong Road, Dalian 116024, China
[2] Institute of Science and Technology Information of China, No.15 Fuxing Road, Beijing 100038, China
[3] Key Laboratory of Rich-Media Knowledge Organization and Service of Digital Publishing Content Institute of China, No.15 Fuxing Road, Beijing 100038, China
hxxia@dlut.edu.cn

Abstract. The rapid advancement of information technology is reforming the ecosystem of the editing and publishing industry. Under the demand for building novel publishing patterns, it is critical to map out the research trends and evolutionary trajectories of the field of editing and publishing comprehensively. In this study, drawing on large-scale journal papers, we construct the knowledge landscape of editing and publishing research field in China, by manifold learning algorithm UMAP based on the deep semantic associations between papers learned by Doc2vec, to visualize the static and diachronic structure of this field. Firstly, with the Gaussian kernel density function to characterize the heterogeneity of spatial distribution of papers, we identify core research topics in the editing and publishing domain. Then, by respectively constructing cumulative and sliced dynamic knowledge maps, we find that over the past 40 years, the research scope has continued to expand, evidenced by the emergence of new research topics along the edges of the map, and meanwhile, topics within the field keep merging and fusing. Furthermore, according to the pattern of research hotspot transitions, the development is roughly divided into four stages. These findings offer valuable insights for researchers in the publishing field and scientific policymakers.

Keywords: Editing and Publishing, Knowledge Map, Document Embedding, Evolutionary Analysis

1 Introduction

The past few decades have witnessed the great progress of the Chinese editing and publishing industry, marked by an expanding scale and growing international influence. Recently, the rapid advance of information technology has disrupted the ecosystem of the publishing industry, with digital technologies bringing integrated publishing, blockchain technology revolutionizing copyright protection, and large language models reforming content production. Faced with this revolution, for the Chinese editing and publishing industry, one of the largest publishing strengths worldwide, it is critical to

© The Author(s) 2024
C. Bai et al. (eds.), *Proceedings of 2023 China Science and Technology Information Resource Management and Service Annual Conference (COINFO2023)*, Advances in Economics, Business and Management Research 293,
https://doi.org/10.2991/978-94-6463-498-3_7

deeply understand the developing patterns of editing and publishing in China. The research field of Chinese editing and publishing, from the development of ancient text compilation to recent digitalized editing and publishing, has accumulated a wealth of theories and practical experience, that could provide meaningful guidance for the future of the Chinese publishing industry. However, as the field of Chinese editing and publishing studies has evolved, due to the vast number of materials and literature, the diversity of researchers, and the proliferation of research institutions, it has become increasingly challenging to fully understand the current research status and grasp the development trend. Therefore, it is necessary to utilize text mining techniques to delineate knowledge structures embedded in massive literature, beneficial for us to better navigate industrial transformations and construct new publishing patterns.

Drawn on 20,000 articles extracted from CNKI, Xu et.al. conducts an analysis of the development trends, research hotspots, and cutting-edge topics in Chinese editing and publishing studies using Citespace[1]. Wang et.al. utilizes high-impact articles from the CSSCI-indexed journals in CNKI and employ the "first-order linear homogeneous difference equation" to categorize the development of China's publishing discipline from 1998 to 2020 into two distinct phases. Additionally, they perform keyword co-occurrence analysis and social network analysis to extract key research hotspots[2]. Although some researchers review parts of this field, such as digital publishing[3], knowledge service[4], and integrative publishing[5], there is still a lack of systematic reviews and bibliometric analyses of the research landscape of the editing and publishing field.

Scientific knowledge landscape visualization has raised much attention. Tools like Citespace[6] and VOSviewer[7] are extensively applied. Methods such as citation networks[8], co-word networks[9], and topic modeling[10] are frequently employed to uncover the knowledge structure. However, these approaches typically remain at the level of keyword and external features, failing to delve into the semantic space of the literature. In recent years, representation learning algorithms, outstanding in learning the semantic features in the large-scale corpus, have enhanced scientific and technical information analysis[11]. Some researchers have proposed a framework for constructing semantic maps (SMAP) from the diverse bibliometric corpora by combining embedding and manifold learning algorithms[12].

In this study, drawing on large-scale Chinese journal papers, we employ the methodological framework of SMAP [12] to comprehensively visualize the static and dynamic knowledge structure of the field of editing and publishing. In particular, this research utilizes document embedding representation learning to capture the implicit semantic features embedded within academic literature abstracts. With the aid of manifold learning algorithms, we convert the high-dimensional semantic proximity into a lower-dimensional space and characterize the knowledge structure using kernel density estimation. Subsequently, the knowledge map is annotated and analyzed using keywords in conjunction with ChatGLM-4, an Open Bilingual Chat LLM[13]. By analyzing the dynamics of the knowledge landscape, the evolutionary process and the hotspot shift are revealed.

2 Method and Data

2.1 Method

This study aims to detect and visualize the static knowledge structure and diachronic evolution of the knowledge landscape of the field of editing and publishing based on document semantics. Here, we employ a new visualization method that integrates embedded representation learning and manifold learning algorithms, SMAP (Semantics Map). The core idea of SMAP is shown in Figure 1. First, the document embedding algorithm captures the implicit semantic features in academic literature, and the manifold learning algorithm projects the high-dimensional semantic structure into a low-dimensional space. Then the obtained plane is characterized by kernel density estimation (KDE) and keyword clouds. The global knowledge landscape is hereto generated. For further evolution, by identifying changes in the distribution density and scope of the diachronic knowledge landscape, we analyze the process of domain formation and the shift in the focus of research. The specific steps mainly include document semantics learning, dimensionality reduction, topic structure characterization and annotation, and evolution analysis.

(1) **Semantic representation learning**. In this study, we primarily utilize the titles and abstracts of scientific literature as the core representation of the research content. Recently, various technical solutions have emerged for the deep semantic representation of text, such as the classic Doc2Vec based on word embeddings[14], LSTM based on recurrent neural networks[15], SentenceBERT based on the Bert model[16], and SPECTER algorithms combining citation relationships[17]. These methods automatically learn the word frequency, order, and semantics of the papers, representing it as a low-dimensional vector. Here, the classic Doc2Vec algorithm is picked. Firstly, the titles and abstracts of scientific papers need to be preprocessed. The corpus cleaning steps include merging titles and abstracts, segmentizing Chinese words, removing symbols (punctuations, numbers, characters), and eliminating noise sentences (high-frequency words, email addresses, website addresses, copyright statements). Then, the popular PV-DM framework is used to train the Doc2vec model. The core idea of the PV-DM method is that the context of a word can be used to predict that word. At first, the corpus is built with paper ID and the preprocessed title and abstract of each paper in the dataset collected. In Doc2Vec, each document is represented as a vector. When training a given document, several local training sets are generated using a sliding window method. In each training iteration of each local training set, a word is selected as the predicted, other words in the local training set and the document vector are as input, and then the document vector is adjusted according to the predicted results. After traversing all local training sets of this given document, the training for the next document begins. As training iterations progress, all document vectors gradually stabilize. Given the unsupervised learning nature of Doc2vec, this study evaluates model quality by self-similarity validation, i.e., to check whether the most similar document predicted by the trained model is itself.

Finally, we obtain all the vectors for the corpus. The high-dimensional semantic vectors of the literature tensor the domain's high-dimensional semantic space, where the spatial distribution of paper points represents the knowledge structure.

(2) **Nonlinear dimensionality reduction**. Next, for the procedure of visualizing the knowledge map, it is necessary to project the abstract high-dimensional semantic space into a low-dimensional space. Common methods for reducing the dimensionality of high-dimensional data include t-SNE[18], PCA, and IPC. Recently, UMAP has gained attention, outstanding in its ability to maintain the global and local topology of high-dimensional vectors[19]. UMAP is a manifold learning algorithm based on Riemannian geometry theory, with the assumption that "close points in high-dimensional space are also close in low-dimensional space". Therefore, in this study, we apply the UMAP algorithm to project the semantic proximity among vectors measured by cosine distance into a two-dimensional plane. Unlike the Euclidean distance measurement which only considers the absolute magnitude of the distance between vectors, the cosine distance measurement focuses on the directional differences between vectors, allowing for measuring the differences in each dimension for document vectors. The smaller the cosine distance, the lower the text similarity. The major hyperparameters of the UMAP algorithm are set based on the empirical values recommended by the Google team[20]. This way, we get the 2-dimensional coordinates of all papers, and the knowledge landscape is then generated.

(3) **Depiction of topic structure**. The spatial distribution of paper points on the plane is quantified by Gaussian kernel density estimation (KDE) and depicted by contour lines. Regions on the map correspond to specific topics, and the proximity of papers on the map reflects the similarity or closeness of their knowledge at the semantic level. Additionally, the closed contour lines on the map depict the heterogeneity of the paper distribution, indicating that the density of papers within the loop is significantly different from the surrounding areas, thus reflecting the topic structure. The kernel density of areas also indicates whether the area is a research hotspot or a promising "blue ocean" for in-depth exploration.

Further, the specific research content of each region on the map is needed to be detected and presented. We primarily use keyword clouds to show the topic contents, which provides a direct and concise visualization. Specifically, we first extract each closed contour loop. Then, we extract keywords of papers within each region loop. Following this, we calculate the importance of the keywords for each region. According to the assumption that keywords that frequently appear in a specific region but are rare elsewhere are more representative, we use TF-IDF[21] to measure the importance. Additionally, we can leverage the powerful content summarization and generation capabilities of the LLM model for in-depth analysis of the topic content. We can consider the abstracts beneath each closed contour loop as a single comprehensive document, and we can use prompts to facilitate our understanding of it. In this paper, we mainly used the ZhiPuQingYan ChatGLM-4 [22] provided by the Tsinghua team for this attempt.

(4) **Evolutionary Analysis**. Observing the temporal semantic map allows for an intuitive understanding of the evolution of the field. After obtaining the 2D low-dimensional representations of all papers in the dataset D, D is sliced by time t $(D_1, D_2 \ldots D_n)$.

Then, the kernel density estimation (KDE) algorithm is performed on each sliced (cumulative) dataset, and then sliced (or cumulative) semantic maps are constructed. For paper vectors that are trained simultaneously, some region on these maps corresponds same semantics. This means that by comparing topography indicated by contour lines between maps in different periods, we can learn about the expansion of the field and the emergence of new topics in evolution. Observations on the cumulative map provide insights into the formation process of the research landscape. Sliced semantic maps over the years, achieved by summarizing the hotspot regions from the maps of each year, can effectively depict the shifts in research hotspots across different developing stages in a certain domain.

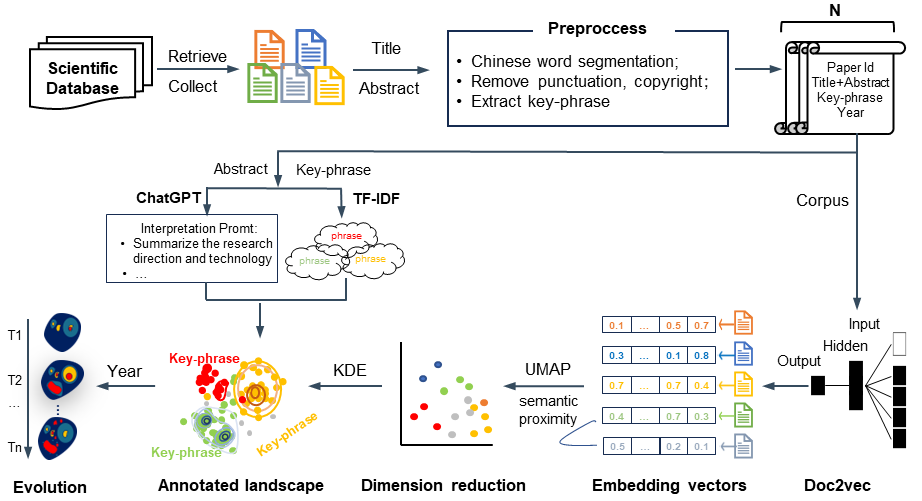

Fig.1. The framework of the visualization of the scientific knowledge landscape based on document semantics representation techniques

2.2 Data

In this study, we primarily intend to visualize the knowledge landscape of the field of editing and publishing in China. Therefore, we use the Chinese Wanfang Database as the data source and search seven representative Chinese academic journals, including "*China Publishing Journal*", "*Chinese Journal of Scientific and Technical Periodicals*", "*Publishing Research*", "*View on Publishing*", "*Editorial Friend*", "*Editors Monthly*" and "*Acta Editologica*". Finally, we collect a total of 76,491 journal articles, spanning from the year 1980 to 2022. To be noted, although the dataset does not encompass all literature, the search methods and data sources used ensure that the dataset includes high-quality scientific literature in the Chinese editing and publishing field. This dataset offers sufficient coverage and representation, enabling subsequent domain analysis.

Figure 2 shows the trend in the number of publications over the years. Firstly, it is observed that the variety of journals gradually increased. There is "*China Publishing Journal*" in the initial phase, and subsequently "*Editorial Friend*," "*Publishing*

Research" and "*Chinese Journal of Scientific and Technical Periodicals*" appear in succession. Second, the yearly number of papers in this field increase from 1980, reaching a peak in 2015 before gradually decreasing thereafter. To some extent, the observed trend in the yearly number of publications reflects the change in research hotness. On the other hand, this trend may be related to the repositioning of certain journals. In Figure 2, most journals have maintained relatively stable publication volumes over the years, except for "*View on Publishing*", which experienced a significant increase in publications from 2010 to 2013.

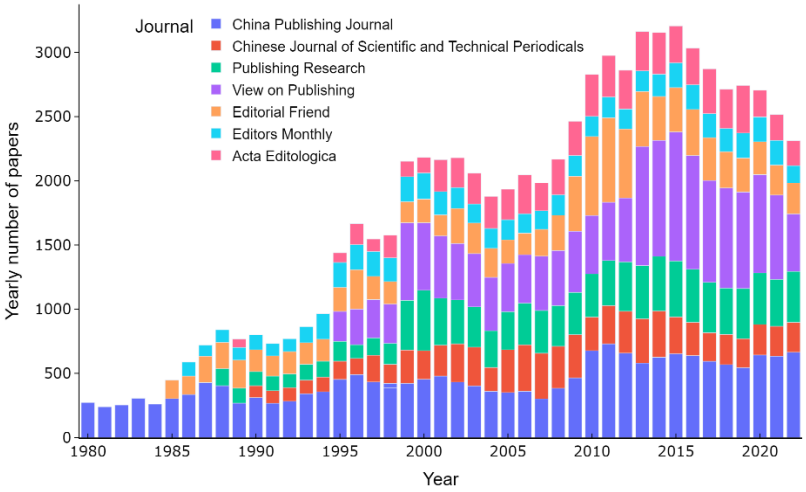

Fig. 2. The yearly Chinese journal papers in the dataset of the field of editing and publishing

3 The Structure of the Knowledge Landscape

We construct the static semantic map of the editing and publishing field, shown in Figure 3, which clearly visualizes the main research topics, the research hotness, and the interconnections among these topics (a detailed and interactive version is available at https://scientific-publishing-world.streamlit.app).

Figure 3a shows the three-dimensional landscape of the field of editing and publishing, where the horizontal plane represents the constructed semantic map and the vertical axis represents the distribution density of journal articles. We could observe that the three-dimensional landscape exhibits a mountainous profile, indicating that the publishing field has formed multiple research topics with varying degrees of research popularity and scope.

To further understand the mainstream research topics represented by the regions on the semantic map, we first annotate the semantic map with keyword word clouds. Here, in ease of presentation, we have translated the original Chinese keywords into English, and the original Chinese results are available at the abovementioned app link. In Figure 3b, thirteen areas with higher literature density are labeled with keyword word clouds.

These detected research topics are “scientific evaluation”(#1), “layout design”(#2), “organization of ancient books and dictionaries”(#3), “literary works”(#4), “press and publishing”(#5), “editor and editing activities”(#6), “textbook construction”(#7), “editing and reviewing of scientific journals”(#8), “social science journals”(#9), “editing and publishing companies(#10), “digital publishing”(#11), “online publishing”(#12), “copyright protection”(#13). Moreover, the contour lines of the literature density reflect that the exploration on “digital publishing” topic(#11) and “scientific journals” topic(#9) are relatively more intense, compared with “textbook publishing”(#7), “cover design”(#2), and “copyright protection”(#13)topics. In addition, by examining the spatial distribution of topics and their relative positions on the map, we can discern their topical connections or knowledge proximity. For instance, the “copyright protection” topic(#13) and “textbook construction” topic(#7) both present a “distant island”-like profile, indicating these two topics is quite distinct from the remaining research within the field. And, the fifth topic located at the center of the map, is characterized by a higher diversity of keywords, which may suggest that the research on this topic is more general and universal.

By further checking the abstracts of paper clusters in this region with ChatGLM, we find that these papers focus on the development, reform, and innovation of the editing and publishing industry, which supports that topic 5 tends to be comprehensive in scope. On the other hand, by utilizing ChatGLM to analyze the combined abstracts of the paper clusters, we compare the respective focuses of several topics containing the keywords “scientific journals” (#11, #10, #8, #6, #5, #1). We find that these topics all relate to scientific journals, with #11 topic emphasizing digital publishing, #10 topic focusing on integrated publishing, #8 topic discussing the editing and reviewing process, #6 topic about editors and editorial staff, #5 topic about innovative paths and #1 topic examining the evaluation and impact of scientific journals. These summarized results demonstrate that the LLM can collaboratively support the understanding of semantics-based visualization.

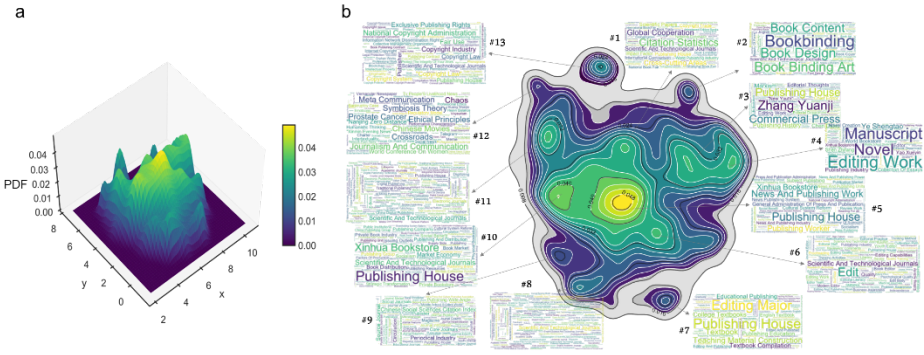

Fig.3. The knowledge landscape of the editing and publishing research field in China

4 The Evolution of Knowledge Landscape

4.1 Cumulative evolutionary semantic map

The changes in the cumulative evolutionary semantic map of the editing and publishing field illustrate the formation process of the current research landscape. Figure 4 presents the cumulative maps for every 5 years, illustrating that the evolution of the editing and publishing field is characterized by a continuous outward expansion from a central "research nucleus" and concurrent internal differentiation.

It is evident that over the past 40 years, the research scope of editing and publishing has continuously expanded from a core topic. The evolution of this field originates with the "ancient literature compilation and lexicography". As the discipline progresses, new research topics begin to emerge. Notably, around the 1985 year, the "textbook construction" topic emerges, and around the 1990 year, the "copyright management" topic surfaces along the edge of maps. Subsequently, topics such as "layout design" and "digital publishing" follow. However, we also observe that topics that emerge in successive periods are not related in content, without a strong semantic connection. In Figure 4, the emergence of new topics in each period seems to be arbitrary, and the topics that appear in adjacent periods are not geographically proximate to each other.

Along with the expansion of the "research facet" of the field, the "topography" depicted by contour lines within the map is also evolving. On one hand, it shows the dynamics of the hot spots of various research topics. On the other hand, it reveals the phenomena of decomposition and fusion courses among topics. For example, around the 2015 year, in the central region of the map, the "research on scientific evaluation of scientific journals" topic stands out.

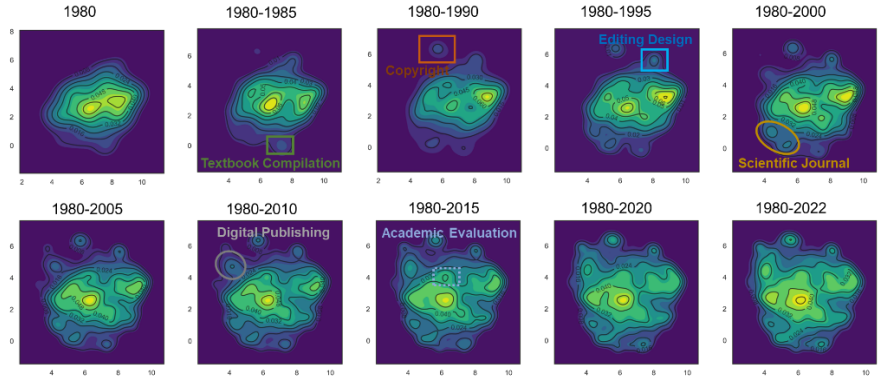

Fig. 4. A cumulative evolutionary semantic map of the editing and publishing field in China

4.2 Sliced Evolutionary Semantic Map

The research hotspots in the field have shifted over time, which could be vividly illustrated by sliced knowledge maps. By aggregating the sliced semantic maps with similar hotspot spatial distributions, we divide the development of the editing and publishing field from 1980 to 2022 year into four stages. As depicted in Figure 5, this field's focus

has shifted from the organization of ancient texts and dictionaries to knowledge services.

The first decade is an initial stage, from 1980-1990, focusing on the compilation and publication of ancient texts, dictionaries, and other fundamental resources. The second decade, from 1990-2000, focuses on the formulation of publishing norms. Enter the 21st century, the publishing industry experienced a boom, characterized by the proliferation of various commercial publishing houses, particularly academic journals such as science and technology journals. In the recent decade, starting from 2010, the field has merged with social media, stepping in a new era of digital publishing.

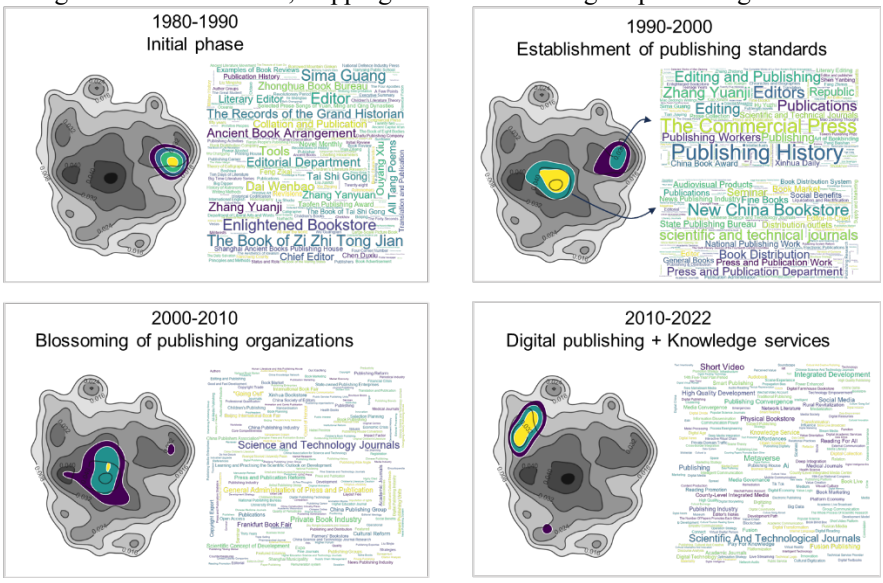

Fig. 5. Visualization of the hotspots transition in the field of editing and publishing

5 Conclusion

Advances in digital technology empower the editing and publishing industry. Grasp the current research landscape and evolutionary trajectory is of great significance for promoting industrial transformation and making technological policy. In this study, drawing on nearly eight thousand papers published in Chinese academic journals, utilizing the document embedding representation learning method Doc2vec and the manifold learning algorithm UMAP, the dynamic and static semantic space is vividly visualized.

Our results show that, firstly, the main research topics in the editing and publishing field are identified, including “copyright protection”, “scientific journals”, “layout design”, and so on, with diverse hotspots shown by the contour lines. Secondly, the cumulative formation process of this research field is revealed. We find the emergence of new topics along the edge of maps during the field expansion and a lack of significant semantic association between topics that emerge in successive stages. Subsequently,

the transitions research hotspots are presented and four stages of the development of the field are clearly detected.

There are still some limitations in this study. Firstly, we merely use academic papers to construct the knowledge landscape. It is worthwhile to combine patent data to comprehensively analyze the research trends in both the industry and academia. Second, this work merely utilizes ChatGLM simply to summarize aggregated abstracts. Further, large models can be attempted to extract knowledge structure. Third, this study solely focuses on the research landscape in China. In the future, comparisons with other countries could shed deep light on the scientific and technical development of Chinese editing and publishing.

References

1. Xu, L.L., Du, L.M., Tao, C.F., and Wu, Y.Q.: Research Hotspots and Visualization Analysis of Editing and Publishing in China.Science-Technology & Publication, 02, 125-132 (2021).
2. Wang, P., and Wang, N.Z.: An Analysis of the Hot Spots of Publishing Research and the Development Trend of the Publishing Industry in China from 1998 to 2020: A Metrological Study based on High-Impact Papers in CSSCI Source Journals. View on Publishing. 22,49-52 (2021).
3. AO, L.Y.: A Bibliometric Analysis of Research Hotspots in Digital Publishing – Based on Academic Journal Papers from CNKI. Media Science and Technology of China, 01, 136-139 (2023).
4. Fan, B., Jia, G.S., Zhang, Z., Fan, L.H., Wang, L., & Zhang, Y.L.: Key Technology of Knowledge Service in Publishing Industry Based on Term Subject Analysis. China Terminology, 25(03), 44-52 (2023).
5. Yang, B.C., and Wang, S.X.: 20 Years of Research on Fusion Publishing in China(2000-2020)--Based on a Knowledge Graph Analysis of Citespace and Gephi. Journal of Huaihua University, 41(01), 122-128 (2022).
6. Chen, C.: CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature, Journal of the American Society for Information Science and Technology. 57, 359-377 (2006).
7. van Eck, N.J., and Waltman, L.: Citation-based Clustering of Publications Using CitNetExplorer and VOSviewer. Scientometrics, 111, 1053-1070 (2017).
8. Velden, T., Yan, S., and Lagoze, C.: Mapping the Cognitive Structure of Astrophysics by Infomap Clustering of the Citation Network and Topic Affinity Analysis. Scientometrics, 111, 1033-1051 (2017).
9. Pourhatami, A., Kaviyani-Charati, M., Kargar, B., Baziyad, H., Kargar, M., and Olmeda-Gomez, C.: Mapping the Intellectual Structure of the Coronavirus Field (2000-2020): A Co-Word Analysis. Scientometrics, 126, 6625-6657 (2021).
10. Suominen, A., and Toivanen, H.: Map of Science with Topic Modeling: Comparison of Unsupervised Learning and Human-Assigned Subject Classification. Journal of the Association for Information Science and Technology, 67, 2464-2476 (2016).
11. Tshitoyan, V., Dagdelen, J., Weston, L., Dunn, A., Rong, Z., Kononova, O., Persson, K. A., Ceder, G., and Jain, A.: Unsupervised Word Embeddings Capture Latent Knowledge from Materials Science Literature. Nature, 571, 95-98 (2019).

12. Zhang, S., Liu, F., Luo, S., and Xia, H.: Smap: Visualization of Scientific Knowledge Landscape Based on Document Semantics, Journal of the China Society for Scientific and Technical Information, 42(01), 74-89 (2023).
13. Zeng, A., Liu, X., Du, Z., Wang, Z., Lai, H., Ding, M., Yang, Z., Xu, Y., Zheng, W., Xia, X., and Others: Glm-130b: An open bilingual pre-trained model, Arxiv Preprint Arxiv:2210.02414 (2022).
14. Le, Q., and Mikolov, T.: Distributed Representations of Sentences and Documents, in Proceedings of Machine Learning Research, Beijing, China, pp. 1188-1196 (2014)
15. Rao, G., Huang, W., Feng, Z., and Cong, Q.: LSTM with sentence representations for document-level sentiment classification, Neurocomputing, 308, 49-57 (2018).
16. Reimers, N., and Gurevych, I.: Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks, Arxiv abs/1908.10084 (2019).
17. Cohan, A., Feldman, S., Beltagy, I., Downey, D., and Weld, D. S.: SPECTER: Document-level Representation Learning using Citation-informed Transformers, Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 2270-2282 (2020).
18. van der Maaten, L., and Hinton, G.: Visualizing Data using t-SNE, Journal of Machine Learning Research, 9, 2579-2605 (2008).
19. McInnes, L., Healy, J., Saul, N., and Großberger, L.: UMAP: Uniform Manifold Approximation and Projection, Journal of Open Source Software, 3, 861 (2018).
20. Understanding UMAP, https://pair-code.github.io/understanding-umap, last accessed, 2022/5/26.
21. Jones, K. S.: A STATISTICAL INTERPRETATION OF TERM SPECIFICITY AND ITS APPLICATION IN RETRIEVAL, Journal of Documentation, 28, 11-21 (1972).
22. Zhipuqingyan, https://chatglm.cn, last accessed, 2024/1/20.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (http://creativecommons.org/licenses/by-nc/4.0/), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.