An Analysis Framework of Research Frontiers Based on The Large-scale Open Academic Graph

Han Huang
School of Information Management, Wuhan University, Wuhan, China
olivia.hanlanlan@qq.com

Hongyu Wang*
School of Information Management, Wuhan University, Wuhan, China
wanghongyu@whu.edu.cn

Xiaoguang Wang
Center for Studies of Information Resources, Wuhan University, Wuhan, China
wxguang@whu.edu.cn

ABSTRACT

[Purpose] As a high-quality and well-structured dataset, the large-scale open academic graph formed under the influence of the open science movement has created new research conditions for research frontier analysis. Constructing the analysis framework of research frontiers based on the large-scale open academic graph can effectively promote the realization of data-driven knowledge discovery and the analysis and decision-making of sci-tech intelligence. **[Approach]** The definition and analysis methods of research frontiers were summarized through related studies, and the data structure of the specific open academic graph was investigated. **[Findings]** The thoughts and steps of research frontier analysis based on the open academic graph were put forward, and an available analysis framework of research frontiers based on the large-scale open academic graph was constructed. **[Value]** The proposed framework can achieve deep, relevant and dynamic analysis of research frontiers in various disciplines based on the emerging large-scale open academic graph. It will provide a novel perspective for performing dynamic analysis across time and space, multidimensional analysis under multiple factors, and multiscale evolution analysis of research frontiers.

KEYWORDS

Analysis Framework; Open Academic Graph; Research Frontier Analysis; Topic Evolution Analysis

ASIS&T THESAURUS

Informetrics; Information and data processes

INTRODUCTION

The increasing diversity of scientific research cooperation and innovation, the increasing complexity of interdisciplinary integration, and the emergence of new disciplines necessitate a comprehensive, rapid, and accurate detection and analysis of the research frontiers in various disciplines so as to help frontline researchers and sci-tech management departments to grasp the situation of scientific and technological innovation in a timely manner and to optimize the allocation of scientific research resources. The research frontier is a clustering structure among the highly interactive papers in various disciplines (Small 1973), which reflects the subject areas that have gradually aroused people's interest and have been discussed and studied by a large number of researchers over time (Kontostathis et al. 2003). With the further development of the Open Science Movement (Evans and Reimer 2009), the large-scale open academic graph published by the major sci-tech information service providers has gradually become the new dataset for the analysis of research frontiers due to the integration of scientific papers as well as their relevant metadata, and the orderly organization of domain knowledge in scientific research. Moreover, it has created a better research condition for the future analysis of research frontiers.

In this paper, the definition and analysis methods of research frontiers were comprehensively summarized through the review of relevant works. Further, an analysis framework of research frontiers based on the large-scale open academic graph was constructed after the data structure investigation of a specific academic graph. Finally, a case study was conducted to demonstrate the logic of the framework.

LITERATURE REVIEW

Since the concept of research frontier was put forward in the 1960s, the multi-dimensional elaboration and definition of its concept connotation had been proposed by many scientometrics and bibliometrics scholars. Coherence, correlation, wholeness, perceivable and ostensive were widely used description properties for research frontier (Goldstein 1999). And Small (Small, Boyack and Klavans 2014) pointed out that novelty and growth are the two most commonly recognized characteristics of it.

Analysis of frontiers in scientific research refers to the process of actively discovering and analyzing emerging research topics, development status of these research topics, and associated structure in the area of scientific research by means of expert judgment and scientometrics methods. At present, two main types of measurement methods are used in the detection and analysis of research frontiers. One is the citation-based method, such as co-citation analysis and bibliographic coupling

* Corresponding Author

83rd Annual Meeting of the Association for Information Science & Technology | October 24-28, 2020.

analysis, and the other is the word-based analysis method, such as word frequency analysis, co-word analysis, and probabilistic topic model.

The premise of analyzing the research frontiers in various disciplines is to extract the subject topics from data sources, such as scientific papers. The subject topics are word representation of the research content in a discipline. The extraction of subject topics can be carried out by means of keyword extraction, through the three main steps of candidate keyword generation, feature engineering, and keyword extraction, driven by the characteristics of multiple dimensions. In the word frequency dimension, the classical term frequency−inverse document frequency algorithm and its derived algorithms can be used to quantify the keywords (Haddoud and Abdeddaïm 2014). In the dimension of co-occurrence characteristics among words, the semantic relationship between candidate words is judged according to the number of co-occurrences of the words (Mihalcea and Tarau 2004). Besides, the number of co-occurrence of citations (Gollapalli and Caragea 2014) and word embedding vector characteristics, such as Word2Vec (Wang, Liu and McDonald 2014), have also been used for keyword extraction and gradually received more attention.

Judging the state and change tendency of subject topics is one of the difficulties in frontier analysis. Numerous methods and models have been proposed to solve this problem, which can be divided into two types according to different granularity: word analysis and topic analysis. The word analysis method includes the burst detection algorithm (Kleinberg 2003) and the co-word analysis method (He and Wang 2015). Moreover, the topic analysis method integrates feature indicators of research frontiers into the topic model and then constructs a classifier to detect new topics (Lee et al. 2015). With the expansion of the data analysis scale and the application of complex network mining ideas, a novel method for research frontier analysis was proposed based on the co-word analysis method. That is, an algorithm was designed to extract the core nodes in the community of co-words in a discipline so as to analyze the evolution process of subject topics and its characteristics (Wang, Cheng and Lu 2014). Wallace (Wallace, Gingras and Duhon 2009), Prabhakaran (Prabhakaran, Lathabai and Changat 2015), and other researchers have shown that the research frontiers in various disciplines can be found more dynamically and comprehensively through the evolutionary analysis of complex networks, such as co-word network, co-citation network, citation network, and author co-network.

The open academic graph, as a type of academic dataset, represents and organizes sci-tech papers and their relevant information and knowledge regarding authorship, affiliations, journals, conferences, and so on, using semantic technologies, such as ontology, resource description framework, and linked open data, to fulfill the representation and organization. Further, it allows open access to the Internet. The typical open academic graph includes academic graphs such as the Springer Group’ s SciGraph, Tsinghua University ’ s Aminer, and Microsoft Academic Graph (MAG). In the process of building academic graphs, the team of Aminer and Microsoft Research Institute have explored the unified modeling method and semantic information matching algorithm of heterogeneous objects in the scientific knowledge network (Tang et al. 2008, 2010), besides the organization, discovery, and merging algorithms of scientific entities (Sinha et al. 2015). At the application level of academic graphs, Robertie (Robertie et al. 2017) and Vaccario (Vaccario et al. 2017) have confirmed the feasibility of frontier analysis based on the open academic graph according to the construction of researcher ’ s portrait and analysis of paper influence.

FRAMEWORK CONSTRUCTION

With the rapid development of emerging technologies, such as natural language processing, big data, and machine learning, realizing the analysis of research frontiers based on open academic graph is feasible nowadays. Therefore, the data structure of SciGraph was investigated in this paper. SciGraph is a graph stored in the form of triples, the model used to organize the data of main categories is shown in Figure 1.

[FIGURE 1]

Figure 1. Graph model of SciGraph

By the end of 2017, SciGraph had released metadata of more than 11 million papers based on the aforementioned model, including nearly 1 billion triples of papers and their relevant information regarding contributions, affiliations, grants, journals, and so on, in 12 categories. Through such a dataset, it is convenient to discover the evolution process of subject topics in the dimensions of communication, attention, and fund, and explore the development tendencies of research frontiers based on multilevel factors, such as grant support, journal publication, and paper citation, and thus realize the accurate analysis of research frontiers.

Based on it, according to the roadmap of “ open data acquisition → academic graph fusion → core topic extraction → multidimensional evolution analysis → frontier analysis → system development & empirical research”, an analysis framework of research frontiers based on the large-scale open academic graph was proposed, which is shown in Figure 2.

[FIGURE 2]

Figure 2. Structure of research framework

In this framework, it was necessary to design unified mapping rules based on the data structure of various academic graphs first to complete standardized information acquisition and graphical storage of the multisource heterogeneous large-scale open academic graph. Then, the entity links, knowledge merge, and alignment between large-scale were complete by the semantics-based knowledge fusion algorithm designed with reference to the Aminer's algorithm through the help of the tools, such as semantic ontology libraries and academic domain ontology. Third, the core topics in the dataset were extracted according to multidimensional features, such as word frequency, word embedding vector, and topic model, and the similar topics were merged through the designed merge algorithm. Moreover, the topics were used as the main line to complete and aggregate the data in the academic graphs. Then, the evolution mechanism and influencing factors of subject topics were analyzed using the correlation analysis algorithm from multiple dimensions, such as communication, funding, attention, and so on. Further, the expression of topics generation, division, fusion, and extinction in the academic graphs were revealed to achieve multidimensional evolution analysis and visual representation of topics in the large-scale open academic graph.

In the next step, a comparative analysis in the evolution process of individuals and groups in journal attention, fund preference, researcher interests, and other terms was completed. And the analysis model based on the multivariate indices of the frontier development tendency was constructed. Finally, a visualized analysis system of research frontiers was developed based on the model, and the comprehensive and accurate analysis of research frontier is realized through continuous system optimization.

CASE ANALYSIS

A case study which took 1,632,329 papers in the field of Medical and Health Sciences released by SciGraph (2019 FEB) in the past 20 years as dataset was conducted to demonstrate the logic of the framework. The SciGraphId, date of publication, title, abstract and funding information of papers were extracted. The distribution of those data is shown in Table 1. And the original data and results of this case can be accessed at https://github.com/wanghongyu94/ASIST2020.

Table 1. Distribution of case data

[TABLE 1]

Next, the RAKE algorithm (Rose et al. 2010) was used to extract several core topics from the title and abstract of each paper with the aid of the Medical Subject Headings (Lipscomb 2000). Then the distribution and evolution of research topics in this field over the past 20 years was visually analyzed by NEViewer (Wang, Cheng and Lu 2014) which is a visualization analysis tool for disciplinary topic networks developed by our team. The result is shown in Figure 3.

[FIGURE 3]

Figure 3. The subject community network evolution map

There are four main dynamic evolution paths of research topics (cell research, diagnosis and treatment methods, public health and function control), and the core keywords of each path are shown in Table 2:

Table 2. Main dynamic evolution paths and core keywords

[TABLE 2]

Preliminary observations show that in the past 20 years, the core topics of scientific papers in the path of cell research focus on the research of different types of cancer, and related studies on the treatment methods and genetic defects has gradually gained attention in recent years. In the public health path, it has gradually evolved from microscopic studies such as blood pressure control and epidemiology to the research of human-centered such as medical service process, health guideline and disease prevention & control. Next, this study will focus on the multidimensional analysis of topics to understand the evolution mechanism of research topics deeply.

Finally, the word frequency statistics were made for each core topic by year in this case, and the word frequency time series of 168,578 different core topics were formed. Then, the Mann-Kendall Test was used to identify the change trend of these time series (Marrone 2020), in order to test whether there was an obvious upward or downward trend in each word frequency time series. Then, the changing trends of the word frequency of core topics were correlated with the number of funds related to this topic. The results of the top 90 topics are shown in Table 3.

Table 3. Correlation analysis result of top 90 topics

[TABLE 3]

According to the number of related funds, the total number and trend distribution of all kinds of core topics were divided into sections statistically, as shown in Table 4.

Table 4. Piecewise statistics of correlation analysis result

[TABLE 4]

It can be found that in most of the core topics with fund support, the proportion showing an upward trend is the highest, and with the decline of the number of funds, the proportion showing an upward trend gradually decreases. This shows that there is a correlation between the funding and the increase of related research topics in a way. In the next step, this study expects to design the analysis model of research frontiers and complete the construction and experiment of the analysis framework based on the changing trend of the funding amount and related topics.

CONCLUSION

Based on the literature review of the research frontiers and the open academic graph, and combined with the data structure analysis of the open academic graph, an analysis framework of research frontiers based on the open academic graph was put forward in this study. Then, the operational logic of this framework is demonstrated by a case study in the field of Medical and Health Sciences.

Overall, the open academic graph has created new research conditions for dynamic analysis across time and space, multiscale evolution analysis, multidimensional analysis under multiple factors, and automated intelligent analysis of research frontiers. With the increasing fusion of knowledge between large-scale academic graphs and other open data in the scientific research area, it is possible to further complete a more comprehensive and accurate analysis for research frontiers and effectively realize the data-driven knowledge discovery in the future.

ACKNOWLEDGMENTS

This work is supported by the National Natural Science Fund (No. 71874129) and the Major Project of National Social Science Fund (No. 17ZDA292) of China. And this paper is supported by the China Scholarships Council (CSC).

REFERENCE

Evans, J. A., & Reimer, J. (2009). Open access and global participation in science. *Science*, 323(5917):1025-1025.

Goldstein J. (1999). Emergence as a Construct: History and Issues. *Emergence*, 1(1):49-72.

Gollapalli, S. D., Caragea, C.(2014). Extracting keyphrases from research papers using citation networks. *Twenty-Eighth AAAI Conference on Artificial Intelligence*, pp. 1629-1635.

Haddoud, M., Abdeddaïm S. (2014). Accurate keyphrase extraction by discriminating overlapping phrases. *Journal of Information Science*, 40(4):488-500.

He, Q. , & Wang, G. (2015). Hotspots evolution and frontier analysis of lean construction research - integrated scientometric analysis using the web of science and scopus databases. *Frontiers of Engineering Management,* 2(2):141-147.

Kontostathis, A., Galitsky, L. M., Pottenger, W. M., Roy, S., & Phelps, D. J. (2003). A survey of emerging trend detection in textual data mining. *Survey of Text Mining Clustering Classification & Retrieval*, 185-224.

Kleinberg, J. (2003). Bursty and Hierarchical Structure in Streams. *Data Mining & Knowledge Discovery*, 7(4):373-397.

Lee, Y. S. , Lo, R. , Chen, C. Y. , Lin, P. C. , & Wang, J. C. (2015). News topics categorization using latent Dirichlet allocation and sparse representation classifier. *2015 IEEE International Conference on Consumer Electronics - Taiwan*, pp. 136-137.

Lipscomb, C. E. (2000). Medical subject headings (MeSH). *Bulletin of the Medical Library Association*, 88(3), 265.

Marrone, M. (2020). Application of entity linking to identify research fronts and trends. *Scientometrics*, 1-23.

Mihalcea, R., Tarau, P.(2004). TextRank: Bringing Order into Texts. *Conference on Empirical Methods in Natural Language Processing*, pp. 404-411.

Prabhakaran, T. , Lathabai, H. H. , & Changat, M. (2015). Detection of paradigm shifts and emerging fields using scientific network: a case study of information technology for engineering. *Technological Forecasting and Social Change,* 91:124-145.

Robertie, B. D. L. , Ermakova, L. , Pitarch, Y. , Takasu, A. , & Teste, O. (2017). A unified approach for learning expertise and authority in digital libraries. *International Conference on Database Systems for Advanced Applications*. Springer, Cham.

Rose, S., Engel, D., Cramer, N., & Cowley, W. (2010). Automatic keyword extraction from individual documents. *Text mining: applications and theory*, 1, 1-20.

Sinha, A. , Shen, Z. , Song, Y. , Ma, H. , & Wang, K. (2015). An Overview of Microsoft Academic Service (MAS) and Applications. *the 24th International Conference*, pp. 243-246.

Small, H. (1973). Co-citation in the scientific literature: a new measure of the relationship between two documents. *J. Am. Soc. Inf. Sci.*, 24(4): 265-269.

Small, H., Boyack, K. W., Klavans, R. (2014). Identifying emerging topics in science and technology. *Research Policy*, 43(8):1450-1467

Tang, J. , Yao, L. , Zhang, D. , & Zhang, J. (2010). A combination approach to web user profiling. *ACM Transactions on Knowledge Discovery from Data, 5*(1):1-44.

Tang, J. , Zhang, J. , Yao, L. , Li, J. , & Su, Z. (2008). ArnetMiner: Extraction and Mining of Academic Social Networks. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 990-998.

Vaccario, G. , Medo, Matú, Wider, N. , & Mariani, M. S. (2017). Quantifying and suppressing ranking bias in a large citation network. *Journal of Informetrics,* 11(3):766-782.

Wallace, M. L. , Gingras, Y. , & Duhon, R. (2014). A new approach for detecting scientific specialties from raw cocitation networks. *Journal of the Association for Information Science & Technology,*60(2):240-246.

Wang, R., Liu, W., McDonald, C.(2014). Corpus-independent generic keyphrase extraction using word embedding vectors. *Software Engineering Research Conference, pp.* 39.

Wang, X. , Cheng, Q. , & Lu, W. (2014). Analyzing evolution of research topics with NEViewer: a new method based on dynamic co-word networks. *Scientometrics,* 101(2):1253-1271.