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

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 analysis methods of research frontiers were reviewed, and the existing problems in the research frontier analysis were summarized through related studies. Then, the data structure of the specific open academic graph was further investigated. **[Findings]** The thoughts and steps of research frontier analysis based on the open academic graph were put forward, and the main algorithms and tools that needed to be used were suggested, from which an 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; Subject Topic Evolution Analysis

ASIS&T THESAURUS

Informetrics; Knowledge and information; 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.

This paper briefly analyze the existing problems in the current analysis of research frontiers through related studies. Further, the thoughts and steps of the analysis of research frontiers based on the open academic graph were summarized after the data structure investigation of a specific academic graph. Finally, an analysis framework of research frontiers based on the large-scale open academic graph was constructed systematically.

LITERATURE REVIEW

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 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.

MAJOR LIMITATIONS OF FRONTIER ANALYSIS

The present study was limited by the influence of data source, data scale, and analysis principle. It failed to achieve deep, relevant, and dynamic analysis of research frontiers in a multidimensional, large-scale, and fine-grained manner, specifically in the following aspects:

- Insufficiency in depth and breadth

Most of the present research collects a certain amount of papers in a specific discipline for analysis, which easily ignores the cross-influence of research topics at different levels and different fields. As a result, some deficiencies in the depth and breadth of frontier analysis are reflected.

- Overlook of dynamic changes

Nowadays, many researchers focus on the scientometrics analysis based on the data of papers collected in a certain period but overlook the dynamic changes of the citation and the distribution of subject topics with time.

- Lack of data sources

The research frontiers in various disciplines are embodied in the data of sci-tech plans, research projects and fund reports, patent literature, scientific papers, and so on. However, most of the present research discovers and mines the research frontiers through the scientometrics analysis of specific categories of data. The source of data is relatively stuffless, which may have a negative impact on the integrity, accuracy, and objectivity of the analysis results.

- Limitation of analysis dimensions

The current research mostly fails to synthetically judge the multidimensional factors for research frontiers, such as sci-tech projects, research funds, academic journals, field scholars, and research institutes, and ignore the synergy and correlation between these factors, which makes the revealing of the research frontiers not so deep and specific.

- Deep dependence on experts

At present, the different dimension features, such as word frequency, word co-occurrence, citation co-occurrence, word similarity and word embedding vector, used in the extraction of subject topics usually depend on experts for selection and weight decision-making, which is difficult to adapt to the dynamic changes of research topics and the comprehensive judgment of features.

FRAMEWORK CONSTRUCTION

With the rapid development of emerging technologies, such as text mining, natural language processing, big data, and machine learning, realizing the analysis of research frontiers based on the open academic graph is feasible nowadays. Moreover, the opportunities for solving the problems that exist in the analysis of research frontiers have been created.

Both Aminer and MAG organize and store the related attributes of scientific papers according to the relational model, and their data structures are relatively simple. On the other hand, SciGraph is a graph stored in the form of triples, which integrates the data of papers and their relevant information regarding authorship, affiliations, funds, journals and conferences, etc. The model used to organize the data of main categories in SciGraph is shown in Figure 1 according to the investigation.

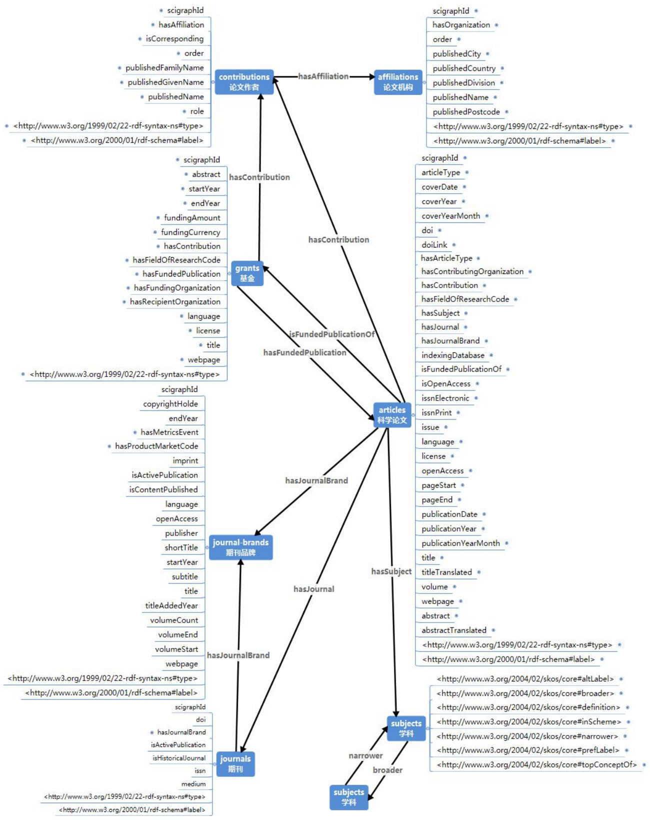


Figure 1. Graph model of SciGraph

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.

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

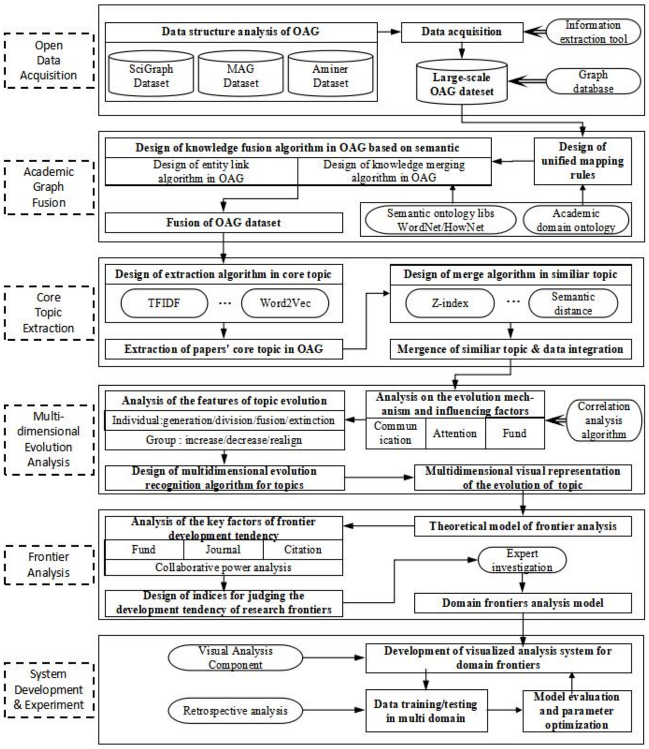

Open Data Acquisition
Data structure analysis of OAG
SciGraph Dataset
MAG Dataset
Aminer Dataset
Data acquisition
Information extraction tool
Large-scale OAG dateset
Graph database
Academic Graph Fusion
Design of knowledge fusion algorithm in OAG based on semantic
Design of entity link algorithm in OAG
Design of knowledge merging algorithm in OAG
Design of unified mapping rules
Fusion of OAG dataset
Semantic ontology libs WordNet/HowNet
Academic domain ontology
Core Topic Extraction
Design of extraction algorithm in core topic
TFIDF
···
Word2Vec
Extraction of papers' core topic in OAG
Design of merge algorithm in similiar topic
Z-index
···
Semantic distance
Mergence of similiar topic & data integration
Multi-dimensional Evolution Analysis
Analysis of the features of topic evolution
Individual:generation/division/fusion/extinction
Group : increase/decrease/realign
Analysis on the evolution mech-anism and influencing factors
Commun ication
Attention
Fund
Correlation analysis algorithm
Design of multidimensional evolution recognition algorithm for topics
Multidimensional visual representation of the evolution of topic
Frontier Analysis
Analysis of the key factors of frontier development tendency
Fund
Journal
Citation
Collaborative power analysis
Theoretical model of frontier analysis
Expert investigation
Design of indices for judging the development tendency of research frontiers
Domain frontiers analysis model
System Development & Experiment
Visual Analysis Component
Development of visualized analysis system for domain frontiers
Retrospective analysis
Data training/testing in multi domain
Model evaluation and parameter optimization

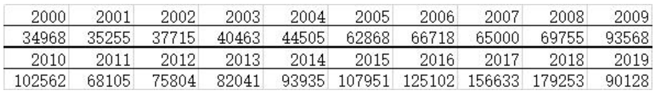
Figure 2. Structure of research framework

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, with the help of the tools, such as semantic ontology libraries and academic domain ontology, the semantics-based knowledge fusion algorithm, which is used to complete the entity links, knowledge merge, and alignment between large-scale academic graphs, was designed with reference to the Aminer's algorithm. 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 collaborative power analysis of the evolution process of individuals and groups in journal attention, fund preference, researcher interests, and other terms was completed. Then, the multivariate indices of the frontier development tendency were designed and the analyzing model of the research frontiers was constructed. Finally, based on the model, a visualized analysis system of research frontiers was developed. Through the training and testing of data in various disciplines, the performance of the system was evaluated and optimized, and the comprehensive and accurate analysis of research frontiers was finally realized.

CASE ANALYSIS

1,632,329 papers in the field of Medical and Health Sciences released by SciGraph (2020 JAN) in the past 20 years were taken as dataset, and a case study was conduct to demonstrate the logic of the framework in this paper.

In the data processing stage, by importing the dataset into the MongoDB, the data whose "article_genre" is "research_article", the "datePublished" is after 2000, and the subject label is "Medical and Health Science" are selected, and the SciGraphId, date of publication, title, abstract and funding information of those papers are extracted. The original data and results of this case can be accessed at https://github.com/wanghongyu94/ASIST2020. The distribution of those data is as follows:



2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
34968	35255	37715	40463	44505	62868	66718	65000	69755	93568
2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
102562	68105	75804	82041	93935	107951	125102	156633	179253	90128

Figure 3. Distribution of case data

Next, with the aid of the Medical Subject Headings (Lipscomb 2000), the RAKE algorithm (Rose et al. 2010) was used to extract several core topics from the title and abstract of each paper in this case. Based on the co-occurrence relationship between core topics, NEViewer (Wang, Cheng and Lu 2014), a visualization analysis tool for disciplinary topic networks developed by our team was applied to visually analyze the distribution and evolution of research topics in the field of Medical and Health Sciences over the past 20 years. The result is shown in figure 4.

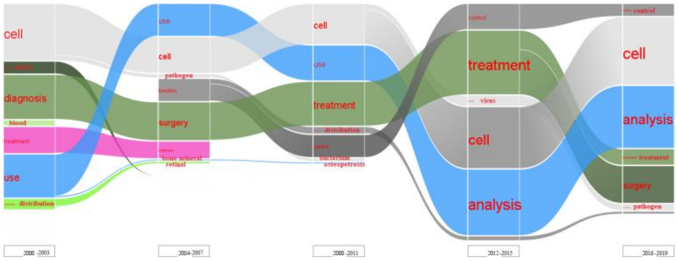
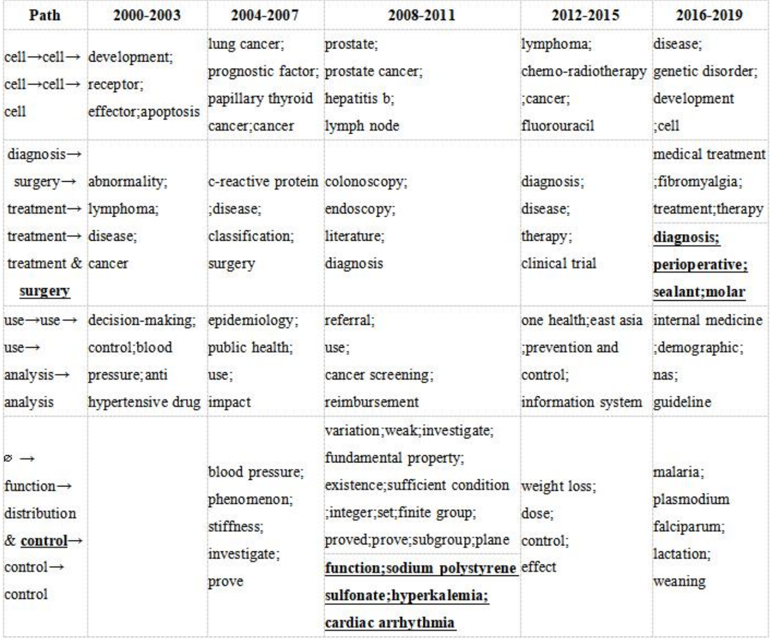


Figure 4. The subject community network evolution map

The result includes four main dynamic evolution paths of research topics (cell research, diagnosis and treatment methods, public health and function control), of which the core keywords contained in the communities of each period are shown in figure 5:

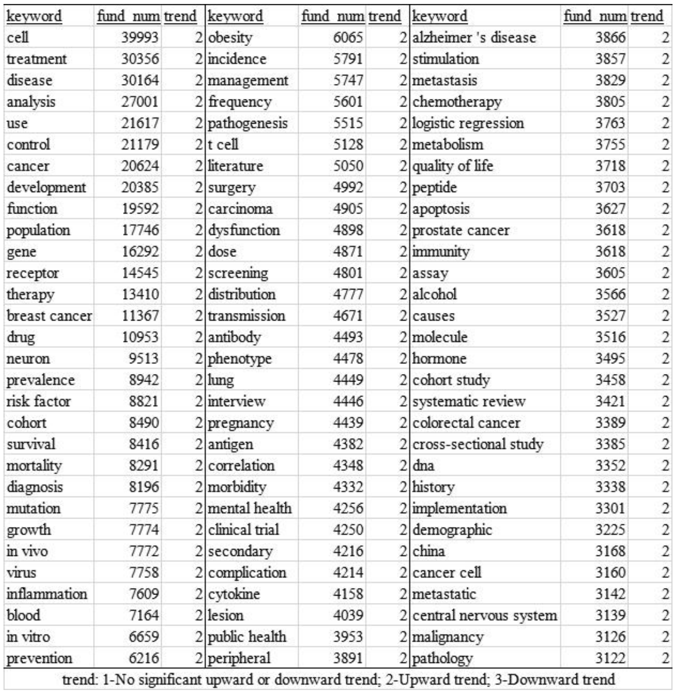


Path	2000-2003	2004-2007	2008-2011	2012-2015	2016-2019
cell→cell→ cell→cell→ cell	development; receptor; effector;apoptosis	lung cancer; prognostic factor; papillary thyroid cancer;cancer	prostate; prostate cancer; hepatitis b; lymph node	lymphoma; chemo-radiotherapy ;cancer; fluorouracil	disease; genetic disorder; development ;cell
diagnosis→ surgery→ treatment→ treatment→ treatment & **surgery**	abnormality; lymphoma; disease; cancer	c-reactive protein ;disease; classification; surgery	colonoscopy; endoscopy; literature; diagnosis	diagnosis; disease; therapy; clinical trial	medical treatment ;fibromyalgia; treatment;therapy **diagnosis; perioperative; sealant;molar**
use→use→ use→ analysis→ analysis	decision-making; control;blood pressure;anti hypertensive drug	epidemiology; public health; use; impact	referral; use; cancer screening; reimbursement	one health;east asia ;prevention and control; information system	internal medicine ;demographic; nas; guideline
∅ → function→ distribution & **control**→ control→ control		blood pressure; phenomenon; stiffness; investigate; prove	variation;weak;investigate; fundamental property; existence;sufficient condition ;integer;set;finite group; proved;prove;subgroup;plane **function;sodium polystyrene sulfonate;hyperkalemia; cardiac arrhythmia**	weight loss; dose; control; effect	malaria; plasmodium falciparum; lactation; weaning

Figure 5. Main dynamic evolution paths and core keywords

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 different scale of topic and conduct multidimensional analysis, so as 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. In this case, the changing trends of the word frequency of core topics were correlated with the number of grants related to those core topics. The results of the top 90 topics are shown in figure 6:

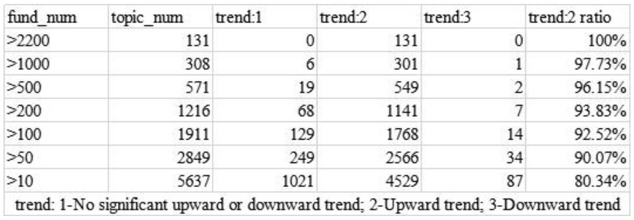


keyword	fund_num	trend	keyword	fund_num	trend	keyword	fund_num	trend
cell	39993	2	obesity	6065	2	alzheimer 's disease	3866	2
treatment	30356	2	incidence	5791	2	stimulation	3857	2
disease	30164	2	management	5747	2	metastasis	3829	2
analysis	27001	2	frequency	5601	2	chemotherapy	3805	2
use	21617	2	pathogenesis	5515	2	logistic regression	3763	2
control	21179	2	t cell	5128	2	metabolism	3755	2
cancer	20624	2	literature	5050	2	quality of life	3718	2
development	20385	2	surgery	4992	2	peptide	3703	2
function	19592	2	carcinoma	4905	2	apoptosis	3627	2
population	17746	2	dysfunction	4898	2	prostate cancer	3618	2
gene	16292	2	dose	4871	2	immunity	3618	2
receptor	14545	2	screening	4801	2	assay	3605	2
therapy	13410	2	distribution	4777	2	alcohol	3566	2
breast cancer	11367	2	transmission	4671	2	causes	3527	2
drug	10953	2	antibody	4493	2	molecule	3516	2
neuron	9513	2	phenotype	4478	2	hormone	3495	2
prevalence	8942	2	lung	4449	2	cohort study	3458	2
risk factor	8821	2	interview	4446	2	systematic review	3421	2
cohort	8490	2	pregnancy	4439	2	colorectal cancer	3389	2
survival	8416	2	antigen	4382	2	cross-sectional study	3385	2
mortality	8291	2	correlation	4348	2	dna	3352	2
diagnosis	8196	2	morbidity	4332	2	history	3338	2
mutation	7775	2	mental health	4256	2	implementation	3301	2
growth	7774	2	clinical trial	4250	2	demographic	3225	2
in vivo	7772	2	secondary	4216	2	china	3168	2
virus	7758	2	complication	4214	2	cancer cell	3160	2
inflammation	7609	2	cytokine	4158	2	metastatic	3142	2
blood	7164	2	lesion	4039	2	central nervous system	3139	2
in vitro	6659	2	public health	3953	2	malignancy	3126	2
prevention	6216	2	peripheral	3891	2	pathology	3122	2

trend: 1-No significant upward or downward trend; 2-Upward trend; 3-Downward trend

Figure 6. Correlation analysis result of top 90 topics

According to the number of related funds, the total number and trend distribution of all kinds of core topics were devided into sections statistically, as shown in figure 7.



fund_num	topic_num	trend:1	trend:2	trend:3	trend:2 ratio
>2200	131	0	131	0	100%
>1000	308	6	301	1	97.73%
>500	571	19	549	2	96.15%
>200	1216	68	1141	7	93.83%
>100	1911	129	1768	14	92.52%
>50	2849	249	2566	34	90.07%
>10	5637	1021	4529	87	80.34%

trend: 1-No significant upward or downward trend; 2-Upward trend; 3-Downward trend

Figure 7. Piecewise statistics of correlation analysis result

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 related 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, this study analyzed the problems existing in the current frontier research areas, such as shortcomings in the depth and breadth, overlook of dynamic changes, lack of data source, limitation of analysis dimension, and deep dependence on experts. Combined with the data structure analysis of the open academic graph, this study further put forward the research thoughts and steps of frontier analysis based on the open academic graph and made suggestions on the main algorithms and tools to be used. Thus, an overall framework of frontier analysis based on the open academic graph was constructed.

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 of China (No. 71874129) and Major Project of National Social Science Fund of China (No. 17ZDA292).

REFERENCE

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.

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.

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

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

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

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

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

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

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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

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

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