Postprint: Document-Topic Semantic Matching Analysis Method Based on Co-word and Word2Vec Weighted Vectors

Authors: Ding Jingda, Chen Yifan, Liu Chao, Cai Wei

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

Abstract

[Purpose/Significance] Co-word analysis, as an important method for topic identification, has certain limitations and deficiencies. Integrating Word2Vec weighted vectors with co-word analysis facilitates clarifying the topic affiliation of specific documents and enables better analysis of the development and evolution of topics. [Method/Process] Building upon topic clustering using co-word analysis, document vectors and cluster topic vectors are calculated separately via Word2Vec weighted vectors, and semantic matching between documents and topics is performed based on cosine similarity. [Results/Conclusion] Empirical analysis in the domestic and international knowledge sharing domain demonstrates that this method can effectively match relevant documents to their corresponding topics and enables dynamic analysis of topic characteristics and development evolution at the document level.

Full Text

An Article-Topic Semantic Matching Analysis Method Based on Co-word and Weighted Word2Vec Vectors

Ding Jingda, Chen Yifan, Liu Chao, Cai Wei
School of Cultural Heritage and Information Management, Shanghai University, Shanghai 200444

Abstract:
[Purpose/Significance] As an important method for topic identification, co-word analysis has certain limitations. Combining Word2Vec weighted vectors with co-word analysis helps clarify the topic attribution of specific documents and enables better analysis of topic evolution. [Method/Process] Based on topic clustering using co-word analysis, this study calculates document vectors and

cluster topic vectors through Word2Vec weighted vectors, and performs semantic matching between documents and topics based on cosine similarity. [Result/Conclusion] Empirical analysis in the field of knowledge sharing at home and abroad shows that this method can effectively match relevant documents to corresponding topics and dynamically analyze topic characteristics and evolution from the document level.

Keywords: Word2Vec; co-word analysis; semantic matching; knowledge sharing; topic evolution
Classification Number: G203
DOI: 10.13266/j.issn.0252-3116.2022.12.010

Scientific literature serves as an important knowledge carrier, containing rich semantic content and topic information. Mining and analyzing the content topics of massive scientific literature not only facilitates the transformation of library and information work from document services and information services to knowledge services, but also helps governments, research institutions, and relevant personnel understand domain topics, trace topic evolution, grasp development trends, and discover potential research topics.

However, co-word analysis, commonly used in literature topic identification research, has limitations such as treating word pair co-occurrence strength equally [**?**] and being unable to detect the topic distribution contained in specific documents [**?**]. Therefore, based on Word2Vec' s ability to represent text semantics, this study combines co-word analysis and Word2Vec to construct a document-topic semantic matching analysis method based on co-word and Word2Vec weighted vectors. This enables measurement and analysis of document-level features such as publication trends, timelines, and topic content evolution for clustering topics based on co-word networks.

2 Related Research Review

2.1 Co-word Analysis

Co-word analysis was proposed by M. Callon et al. in the 1980s [**?**]. It utilizes the co-occurrence of professional terms or noun phrases in literature collections, extracts these word co-occurrence relationships on a large scale, and uses clustering methods to simplify the complex co-word network relationships between terms into relationships among a relatively small number of clusters [**?**]. This approach thereby intuitively expresses topics in literature collections with unclear relationships.

Since its inception, scholars have discussed problems around the principles and application processes of co-word clustering analysis, such as ignoring inter-word relationships, ignoring the importance of words in documents, and results being independent of specific documents [**?**, **?**]. In response, researchers have proposed

corresponding improvements, such as co-word analysis methods based on multi-attribute weighting of documents [?] and improving co-word analysis clustering results using edge community detection algorithms [?].

2.2 Topic Identification Methods Combining Word2Vec

Word2Vec is an open-source word vector training tool developed by Google in 2013 that can transform text information from unstructured form into vectorized form [?]. The generated word vectors are semantically relevant and pay more attention to contextual logic [?], making related or similar words closer in distance.

Introducing Word2Vec and other semantic models into topic identification mainly involves two combination approaches. The first is model-level fusion of topic identification models with Word2Vec word vectors to improve topic identification effectiveness. For example, Yan Duanwu et al. found that combining Word2Vec word vectors with LDA document-topic distribution can more comprehensively and accurately describe the semantic information of microblog texts [?]. Wang Yingze et al. used the Word2Vec model to transform text collections into vocabulary relationship matrices as input data for LDA models for topic identification, analyzing the topic characteristics of disruptive technology policy texts in the EU, UK, and US [?]. C. E. Moody embedded LDA into the Word2Vec learning process, learning not only word embeddings but also topic representations and document representations simultaneously, improving the semantic conciseness of LDA-generated topics [?]. Wang Weijun et al. used Word2Vec to map keyword co-occurrence relationships into low-dimensional vector space, finding that this method can not only evaluate keyword importance in co-occurrence networks but also quantify the co-occurrence strength between disciplinary keywords [?].

The second approach uses word vectors for more detailed topic analysis through matching text-to-text and word-to-word correlations. Examples include Yan Shengfeng using Word2Vec to identify similar topic words under the DTM topic model, achieving synonym merging in topic words [?]. Zhou Yunze et al. selected the 10 topic words with the highest membership probability in LDA-identified topics and combined them with Word2Vec word vectors to represent topic vectors for similar topic matching [?]. C. Li et al. also proved that combining Word2Vec with LDA models using weighted vectors can effectively represent technical topic features as low-dimensional dense vectors and achieve semantic matching between documents and topics using cosine similarity [?], enabling more refined semantic modeling.

In summary, the organic combination of Word2Vec with topic words and text vocabulary can effectively represent the semantic features of topics or texts, enabling finer-grained semantic association and analysis. Current research mainly focuses on combining LDA with Word2Vec, but the LDA topic model is more suitable for long texts and has poor topic identification effects for short texts.

Additionally, the number of topics in the model needs to be manually determined based on perplexity curves. Although there are some improvements for these problems [**?**], relatively mature solutions have not yet been formed. Moreover, research combining co-word analysis with Word2Vec is relatively scarce, mainly using Word2Vec to learn or replace co-occurrence relationships [**?**, **?**], with less utilization of Word2Vec' s advantages in text matching. Considering that topics obtained through co-word analysis only manifest as clusters of different keywords and the results are independent of documents, to overcome the limitation that the topic distribution contained in any document cannot be detected [**?**], this study attempts to combine co-word analysis with weighted Word2Vec. This approach leverages Word2Vec's ability to represent text semantics to construct a topic-document semantic matching analysis method based on co-word analysis and Word2Vec weighted word vectors for document-level feature measurement and evolutionary analysis of topics.

3 Method Construction and Topic Measurement

3.1 Method Construction

To solve the problem that co-word analysis cannot perform measurement at the document level, this study applies the Word2Vec model to co-word analysis to achieve topic-document similarity matching under co-word networks, thereby assigning different documents to corresponding topics. First, use keyword information from bibliographic data to build a domain dictionary required for word segmentation, perform word segmentation and part-of-speech filtering after data cleaning, select high-frequency words to build a keyword co-occurrence network and conduct topic clustering. Second, use titles, abstracts, and keywords as text data to train Word2Vec word vectors, and construct topic vectors and document vectors based on word vectors. Finally, implement document-topic matching according to established rules and select topic measurement indicators for result measurement and analysis.

3.1.1 Constructing the Co-word Matrix Using Python' s jieba segmentation package, extract keywords from bibliographic data to establish a domain dictionary required for segmentation, perform segmentation and part-of-speech screening, and build a co-word matrix in three steps: (1) Synonym merging; (2) High-frequency keyword selection; (3) Matrix construction.

For synonym merging, calculate the character overlap between keywords using $(word_i \cap word_j)/(word_i \cup word_j)$, where $word_i$ represents the character set in keyword i. For example, the character set for the keyword "structural equation model" is {结, 构, 方, 程, 模, 型}. For words with high character overlap, supplement with manual screening to obtain synonyms.

For high-frequency keyword selection, this study uses Price' s formula $M = 0.749 \times \sqrt{N_{max}}$, where N_{max} is the occurrence frequency of the most frequent keyword. Cluster the co-word matrix to extract different research topics, where

the keywords under each topic constitute the topic words for that topic.

3.1.2 Document-Topic Matching The Word2Vec word vector model is essentially a three-layer neural network model with "input layer - hidden layer - output layer" [**?**], as shown in Figure 1 [Figure 1: see original paper]. $w(t)$ is the target word, and its context words are $w(t-r), ..., w(t-1), w(t+1), ..., w(t+r)$. The model has two learning methods: CBOW (Continuous Bag of Words) and Skip-gram. The Skip-gram model predicts the context of the target word based on the target word itself.

This study adopts the Skip-gram learning method, using the combination of title, abstract, and keywords (instead of a single document) as the training dataset for the Word2Vec model. For $topic_t$ obtained from co-word matrix clustering, use the trained Word2Vec model to generate word vectors for each topic word in $topic_t$, and obtain the vector representation of $topic_t$ through weighted summation of word vectors based on word frequency relationships:

$$W_{topic_i}(t = 1, 2, ..., T) = w_1 \times w2v_{k1} + w_2 \times w2v_{k2} + ... + w_{kt} + w2v_{kkt}$$

where $w2v_{ki}$ represents the Word2Vec word vector of topic word k_i in $topic_t$, T represents the number of topics, kt represents the number of topic words in $topic_t$, and w_i is the weight of topic word k_i, which is the ratio of the frequency of topic word k_i to the total frequency of all topic words in that topic.

For document vectorization, we still use the combination of title, abstract, and keywords (instead of a single document) as the data source. However, the abstract may contain some high-frequency irrelevant words. To better measure the importance of vocabulary, we use TF-IDF to weight the word vectors of each word in the data source to obtain the document vector $w2v_{tfidf\_di}$, which reduces the influence of high-frequency words with low discriminative power [**?**] and improves the feature representation effect of Word2Vec [**?**].

Finally, calculate the similarity between each document and various topics through cosine similarity:

$$Similar_{topic_t\_di} = cosine(W_{topic_t}, w2v_{tfidf\_di})$$

For T topics and D documents, a total of $T \times D$ calculations are required. The document-topic matching rules are: (1) If a document d_i has $Similar_{topic_t\_di} \geq \beta$ for any topic $topic_t$, then the document belongs to $topic_t$; (2) If a document d_i has $Similar_{topic_t\_di} < \beta$ for all topics $topic_t$, then the document only belongs to the topic with the maximum $Similar_{topic_t\_di}$. Through this method, one document can be assigned to different topics, and each document may correspond to more than one topic, which aligns with the actual situation since many documents may be related to multiple topics.

3.2 Topic Measurement and Evolution Analysis

3.2.1 Topic Feature Measurement Common indicators for topic feature measurement include topic strength [**?**], novelty [**?**], influence [**?**], interdisciplinarity [**?**], and attention [**?**]. Overall, relevant measurement indicators typically incorporate the number of documents corresponding to the topic and their publication dates. Therefore, this study uses three indicators—topic strength, attention, and novelty—to examine the characteristics of each topic.

(1) Topic Strength (Strength Index, SI). Topic strength is the most intuitive manifestation of whether a topic is popular. From a quantitative perspective, the more documents a topic accumulates, the greater the effort researchers have invested in it, the more profound its influence in the academic field, and the stronger the topic' s strength:

$$SI_{topic_t} = \Sigma(1) \text{ from } i = 1 \text{ to } D, \text{ where document } i \text{ belongs to } topic_t$$

(2) Attention (Attention Index, AI). Attention is a dynamic process that needs to be described from both temporal and quantitative dimensions. From the temporal dimension, due to limited researcher attention and factors such as the topic' s own development status and social changes, researchers' attention to a topic fluctuates over time. From the quantitative dimension, attention is equivalent to the annual topic strength, i.e., the number of documents under the topic each year. Using the document-topic matching method to obtain the number of documents produced each year under each topic can quantify researchers' attention to the topic:

$$AI = SI_{year\_topict}$$

where SI_{year_topict} represents the number of documents belonging to $topic_t$ each year.

(3) Novelty. As documents "age," their content becomes increasingly outdated, and their value as information sources continuously decreases. The influx of new documents, accompanied by potentially new theories, methods, and perspectives, also accelerates the depreciation of existing documents. Therefore, after obtaining the corresponding documents for a topic through the document-topic matching method, we can further measure the novelty of these documents as an important indicator for judging the topic' s development potential. The year of first publication is a common indicator for revealing the age of documents. A topic' s novelty can be represented by the median publication year of documents belonging to that topic. A larger median indicates that most documents in the topic were published more recently, representing higher potential for new achievements.

3.2.2 Topic Evolution Analysis Compared with traditional co-word analysis that studies topic evolution in stages, this research can directly analyze the development trajectory of topics from the temporal dimension. The specific method is: after obtaining the documents corresponding to each research topic using the above method, divide the documents under the topic by year, and use the keywords of documents each year as the corpus to calculate the TF-IDF values of keywords between different years. Sorting by TF-IDF values in descending order yields the core keywords for each topic each year. Through keyword burst analysis, we can gain a dynamic macro understanding of the topic' s research trajectory, supplemented by content analysis of documents with corresponding keywords under the topic to understand the topic' s development and evolution in detail.

4 Empirical Analysis

With the advent of globalized knowledge economy, the production, processing, innovation, and application of knowledge have increasingly become the driving force for economic growth and social development. Knowledge is regarded as a key strategic resource for both organizations and individuals [**?**], and knowledge sharing, as a critical process for sharing, utilizing, and creating knowledge, has attracted attention from enterprises and scholars. However, the growing volume of literature makes it difficult to grasp core knowledge. To comprehensively understand the knowledge system and development frontiers of knowledge sharing research, identify research entry points, and enhance enterprise competitiveness, this study takes the "knowledge sharing" domain as an example to empirically analyze the document-topic matching method and conduct topic feature measurement and evolution analysis.

4.1 Data Sources and Processing

Chinese data were obtained from CNKI, searching with "knowledge sharing" as the theme, limited to core journals, CSSCI, and CSCD sources, with a retrieval date of April 20, 2021, yielding 5,481 journal papers. After removing irrelevant data such as announcements and reports, 5,132 journal papers remained.

English data were obtained from the Web of Science Core Collection, searching with "knowledge sharing" as the theme, limited to English language, with a time span of 1996-2020 (limited by database access permissions, our institution' s IP can only retrieve data from 1996 onwards. Knowledge sharing research originated in 1990, but publications from 1990-1995 were minimal [**?**]), with a retrieval date of April 20, 2021, yielding 5,813 journal papers. After removing irrelevant literature, 5,625 journal papers remained. The annual distribution of literature is shown in Figure 2 [Figure 2: see original paper]. The number of domestic papers gradually declined after 2010, while the number of foreign papers surged suddenly in 2015 and continued to rise each year thereafter.

4.2 Co-word Clustering

In the process of constructing the keyword co-occurrence network, since "knowledge sharing" is the theme word of this study and "knowledge management" has overly broad meaning and excessively high frequency that is not conducive to clustering, these two terms were removed in subsequent research. The remaining keywords were cleaned, and after multiple experiments, keywords with overlap above 0.6 were selected for manual screening to achieve synonym merging.

Price's formula was then used to calculate the high-frequency keyword threshold. Words appearing more than 14 times (domestic) and 11 times (foreign) were selected as high-frequency keywords, with domestic papers containing 125 such words and foreign papers containing 277. The keyword co-occurrence matrix was clustered, and the strength distribution of clustering results is shown in Figure 3 [Figure 3: see original paper].

4.3 Document-Topic Semantic Matching

Using Python's jieba and nltk libraries for data processing, "jieba.load_{userdict}()" loaded the custom domain dictionary, and pytorch was used to implement the Word2Vec word vector training model. Then, the word frequency and TF-IDF values of each keyword were calculated, and combined with trained word vectors for weighted summation to obtain topic vectors and document vectors respectively. Cosine similarity was used to match topic vectors with document vectors. After multiple experiments, it was found that setting the matching threshold to 0.62 for domestic literature and 0.24 for foreign literature yielded good document-topic classification results.

Tables 1 through 4 list 10 documents with highest similarity and 10 with lowest similarity under the domestic "tacit knowledge and explicit knowledge" topic and the foreign "social media" topic. It was found that titles of documents with high similarity to the topic vector often contain core keywords, while documents with low similarity can mostly serve as extended research for the topic.

4.4 Topic Feature Analysis

4.4.1 Topic Strength and Attention The topic strength distribution for domestic and foreign knowledge sharing research is shown in Figure 3. Domestic research focuses on knowledge sharing influencing factors and knowledge sharing models and performance, while foreign research emphasizes innovation value-added from knowledge sharing and collaborative research between organizations and individuals.

Comparing the attention changes for knowledge sharing topics (Figure 4 [Figure 4: see original paper]), for domestic research, except for the influencing factors topic, attention to other topics began to decline around 2010. For example, before 2009, attention to knowledge sharing models and performance research was higher than that for influencing factors, but related research attention declined

year by year after 2009, while knowledge sharing influencing factors research received more attention and maintained high and stable output each year, becoming the current mainstream research.

For foreign research, since foreign knowledge sharing is still in a growth period with emerging results, the strength size and attention degree show similar changes. Since 2007, research on innovation value-added based on knowledge sharing has received increasing attention and developed into the current mainstream, while collaborative research also surpassed social media in 2017 to become the second most concerned topic for foreign researchers.

Overall, compared with domestic knowledge sharing research, foreign research pays more attention to technological iteration, inter-organizational cooperation, knowledge, product, and service innovation value-added from knowledge sharing, and knowledge sharing based on social media, with stronger research progression shown by a slow accumulation to rapid rise in attention. Domestic knowledge sharing research content focuses more on influencing factors and various applications, showing greater fluctuation in attention, with research on knowledge sharing technology and social media-based knowledge sharing being emphasized.

4.4.2 Topic Novelty The publication year distribution of documents corresponding to each research topic is shown in Figure 6 [Figure 6: see original paper]. The median line of foreign box plots is overall higher than that of domestic box plots, indicating that foreign knowledge sharing research has higher novelty overall.

Specifically, domestic knowledge sharing topic novelty is mainly distributed in 2010-2013, with influencing factors and knowledge transfer between organizations and teams having relatively new publication years. The lower novelty of knowledge sharing models and performance research and intellectual property research may be due to earlier starts, but with the introduction of new technologies and relevant laws and regulations, new research topics continue to emerge.

Foreign knowledge sharing topic novelty is mainly distributed in 2011-2014, with influencing factors and collaboration receiving continuous attention from researchers in recent years, while knowledge sharing technology has lower novelty, possibly because related technologies are relatively mature and research results are gradually being applied to other topics.

Through novelty measurement of domestic and foreign topics: (1) Foreign research is overall more novel; (2) Researchers' research and application of knowledge sharing influencing factors are continuously updated and improved, and knowledge transfer and cooperation coordination between individuals and organizations still maintain high attention; (3) There is no direct relationship between the starting time of topic research and topic novelty; (4) Domestic research on influencing factors and foreign research on innovation value-added and collaboration coordination all have high strength and novelty, making them more likely to produce cutting-edge research directions or development trends.

4.5 Topic Evolution Analysis

Figure 7 [Figure 7: see original paper] plots the attention of foreign researchers under the "collaborative coordination" topic and lists TOP5 keywords for some years. Foreign research on collaborative coordination mainly focuses on how individuals, organizations, and groups achieve coordination for efficient knowledge sharing.

Over time, research objects have gradually diffused from enterprises to cities, nations, and virtual communities, and research questions have become more diversified. In the initial stage of topic development, system platforms were used as media to study strategic cooperation between enterprises and problems faced in knowledge sharing practice. "Multi-agent systems" based on ontology methods became emergent topic words, and establishing complete agent systems could achieve coordinated operation among various links within organizations, with process optimization also helping knowledge sharing implementation [**?**].

In 2007, knowledge sharing theory and technology were introduced into higher education and attracted researcher attention. In traditional classroom teaching, information and communication technologies were used to enhance classroom collaboration and group interaction [**?**], with "improved classroom teaching" and "multidisciplinary design" becoming core vocabulary, while research on traditional organizations focused on enhancing knowledge sharing among "stakeholders" for efficient operation.

Since 2011, research on using knowledge sharing to promote sustainable development of cities [**?**] and public sectors began to emerge. On the other hand, with increasingly severe climate issues, bringing knowledge sharing into climate change adaptation research would promote decision-making execution and response to emergencies, such as learning from successful climate governance cases in different countries and sharing theoretical frameworks and data used [**?**]. Additionally, enabling doctors, patients, and multiple parties to collaboratively participate in treatment and care through knowledge sharing [**?**] became a novel research direction in 2016.

In 2020, affected by the COVID-19 pandemic, organizational work patterns and environments changed significantly, and how to improve knowledge sharing for students, office workers, and medical personnel in online modes requires further discussion and practice. Knowledge sharing related to pandemic information and prevention in social media and virtual communities [**?**] received widespread attention from researchers.

5 Conclusion

This study proposes a document-topic matching analysis method combining Word2Vec weighted vectors and co-word analysis, and conducts empirical analysis using the knowledge sharing domain at home and abroad to compensate for the limitations of co-word analysis in document-level measurement. First, nat-

ural language processing and text mining techniques were used to clean the bibliographic data of knowledge sharing literature. Second, co-word analysis was used to obtain relevant research topics in the knowledge sharing field. Then, weighted Word2Vec word vectors were used to match documents with corresponding research topics.

Empirical analysis results show that this method can obtain documents highly related to research topics. Compared with traditional co-word analysis, this method can not only detect topic evolution macroscopically but also evaluate topic development status from the document perspective using existing topic measurement indicators, combined with topic word burst analysis to deeply analyze the development trajectory and dynamic evolution of topics.

The limitation of this study is that this is an unsupervised method where thresholds need to be adjusted subjectively based on matching results. While higher thresholds can improve the accuracy of topic-document matching, they may also cause potential topics of some documents to be ignored. In the future, ideas from supervised topic models such as Label-LDA and MedLDA can be referenced, combining observable document external feature information such as publication location and authors to annotate documents for automated generation of optimal thresholds.

References

[1] Ba Zhichao, Li Gang, Zhu Shiwei. Research on keyword selection and semantic measurement in co-occurrence analysis [J/OL]. [2022-03-22]. https://arxiv.org/abs/1310.4546v1.

[2] Zhou Liqin, Xu Jian, Ba Zhichao, et al. Comparative study on hypertension topic detection and evolution trends based on SNA and DMR methods [J]. Library and Information Service, 2018, 62(13): 82-91.

[3] Callon M, Courtial J P, Turner W A, et al. From translations to problematic networks: an introduction to co-word analysis [J]. Social science information, 1983, 22(2): 191-235.

[4] Zhong Weijin, Li Jia, Yang Xingju. Research on co-word analysis (III) —Principles and characteristics of co-word clustering analysis [J]. Journal of Intelligence, 2008(7): 118-120.

[5] Li Gang, Ba Zhichao. Research on several issues in co-word analysis process [J]. Journal of Library Science in China, 2017, 43(4): 93-113.

[6] Li Feng. Clustering analysis based on core keywords —Also on the shortcomings of co-word clustering analysis [J]. Information Science, 2017, 35(8): 68-71, 78.

[7] Sun Haisheng. Research on improving co-word analysis clustering results using edge community detection algorithms [J]. Library and Information Service, 2016, 60(10): 123-129.

[8] Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality [EB/OL]. [2022-03-22]. https://arxiv.org/abs/1310.4546.

[9] Qiu Huilin, Shao Bo. Research on scientific research hotspot identification methods in multi-source data environment [J]. Library and Information Service, 2020, 64(5): 78-88.

[10] Yan Duanwu, Mei Xirui, Yang Xiongfei, et al. Research on microblog text topic clustering based on fusion of topic model and word vector [J]. Modern Information, 2021, 41(10): 67-76.

[11] Wang Yingze, Hua Bolin. Research on topic modeling analysis of disruptive technology policy text data in European and American countries [J/OL]. Information Studies: Theory & Application, 2022: 1-14 [2022-03-22]. http://kns.cnki.net/kcms/detail/11.1762.g3.20220225.1702.002.html.

[12] Moody C E. Mixing dirichlet topic models and word embeddings to make lda2vec [J/OL]. [2022-03-23]. https://arxiv.org/abs/1605.02019.

[13] Wang Weijun, Yao Chang, Qiao Ziyue, et al. A method for discovering interdisciplinary knowledge of NSFC based on word embedding —Taking "artificial intelligence" and "information management" as examples [J]. Journal of the China Society for Scientific and Technical Information, 2021, 40(8): 831-845.

[14] Yan Shengfeng. Temporal modeling and evolution analysis of public policy text integrating word vector semantic enhancement and DTM model —Taking the "big data field" as an example [J]. Information Science, 2021, 39(9): 146-154.

[15] Zhou Yunze, Min Chao. Emerging technology identification based on LDA model and shared semantic space —Taking autonomous vehicles as an example [J/OL]. Data Analysis and Knowledge Discovery, 2021: 1-16 [2022-03-25]. http://kns.cnki.net/kcms/detail/10.1478.g2.20211206.1917.007.html.

[16] Li C, Guo J, Lu Y, et al. LDA meets Word2Vec: a novel model for academic abstract clustering [C]//Companion of the Web Conference 2018. Republic and Canton of Geneva: CHE, 2018: 1699-1706.

[17] Jiang Tian, Liu Xiaoping, Liu Huizhou. Research on LDA noise topic filtering based on keyword relevance index (KRI) [J]. Library and Information Service, 2020, 64(3): 92-99.

[18] Wang Tingting, Han Man, Wang Yu. Optimization of LDA model and its topic number selection research —Taking scientific literature as an example [J]. Data Analysis and Knowledge Discovery, 2018, 2(1): 29-40.

[19] Huang L, Chen X, Zhang Y, et al. Identification of topic evolution: network analytics with piecewise linear representation and word embedding [J]. Scientometrics, 2022, 127(2): 1-31.

[20] Yu Qiuyu, Xu Yuequan. Empirical research on high-frequency word threshold determination methods in co-word analysis —Taking high-frequency word

selection in COVID-19 literature as an example [J]. Information Science, 2020, 38(9): 90-95.

[21] Bai Rujiang, Liu Bowen, Leng Fuhai. Research on identifying future emerging scientific research frontiers based on multi-dimensional indicators [J]. Journal of the China Society for Scientific and Technical Information, 2020, 39(7): 747-760.

[22] Tang X, Wan Y, Liu Y, et al. Chinese spam classification based on weighted distributed characteristic [C]//Proceedings of the 2017 Chinese Automation Congress. Jinan: 2017: 6618-6622.

[23] Bai Jingyi, Yan Duanwu, Chen Qiong. Research on emerging topic trend prediction based on topic model and curve fitting [J]. Information Studies: Theory & Application, 2020, 43(7): 130-136, 163.

[24] Wu Yiping, Yu Chunliang, Qu Jiabin, et al. Research on university paper research frontiers and evolution trends from the perspective of text topics [J]. Information Science, 2021, 39(5): 156-162, 183.

[25] Huang Lu, Zhu Yihe, Zhang Yi. Research on emerging technology topic prediction based on weighted network link prediction [J]. Journal of the China Society for Scientific and Technical Information, 2018, 37(7): 686-694.

[26] Fan Shaoping, An Xinying, Yan Guilai, et al. Research on frontier topic identification methods in medical field [J]. Journal of the China Society for Scientific and Technical Information, 2019, 38(4): 335-341.

[27] Liu Ziqiang, Xu Haiyun, Yue Lixin, et al. Topic diffusion evolution lag effect for research frontier prediction [J]. Journal of the China Society for Scientific and Technical Information, 2018, 37(10): 979-988.

[28] Xiong Huixiang, Li Yueyan. Research on researcher recommendation and cross-language paper recommendation model based on Word2vec [J]. Information Science, 2019, 37(12): 19-26.

[29] Castaneda D I, Cuellar S. Knowledge sharing and innovation: a systematic review [J]. Knowledge and process management, 2020, 27(3): 159-173.

[30] Zhang Chunyang, Liang Qihua. Analysis of current status and development trends of knowledge sharing science based on Web of Science [J]. Library Science Research, 2016(18): 20-29.

[31] Kock N, Davison R. Can lean media support knowledge sharing? investigating a hidden advantage of process improvement [J]. IEEE transactions on engineering management, 2003, 50(2): 151-163.

[32] Looi C K, Chen W. Community-based individual knowledge construction in the classroom: a process-oriented account [J]. Journal of computer assisted learning, 2010, 26(3): 202-213.

[33] Shen L Y, Ochoa J J, Shah M N, et al. The application of urban sustainability indicators —a comparison between various practices [J]. Habitat international, 2011, 35(1): 17-29.

[34] Johanna M, Natasha K, Arnoldo M K, et al. Climate adaptation research for the next generation [EB/OL]. Climate and development, 2013: 189-193 [2022-03-25]. https://www.tandfonline.com/doi/full/10.1080/17565529.2013.812953.

[35] Georgia T B N, Tracey B, Andrea M, et al. Patients' perceptions of participation in nursing care on medical wards [J]. Scandinavian journal of caring sciences, 2016, 30(2): 260-270.

[36] Edgheim F, AbuAlqumboz M, Mouzughi Y. Covid-19 transition, could Twitter support UK Universities? [J/OL]. Knowledge management research & practice, 2020: 1-6 [2022-03-25]. https://www.tandfonline.com/doi/full/10.1080/14778238.2020.1848364.

[37] Sakusic A, Markotic D, Dong Y, et al. Rapid, multimodal, critical care knowledge-sharing platform for COVID-19 pandemics [J]. Bosnian journal of basic medical sciences, 2020, 21(1): 93-97.

Author Contributions

Ding Jingda: Paper topic selection, research framework formulation, paper writing;
Chen Yifan: Data processing and analysis, research framework formulation, paper writing;
Liu Chao: Research framework formulation;
Cai Wei: Data collection and analysis.

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