Priority Ranking of Disciplinary Research Topics Based on Fusion of Publication and Citation Trends: A Case Study of Information Science Topics in China (Postprint)

Authors: Li Xiuxia, Process crystallization, Han Xia

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

Abstract

[Purpose/Significance] Topic ranking is not only a fundamental issue in information retrieval and information organization research, but also an important task in library subject services. Effective ranking of research topics within a disciplinary field can help researchers and research management departments effectively grasp the research trends of the discipline, accurately position research directions, and make rapid research decisions.

[Method/Process] Based on trend analysis, this paper proposes a priority ranking algorithm for disciplinary research topics. First, on the basis of topic extraction, each research topic is classified into four subcategories according to research level based on publication trends and citation trends: impoverished topics, hot topics, cold topics, and overheated topics. Then, priority ranking is performed for topic terms within each subcategory respectively.

[Results/Conclusion] Experiments in the field of information science demonstrate that the priority ranking algorithm proposed in this paper can comprehensively, finely, and deeply display the development levels of research topics in a disciplinary field, and this method can provide a new perspective for achieving dynamic intelligence analysis from the temporal dimension.

Full Text

Preamble

Vol. 63 No. 11 June 2019

The Prioritization of Subject Research Topics Based on the Integration of Publication Trends and Citation Trends: Taking the Subject

of Information Science in China as an Example

Li Xiuxia[1], Cheng Jiejing[2], Han Xia[1]

[1]School of Communication, Qufu Normal University, Rizhao 276826
[2]Institute of Social Development, Yangzhou University, Yangzhou 225008

Abstract

[**Purpose/Significance**] Topic ranking is not only a fundamental issue in information retrieval and information organization research, but also an important task in library subject services. Effective ranking of research topics in a discipline can help researchers and research management departments grasp the research landscape of the field, accurately locate research directions, and make scientific decisions quickly. [**Method/Process**] This paper proposes a priority ranking algorithm for disciplinary research topics based on trend analysis. First, following topic extraction, each research topic is classified into four subcategories according to publication trends and citation trends: impoverished topics, hot topics, cold topics, and overheated topics. Then, priority ranking is performed for topic terms within each subcategory. [**Result/Conclusion**] Experiments in the field of information science demonstrate that the proposed priority ranking algorithm can comprehensively, finely, and deeply display the development levels of research topics in a discipline. This method provides a new perspective for achieving dynamic intelligence analysis from the time dimension.

Classification Number: G250.2
Keywords: publication trend; citation trend; research topic; prioritization
DOI: 10.13266/j.issn.0252-3116.2019.11.010

With the deepening of disciplinary research and the expansion of interdisciplinary studies, academic literature has grown exponentially, and research topics continue to evolve and update. Faced with massive literature resources covering diverse topics, how to quickly and accurately grasp the development levels of disciplinary research topics and determine research directions has become a significant challenge for scientific researchers [1]. Currently, many scholars use individual documents as the basic unit to provide information services through literature ranking, offering effective and reliable decision-making references for researchers to select topics [2-3]. Although literature ranking can identify valuable documents and authoritative authors in a field, this type of information service is far from sufficient given the continuous deepening and expansion of disciplinary research. Therefore, this study proposes a method for prioritizing disciplinary topics based on trend analysis, conducting in-depth mining of literature content to provide researchers with deeper and finer-grained information services [4].

The main objectives of this study are: (1) to provide definitions for relative citation volume, citation trend, and publication trend, establishing a theoretical foundation for dynamically analyzing the evolution of disciplinary research topics from the time dimension; (2) to propose a priority ranking method for disciplinary research topics based on trend analysis, providing theoretical support for discovering the fine-grained development trends of research topics; and (3) to apply the proposed ranking algorithm to prioritize research topics in Chinese information science, offering effective and reliable decision-making references for research management departments in formulating research plans and for researchers in selecting topics.

2 Related Research

2.1 Bibliometric-Based Topic Identification

This category of research includes term frequency analysis [5], co-word analysis [6], and co-word clustering analysis [7]. These methods are essentially based on high-frequency keywords to identify research topics of document clusters and discover research hotspots in the field. Keywords provide a high-level summary and condensation of a document's research subject, content, and methodology, reflecting the logical relationships or innovative breakthroughs of the research. High-frequency keywords represent hot topics and frontier directions in a discipline—the higher the frequency, the greater the research attention, typically constituting research hotspots [8]. Therefore, research hotspots and development trends in a field can be identified by statistically analyzing keyword frequencies. Bibliometric-based topic identification has been widely applied in identifying research hotspots and analyzing research structures due to its strong versatility and simple analytical tools. However, its limitations include strong subjectivity in keyword selection, lack of semantic relationships between keywords, and omission of low-frequency keywords that may represent emerging research topics, resulting in less than ideal effects in revealing the knowledge structure of a field [9].

2.2 Machine Learning-Based Topic Mining

Topic mining originated from the Vector Space Model (VSM) proposed by G. Salton et al. [10] in 1975. VSM represents text as vectors in geometric space, facilitating the calculation of similarity between texts and the relationship between keywords and documents. In 1990, S.C. Deerwester et al. [11] introduced Latent Semantic Analysis (LSA), successfully incorporating "semantics" into text topic mining for the first time. In 1999, T. Hofmann [12] proposed the Probabilistic Latent Semantic Analysis (PLSA) model using the expectation-maximization algorithm, incorporating machine learning into text topic extraction. In 2003, D.M. Blei et al. [13] introduced prior probabilities into latent semantic analysis and proposed the Latent Dirichlet Allocation (LDA) model based on PLSA. The LDA model assumes that words are generated from a mixture of topics, with each topic being a multinomial distribution over a fixed vocabulary shared by all documents in the collection, and each document having a specific topic proportion sampled from a Dirichlet distribution. In practice, topic extraction can be achieved as long as the document collection is determined and the number of latent topics is specified. Currently, the LDA model has become a widely used topic mining model, giving rise to a series of derivative methods [14-17]. Compared with bibliometric-based topic identification, machine learning-based topic mining can not only discover more comprehensive topics with more specific and explicit content descriptions [18-19], but also establish tighter semantic connections among keywords within topics, achieving better results in mining correct topics from documents with vague semantic relationships and rough logical structures [20].

2.3 Machine Learning-Based Topic Ranking

Some scholars have borrowed the PageRank algorithm from web ranking to rank scientific topics. For example, Jiang Zhuoren et al. [27] used the PageRank algorithm to measure and rank the importance of Chinese and English scientific topics. However, since PageRank is based on link analysis and cannot effectively handle topic-based queries, the calculation results often deviate from the actual query topics.

Effective ranking of disciplinary research topics can help researchers and research management departments grasp research trends, accurately locate research directions, and make quick decisions. This is significant and widely applicable. However, current algorithmic research on ranking disciplinary research topics mainly achieves topic ranking based on term frequency or associations between topic terms following text mining, with no studies considering user demand factors. Therefore, based on previous research, this study combines bibliometric and topic mining methods to achieve reasonable ranking of disciplinary research topics from two perspectives—readers and researchers—through two dimensions: publication trends and citation trends.

3 Research Steps and Methods

3.1 Topic Extraction and Topic Number Determination

Topic extraction involves extracting research topics from academic literature in a discipline. Generally, academic literature titles provide core information such as research content, methods, and objectives, while keywords offer high-level summaries of core content. As mentioned earlier, the LDA model has excellent capability for mining latent topics in text and can identify hidden topic information in large-scale document collections or corpora [22]. It has been applied in topic extraction, hotspot mining, text classification, user recommendation, and other fields. Therefore, this study adopts the LDA model to extract research topics from literature titles and keywords.

In disciplinary literature topic extraction, determining the appropriate number of topics is crucial. Too few topics cannot cover the full research landscape of the discipline, while too many lead to redundant analysis. Relying on author or expert recommendations to determine topic numbers introduces subjectivity. Therefore, this study uses the average similarity between all topics to measure the stability of topic structure. The average cosine value ranges between 1 and 0; the smaller the average similarity between topics, the better the corresponding topic structure [28].

3.2 Topic Word Citation Level Determination

For convenience of description, we make the following assumptions for a given discipline: suppose there are M topics, where any topic m contains k topic words, and there are N papers corresponding to a certain topic word in a given year. The publication volume for each time period corresponding to a topic word is represented by vector Ljn, where Ljn = (lj1, lj2, …, ljn).

First, we calculate the relative citation volume RCj for a topic word in a given year. Then, we calculate the total relative citation volume TC for all topic words in the discipline for that year. Next, we compute the sample standard deviation d based on the values of RCj and TC. Finally, we determine the citation level qjm for the j-th (j = 1, 2, …, k) topic word in a given year. The formulas are as follows:

Formula (1):
RCj represents the ratio of all citations for the j-th (j = 1, 2, …, k) topic word's corresponding documents to the total number of documents for that topic word in the same year. In the formula, Ci represents the citation count of the i-th (i = 1, 2, …, N) document. To avoid division by zero when publication or citation volume is zero, we add 1 to both numerator and denominator.

Formula (2):
TC represents the ratio of citations for all topic words' corresponding documents in a given year to the total number of documents for all topic words under topic m in that year. In the formula, Pij represents the i-th document of the j-th topic word, and Cij represents the citation count of the i-th document of the j-th topic word.

Formula (3):
d represents the sample standard deviation of k topic words in a given year.

Formula (4):
qjn represents the citation level of the j-th topic word in a given year, where n represents the time period (n = 1, 2, …). Each topic word has a citation level for each year.

3.3 Topic Word Citation Trend and Publication Trend

Based on the citation levels of topic words, we construct citation level vectors for n time periods for each topic word, i.e., Qjn = (qj1, qj2, …, qjn). All topic words correspond to the same time vector, i.e., Y = (y1, y2, …, yn).

We perform Spearman correlation analysis between the citation level vector Qjn and the time vector Y for each topic word to obtain the Spearman correlation coefficient QRj for each topic word relative to time. The Spearman correlation coefficient indicates the correlation direction between the citation level vector Qjn and citation time Y. If Y increases and Qjn tends to increase, QRj is positive; if Y increases and Qjn tends to decrease, QRj is negative; if QRj is zero, it indicates that Qjn shows no tendency as Y increases. The magnitude of QRj reflects the trend of increasing or decreasing reader demand for each topic word, which is recorded as the citation trend of each topic word.

We count the publication volume of each topic word in different time periods, construct a "publication volume-time" matrix with "time" as rows and "topic words" as columns, and represent the publication volume of each time period corresponding to a topic word with vector Ljn. We perform Spearman correlation analysis between Ljn and the publication time Y to obtain the Spearman correlation coefficient LRj. LRj reflects the correlation direction between Ljn and publication time Y. A positive LRj indicates that Ljn tends to increase as Y increases; a negative LRj indicates that Ljn tends to decrease as Y increases; an LRj of zero indicates that Ljn shows no change tendency. The magnitude of correlation coefficient LRj reflects the increasing or decreasing research trend of researchers on each topic word, which is recorded as the publication trend of each topic word.

3.4 Topic Word Priority Ranking

3.4.1 Topic Word Priority Classification Based on the definitions of relative citation volume, publication trend, and citation trend, we calculate the publication trend LRj and citation trend QRj for topic words corresponding to disciplinary research topics. According to different values of LRj and QRj, research topics are subdivided into four sub-topic categories, each representing different research levels. The classification criteria are as follows:

1. When the publication trend of a topic word decreases while its citation trend increases, demand exceeds supply, indicating that related research is in an impoverished state and urgently needs increased research attention. Therefore, such topic words are defined as impoverished topic words, and research on these topics requires immediate guidance and support, representing the highest research level.

2. When both publication and citation trends of a topic word increase, demand is relatively and rapidly increasing while supply is also growing rapidly, indicating hot topic words in the discipline. Research on such topics can meet demand, so the research level should be lower than that of impoverished topics.

3. When both publication and citation trends of a topic word decrease, both demand and supply for the research topic are low, indicating cold point topic words. Topics with relatively low demand do not require excessive

support, so the research level is lower than that of hot topics.

4. When the publication trend of a topic word increases while its citation trend decreases, supply exceeds demand, indicating that research on the topic word is growing relatively fast and showing signs of overheating. Therefore, research on such topics should be appropriately controlled, representing the lowest research level.

The specific representation is: - LRj $<$ 0, QRj $>$ 0 (Category , impoverished topic) - LRj $>$ 0, QRj $>$ 0 (Category , hot topic) - LRj $<$ 0, QRj $<$ 0 (Category , cold topic) - LRj $>$ 0, QRj $<$ 0 (Category , overheated topic)

Formula (5)

3.4.2 Sub-topic Ranking Priority ranking of topic words within each subcategory is achieved through operations on publication trend LRj and citation trend QRj. The ranking is based on a custom operation relationship rj = LRJ QRJ, where " " is a custom operator [28] that must be defined according to the different distribution characteristics of the data when applied.

4 Experiments and Effect Evaluation

4.1 Data Source

The data for this study were selected from journals indexed in the Chinese Social Sciences Citation Index (CSSCI) source journals. CSSCI journals represent the highest academic level in various fields of social sciences in China, and literature published in these journals basically covers research topics across disciplines. Among them, there are 20 journals in the category of "Library, Information, and Documentation Science," including 10 information science journals (including dual-category journals in library and information science): *Journal of the China Society for Scientific and Technical Information*, *Library and Information Service*, *Journal of Intelligence*, *Library and Information Knowledge*, *Information and Documentation Services*, *Data Analysis and Knowledge Discovery*, *Information Studies: Theory & Application*, *Information Science*, *Library and Information*, and *Modern Information*. To simultaneously obtain titles and keywords from the above journals for subsequent topic extraction, the author conducted a comprehensive search of these 10 information science journals in the CNKI database. The search time range was from June 2013 to May 2018, retrieving 13,559 documents. After removing irrelevant data such as conference announcements, table of contents, and calls for papers, 12,377 valid documents were obtained. The titles, keywords, and other information from these documents were downloaded as experimental data.

4.2 Determination and Extraction of Research Topic Numbers in Chinese Information Science

Before conducting LDA topic extraction, data preprocessing is required. First, the author used the Chinese word segmentation system NLPIR (also known as ICTCLAS) from the Institute of Computing Technology, Chinese Academy of Sciences, to segment the sample data. Then, part-of-speech filtering and stop-word filtering were applied to remove words irrelevant to modeling, obtaining the text corpus needed for the experiment.

By selecting different topic numbers and calculating the average similarity between topics, it was found that when the number of topics is 10, the average similarity between topics is minimized and the topic structure is most stable, as shown in Figure 1 [Figure 1: see original paper].

Using the open-source package JGibbLDA [27], LDA topic modeling was performed on the data with the number of topics set to 10, α and β set to 0.1 and 0.02 respectively, to extract topics. Based on the topic words in the topic lists generated by the LDA model for each research topic, and according to the author's understanding of disciplinary research topics, 10 research topic labels were manually determined, as shown in Table 4 .

4.3 Priority Classification of Sub-topics Under 10 Research Topics

To maintain consistency with the data above, the "Information Science, Information Work" category under the "Information Technology" classification in the CNKI database literature classification directory was selected. Through the "More" option, "Index" was selected. First, simple merging of topic words under each category was performed, such as merging "comparative analysis" with "contrastive analysis," "network public opinion," "network 舆论" with "network public opinion dissemination," "open government data" with "government open data," "satisfaction" with "user satisfaction," "visualization" with "visual analysis," and "evaluation index system" with "index system." Then, the processed topic words were input one by one, and the "academic attention" (i.e., publication volume) and "academic dissemination" (i.e., citation volume) corresponding to each topic word each year were retrieved and recorded. Considering the time lag of citations relative to publications, the publication time was set from June 2013 to May 2017, and the citation time from June 2014 to May 2018. Based on the definitions of relative citation volume, publication trend, and citation trend given in Sections 3.2 and 3.3, the publication trend LRj and citation trend QRj for each category of topic words were calculated. According to the different values of LRj and following the sub-topic classification method given in Section 3.4, topic words under each research topic were divided into four subcategories: impoverished topics, hot topics, cold topics, and overheated topics (some research topics were divided into three subcategories). The classification results are shown in Table 1 .

4.4 Topic Word Ranking Under Subcategories Within 10 Research Topics

Using the custom priority ranking algorithm for research topics given above, topic words under different research topics and subcategories were ranked for priority. Analysis of the publication trend LRj and citation trend QRj characteristics of topic words in each subcategory revealed that: some topic words showed negative correlation between QRj and LRj, while others showed positive correlation; additionally, citation trends QRj mostly clustered in the ranges (-1, -0.7) and (0.7, 1), with large gaps between maximum and minimum values. Meanwhile, to ensure that the ranking of rj values for the four subcategories is consistent with their priority levels and that all topic words have positive rj values, after repeated experiments, the following priority ranking algorithm was designed:

Formula (6):
rj = 13QRj - LRj + 14 (when QRj and LRj show opposite trends)
rj = 2QRj - LRj + 14 (when QRj and LRj show the same trend)

Using the above custom algorithm, the rj value for each topic word under each subcategory was calculated, with results shown in Table 1.

4.5 Effect Evaluation

Since there is currently no widely recognized evaluation method for disciplinary topic ranking, and no ranking studies of Chinese information science research topics have been found, this paper employs two methods to evaluate the ranking effectiveness: rationality analysis of ranking results and comparative experimental analysis.

4.5.1 Rationality Analysis of Ranking Results Taking "Topic5th: Open Data" as an example, we analyze the rationality of the ranking results. This category's first subcategory (Category) contains four topic words, representing two sub-directions: user experience of information services and knowledge management platform construction. Currently, the relevant research theories on information services and knowledge management are nearly complete, lacking innovative development approaches and concepts, thus showing a 逐年递减 trend in publication volume. However, in an era that advocates virtual reality environments and "people-oriented" concepts, people are increasingly focusing on user experience, and the demand for library services is gradually shifting from information services to knowledge services. It is evident that setting this subcategory as the highest research level is reasonable, and relevant institutions and departments should take measures to increase support for research in these two sub-directions to meet the rapidly growing demand for this subcategory of research topics.

The second subcategory (Category) includes topic words mainly studying recommendation systems, cloud computing, mobile services, government open data, and other issues. With the rapid development and widespread adoption of Internet technology, mobile technology, and IoT technology, the library and information science community attaches great importance to data openness and the construction of application platforms, emphasizing the position of mobile terminals in information services. This aligns with the 逐年递增 trends in both publication and citation volumes corresponding to the above topic words. However, if a discipline focuses too much on a particular research direction, it is not conducive to balanced disciplinary development. Therefore, the research level of this subcategory should be set lower than that of Category to remind researchers to remain calm and guard against research overheating.

The third subcategory (Category) contains only one topic word: personal information. Both its publication and citation volumes are 逐年递减. Although people pay more attention to network information security and user privacy protection in the network information age, library and information science does not involve as many privacy security issues as other disciplines (such as computer science). Therefore, it is natural for this subcategory to become a research cold spot in library and information science, with a relatively low research level.

The fourth subcategory (Category) represents research directions mainly focusing on digital resources. Resource management and construction have always been hot topics in library and information science and are the most proficient research directions in this field. In the digital era, research on resource digitization (such as digitization of information resources, digital resource organization and management, and digital resource utilization) has become a hot research direction in library and information science, belonging to overheated research topics. Relevant institutional departments (such as project approval agencies and library and information science academic journals) should take corresponding measures to appropriately control the research volume of this part of topics, hence setting the research level of this subcategory as the lowest.

4.5.2 Comparative Experimental Analysis For convenient comparison, this paper also conducted topic word ranking based on co-word clustering analysis using the same dataset, i.e., ranking topic words through processes such as co-word analysis, co-word clustering, and social network analysis. The specific process was: importing the bibliographic information (including keywords) of the above 10 journals into Bicomb, extracting keywords from each document, and obtaining 27,057 keywords after normalization through merging and deletion. Selecting 244 high-frequency keywords with occurrence frequency ≥ 20 to generate a co-word matrix, a similarity matrix of 244 keywords was obtained through correlation analysis. Finally, the co-word matrix was imported into VOSviewer for social network analysis. For comparison purposes, the number of topics was set to 10, dividing the 244 keywords into 10 categories. Keywords within each category were sorted according to their weight values in VOSviewer, as shown in Table 2 .

For ease of expression, the ranking method based on trend analysis proposed in this paper is called Method A, and the ranking method based on co-word clustering analysis is called Method B. Due to different theoretical foundations, the ranking results of the two methods show significant differences in: topic words within each topic, number of topic words, topic labels, and intra-class hierarchy.

Comparison reveals that Method A has the following obvious advantages over Method B:

(1) Theoretical Foundation Advantage. Method B is based on statistical methods to obtain high-frequency words in the discipline, ignoring a large number of low-frequency words and emerging topic words appearing in the long-tail position. The method itself is subjective and incomplete, leading to unobjective and incomplete topics. Method A uses the LDA model based on probabilistic inference, which has rigorous mathematical theoretical foundations, thus extracting more comprehensive and reliable research topics.

(2) Clustering Hierarchy Advantage. Method B only clusters topics at a single level of research content, while Method A not only distinguishes research topics in terms of content but also performs finer-grained research priority classification for each research topic. That is, based on the division of research topic content, each research topic is further subdivided into four levels: research impoverished points, research hotspots, research cold spots, and research overheated points (some topics are divided into three levels). This represents clustering at two levels—research content and research level—making the clustering effect more refined.

(3) Comparison of Ranking Results. Method B presents ranking of research topic popularity, only showing readers the research hotspots in the discipline and providing only one type of information: research development trends. For example, in Table 2, the top-ranked keywords in each topic represent the research hotspots of that topic. Method A can not only display disciplinary research hotspots and overheated topics (e.g., Category in Table 1 belongs to research hotspots, Category belongs to overheated topics) but also provide research levels for disciplinary topics. For instance, in Table 1, topic words with higher rj values in each topic have higher research levels. Thus, Method A provides more comprehensive information.

(4) Clustering Nature Advantage. Method B belongs to hard clustering, where a keyword appears in only one topic category. Method A belongs to soft clustering, where a keyword can appear in different categories. For example, "resource aggregation" belongs to both the research content of "topic3th: network resources" and the research scope of "topic5th: open data." This method is consistent with the diversity of keyword content orientation, making the clustering results more reasonable.

5 Research Contributions

The main contributions of this study are:

(1) Definitions of relative citation volume, publication trend, and citation trend are provided. Relative citation volume considers the impact of publication volume on citation volume, breaking through the limitation of viewing topic development solely from citation volume, and can objectively present the development trends of disciplinary research topics. Publication trend reflects the current research status of topics, while citation trend reflects the degree of attention topics receive. Combining the two can analyze the development trends of research topics from two different perspectives: researchers and readers.

(2) A research topic priority ranking method is proposed. This ranking method breaks through the 思路 of "uniform ranking of all research topics." First, different research topics are divided into four research levels based on publication trend and citation trend. Then, topic words under the four levels are ranked according to the given ranking algorithm. This not only can display the full picture of disciplinary research topics in detail but also can specifically present the degree to which disciplinary research topics are studied and concerned.

(3) Research priority classification of Chinese information science topics is conducted. By calculating topic similarity, information science research topics are divided into 10 categories. Using the proposed research topic priority ranking method, research topics are classified into four categories: impoverished topics, hot topics, cold topics, and overheated topics. The research results can provide effective and reliable decision-making references for scientific research institutions in this discipline to formulate research plans and for researchers to determine research directions.

This study is an exploratory research on priority ranking of disciplinary research topics. The ranking algorithm itself still has certain limitations, and the verification of research results has not been well resolved, requiring further discussion.

**(1) In any discipline, due to possible shifts in researchers' preferences and changes in the number of readers, the publication volume and reader demand for research topics will change to some extent, and publication trends and citation trends will change accordingly. Moreover, with the continuous development of the discipline and increasingly frequent disciplinary exchanges, new research topics will continue to emerge. All these factors will affect the ranking results of disciplinary research topics. Therefore, research on ranking disciplinary research topics should be a continuous process. The ranking method given in this paper can only show the current research landscape to relevant departments and researchers and provide references for recent research topic selection.

**(2) The topic word priority ranking algorithm in this study, i.e., Formula (6), is given based on the dataset of this study and needs to be customized according to specific data characteristics when applied.

**(3) Since there is currently no widely recognized verification method for disciplinary topic ranking, and no ranking studies of Chinese information science research topics have been found, this study only conducted co-word clustering analysis on the same dataset to compare the advantages of the proposed method in terms of theoretical foundation, clustering nature, clustering hierarchy, and ranking results. However, whether the ranking results are consistent with expert judgment has not been reasonably verified.

References

[1] Zhang Xiaolin. Disruptive trends in digital libraries[J]. Journal of Library Science in China, 2011, 37(195): 4-11.
[2] Liu Xin. Research on ranking methods for scientific literature based on reading value[D]. Dalian: Dalian University of Technology, 2010.
[3] Wang Yanpeng. Research progress on topic discovery and evolution of scientific literature based on topic models in China[J]. Library and Information Service, 2016, 60(3): 130-137.
[4] Guan Peng, Wang Yuefen. Scientific literature topic mining based on LDA topic model and life cycle theory[J]. Journal of the China Society for Scientific and Technical Information, 2015, 34(3): 286-299.
[5] Donohue JC. Understanding scientific literature: a bibliographic approach[M]. Massachusetts: The MIT Press, 1973.
[6] Tang Guoyuan, Zhang Wei. Research progress and analysis of disciplinary topic evolution based on co-word analysis[J]. Library and Information Service, 2015, 59(5): 128-136.
[7] Chen Shiji, Wang Xiaomei. Research on topic identification of document clusters based on C-value and TF-IDF[J]. Journal of the China Society for Scientific and Technical Information, 2009, 28(6): 821-826.
[8] Chen Yongyue, Tian Wenfang, Wu Jinhong. Research on visual analysis methods for tracking research hotspots and predicting trends in subject areas[J]. Information Studies: Theory & Application, 2017, 40(6): 117-121.
[9] Ba Zhichao, Li Gang, Zhu Shiwei. Research on keyword selection and semantic measurement methods in co-occurrence analysis[J]. Journal of the China Society for Scientific and Technical Information, 2016, 35(2): 197-207.
[10] Salton G, Wong A, Yang CS. A vector space model for automatic indexing[J]. Communications of the ACM, 1975, 18(11): 613-620.
[11] Deerwester SC. Indexing by latent semantic analysis[J]. Journal of the American Society for Information Science, 1990, 41(6): 391-407.
[12] Probabilistic latent semantic analysis[EB/OL]. [2019-02-26]. https://en.wikipedia.org/wiki/Latent_{{{sem
[13] Latent dirichlet allocation[EB/OL]. [2019-02-26]. https://max.book118.com/html/2017/0724/123956096.s
[14] Topics over time: a non-Markov continuous-time model of topical trends[EB/OL]. [2019-02-26]. https://link.springer.com/chapter/10.1007%2F978-3-642-12275-0{39}.
[15] Extracting multilingual topics from unaligned comparable corpora[EB/OL]. [2019-02-26]. https://wenku.baidu.com/view/a1b8c42d4b73f242336c5fce.html.
[16] Li Xiangdong, Ba Zhichao, Huang Li. A text feature selection method

based on weighted LDA model and multi-granularity[J]. New Technology of Library and Information Service, 2015, 31(5): 42-48.
[17] Wang Yuefen, Fu Zhu, Chen Bikun. Domestic knowledge flow research using LDA topic model: a perspective of disciplinary classification topic extraction[J]. New Technology of Library and Information Service, 2016, 32(4): 8-19.
[18] Wang Lianxi, Cao Shujin. Comparative analysis of network public opinion research topics from an interdisciplinary perspective: taking domestic library and information science and journalism and communication as examples[J]. Journal of the China Society for Scientific and Technical Information, 2017, 36(2): 159-169.
[19] Xu Tengteng, Huang Hengjun. An improved Supervised-LDA text model and its application[J]. Computer Engineering, 2018, 44(1): 69-73, 78.
[20] Qu Jingye, Chen Zhen, Hu Yinan. Comparative study of co-word analysis and LDA model analysis in text topic mining[J]. Information Science, 2018, 36(2): 18-23.
[21] Wang CH, Zhang M, Ru LY, et al. Automatic online news topic ranking using media focus and user attention based on aging theory[C]//Proceedings of the 17th ACM conference on information and knowledge management. New York: ACM Press, 2008: 1033-1042.
[22] Jiang Xiaowei, Wang Jianmin, Ding Guiguang. Topic ranking method for discovering important topics in microblogs based on topic model[J]. Journal of Computer Research and Development, 2013, 50(5): 179-185.
[23] Liu Peiyu, Hou Xiuyan, Zhu Zhenfang. Microblog hotspot topic discovery based on joint heat ranking[J]. Computer Science and Exploration, 2016, 10(4): 574-581.
[24] Cui W, Liu S, Tan L, et al. Textflow: towards better understanding of evolving topics in text[J]. IEEE transactions on visualization and computer graphics, 2011, 17(12): 2412-2421.
[25] Xiao Zhibo. Research on ranking topic model and its application[D]. Dalian: Dalian Maritime University, 2014.
[26] Xiao Z, Che F, Miao E, et al. Increasing serendipity of recommender system with ranking topic model[J]. Applied mathematics and information sciences, 2014, 8(4): 2041-2053.
[27] Jiang Zhuoren, Gao Liangcai, Zhao Xing. Comparative study on the correlation of Chinese and English science and technology topic ranking: taking the computer field as an example[J]. Journal of the China Society for Scientific and Technical Information, 2017, 36(9): 940-953.
[28] Alam MM, Ismail MA. RTRS: a recommender system for academic researchers[J]. Scientometrics, 2017, 113(1): 1-24.

Author Contributions

Li Xiuxia: Proposed research ideas, designed research plan, wrote paper;
Cheng Jiejing: Designed research plan and revised paper;
Han Xia: Collected and processed data.

ChinaRxiv

The Prioritization of Subject Research Topics Based on the Integration of Publication Trends and Citation Trends: Taking the Subject of Information Science in China as an Example

Li Xiuxia[1], Cheng Jiejing[2], Han Xia[1]

[1]School of Communication, Qufu Normal University, Rizhao 276826
[2]Institute of Social Development, Yangzhou University, Yangzhou 225008

Abstract: [Purpose/significance] Topic sorting is not only the basic problem for information retrieval and information organization, but also an important work of subject service. The effective sorting of subject field research topics can help researchers and research management departments grasp the research situation of the subject field effectively, locate the direction of scientific research accurately and make scientific research decisions quickly. [Method/process] This paper proposes a priority ranking algorithm for disciplinary research topics based on the combination of topic extraction and trend analysis. Then it takes the research topics of Library and Information Science as an example to extract the research topics of the sample literature, and each research topic is divided into four sub-topics: poor theme, hot topic, cold point theme, and overheated topic. Next, priority ranking is carried out in subclasses. [Result/conclusion] The empirical results show that the priority ranking algorithm can display the development level of research topics in an all-round, fine-grained and deep way. This method provides a new perspective for realizing dynamic intelligence analysis from time dimension.

Keywords: writing trend; citation trend; research topic; prioritization