Research on Hotspot Detection Models for Professional Domains Based on Multi-source Data: Postprint

Authors: Wang Xiaoguang, Wang Hongyu, Huang Han

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

Abstract

Purpose/Significance: Aiming at the topic selection decision-making challenges in professional domain publishing for the publishing industry, this study integrates multi-source publicly available information and news dynamics from the internet, detects hotspots within professional domains through multi-dimensional intelligence analysis, realizes data-driven publishing topic selection decisions, and lays a solid foundation for the digital transformation and development of the publishing industry.

Method/Process: An intelligence analysis model is designed for hotspot detection in professional domains oriented toward publishing topic selection decisions. The model comprises two processes: hotspot discovery and heat evaluation. The hotspot discovery process identifies hotspots within professional domains through word frequency statistics and word growth velocity algorithms; the heat evaluation process designs and calculates a series of indicators from two dimensions—content level and dissemination level—to evaluate and rank the identified hotspots.

Results/Conclusion: Using 36,550 multi-source Chinese information items from the information, communications, and technology (ICT) field between January and April 2018 as a sample, hotspot detection experiments were conducted. The experimental results demonstrate that the designed hotspot detection model can effectively detect hotspots within professional domains, assisting the publishing industry in making scientific topic selection decisions for professional domain publishing.

Full Text

Preamble

Vol. 63 No. 14, July 2019
Research on a Professional Field Hotspot Detection Model Based on Multi-source Data
Wang Xiaoguang[1], Wang Hongyu[1], Huang Han[2]
[1]Center for Studies of Information Resources, Wuhan University, Wuhan 430072
[2]School of Information and Safety Engineering, Zhongnan University of Economics and Law, Wuhan 430072

Abstract

[**Purpose/Significance**] To address the topic selection decision-making problem in professional publishing for the publishing industry, this study integrates publicly available information and dynamics from the internet through multi-source consolidation. By conducting multi-dimensional intelligence analysis to detect hotspots within professional fields, it achieves data-driven publishing topic selection decisions, laying a solid foundation for the digital transformation and development of the publishing industry. [**Method/Process**] An intelligence analysis model was designed for hotspot detection in professional fields oriented toward publishing topic selection decisions. The model consists of two processes: hotspot discovery and heat evaluation. The hotspot discovery process identifies hotspots within professional fields through word frequency statistics and word growth rate algorithms. The heat evaluation process designs and calculates a series of indicators from two dimensions—content and dissemination—to evaluate and rank the identified hotspots. [**Result/Conclusion**] Using 36,550 pieces of multi-source Chinese information in the information, communications, and technology (ICT) field from January to April 2018 as a sample, hotspot detection experiments were conducted. The results demonstrate that the designed hotspot detection model can effectively detect hotspots within professional fields and assist the publishing industry in making scientific topic selection decisions for professional domains.

Keywords: topic selection decision; hotspot detection; hotspot discovery; heat calculation; heat evaluation

Introduction

The advent of the "Internet+" and big data era has brought new opportunities and challenges to traditional industries. The traditional publishing industry must adapt to the digital wave and leverage information technology to rapidly acquire industry data and grasp consumer and market dynamics. This requires comprehensive and effective detection and analysis of trending topics and public reading preferences to provide consumers with valuable dynamic content. Book topic planning, as the front-end editorial process in publishing workflows,

involves planning editors from professional field publishers comprehensively analyzing market distribution and sales data as well as publicly available information dynamics to make effective topic selection decisions. Through extensive integration and in-depth analysis of multi-source information, professional field hotspot detection and analysis can be completed, thereby assisting the publishing industry in making data-driven scientific topic selection decisions and laying a solid foundation for digital development. Currently, digital platforms built within the publishing industry mostly focus on terminal marketing services such as e-commerce sales and self-media operations, while lacking effective data analysis and support in the front-end editorial stages like topic planning. Even on the "Kaijuan" data platform widely used in the publishing industry, monitoring is limited to distribution and sales data of published books among publishers, distributors, and retailers, without extending to publicly available information dynamics on the internet.

Literature Review

1.1 Detection of Disciplinary Research Hotspots

When detecting hotspots in specific disciplinary fields, researchers typically use academic literature as the study object, employing bibliometrics, word frequency analysis, and co-word analysis methods. Through word frequency or citation analysis of research literature in a discipline, and conducting co-word analysis or cluster analysis on high-frequency or high-growth-rate keywords, research hotspots in that discipline can be discovered. Commonly used tools include SATI for bibliographic analysis, SPSS for statistical analysis, and CiteSpace for citation visualization. Although most research hotspots in disciplinary fields appear in academic papers, with the continuous growth of digital and online resources related to various disciplines, the research objects for detecting disciplinary research hotspots have gradually expanded to include various information resources beyond scientific literature. For hotspots in these data sources, word frequency analysis and co-word analysis can also be used for detection.

From the perspective of research content in disciplinary fields, hotspots can be divided into general popular research hotspots and potentially important research hotspots. General popular hotspots often focus on newly emerging theoretical concepts and important technologies, with a relatively large number of researchers and publications. Overall, these documents can reflect the research focus of a period, and such hotspots may also have high popularity in other professional fields. Potentially important research hotspots often have strong specialization, and even within the same period, there are significant quantitative differences across different professional fields. In terms of quantity, such research literature often does not dominate. Conversely, only high-quality professional research literature pays more attention to such hotspots. Therefore, keywords in highly cited literature should better reflect potentially important research hotspots than those in less cited literature. Extended to multi-source information dynamics in professional fields, popular hotspots refer to hotspot

topics that are already in a popular state, occupying a relatively large quantity and high frequency in multi-source scientific and technological information, and receiving significant policy attention. Potential hotspots refer to some new keywords, subject terms, and concepts that have gained considerable attention in the latest scientific literature, government bulletins, industry news, and professional institution dynamics, and may subsequently transform into popular themes.

1.2 Identification and Analysis of Public Opinion Hotspots

When conducting public opinion monitoring and analysis, information is typically revealed from different levels such as words, topics, and events. At the micro level, sorting and displaying important keywords in public opinion texts is commonly used to complete monitoring and subsequent analysis. Word frequency statistics and word importance ranking methods are often used for extracting important keywords. At the meso level, topic models such as Latent Dirichlet Allocation (LDA) and clustering methods such as Self-Organizing Maps (SOM) are often employed to reveal public opinion viewpoints at the topic granularity. Identifying and extracting hotspots from public opinion texts is fundamental work for public opinion monitoring and analysis.

Hotspot identification and extraction typically use word frequency statistics and importance ranking methods, which involve extracting keywords from the basic dataset, counting the frequency of each keyword, obtaining a keyword list, calculating word weights, sorting them in descending order of importance, and selecting a certain number of keywords. During keyword extraction, the original text generally needs to undergo preprocessing such as Chinese word segmentation and part-of-speech tagging, followed by selecting appropriate strategies to filter the segmentation results based on subject term tables or stop word tables in the relevant professional fields.

Since some basic vocabulary commonly used in certain professional fields is insufficiently important to be considered a hotspot, the TF-IDF algorithm is typically used to calculate word weights to highlight more important words. The TF-IDF algorithm comprehensively considers the frequency of each keyword in the current data and its frequency in the entire dataset for comprehensive word weight calculation, which can eliminate the impact of common words on subsequent hotspot identification and detection as much as possible. However, for recent and sudden potential important hotspots, due to their high keyword dispersion and large differences from most text set keywords, the traditional TF-IDF algorithm is not very suitable. To address the weak signal recognition problem of the TF-IDF algorithm, researchers have designed indicators such as the growth coefficient of subject terms within a single time window and the word growth rate across time windows to quantify the importance of such words, achieving certain results. In some studies, to better reveal the integrity of public opinion monitoring results, further clustering of extracted keywords is performed through topic models and semantic analysis methods. By mining

associations between different keywords within clusters, macro analysis of public opinion hotspots is achieved.

1.3 Heat Calculation and Evaluation

In hotspot heat calculation, besides common indicators such as word importance, number of related documents, citation counts of related literature, and word growth rate, other indicator calculation algorithms can also be used to quantify heat. For example, Kleinberg's Burst Detection algorithm, proposed in 2002, is often used to calculate the burst weight index of focal words with suddenly increased relative growth rates in document streams. Kleinberg believed that document occurrence is not a smooth growth process but rather a jump growth process within a certain period. Any word in documents can be described as being in either an inactive state or a bursty state, with the bursty state level increasing according to the intensity of the jump. The larger the state value, the more active the word is during that period. Therefore, the Burst index can fully reflect the heat of various hotspot themes within a time period.

After completing the calculation of multiple indicators measuring hotspot heat, a comprehensive evaluation of the heat reflected by these indicators is needed to ultimately determine the hotspot ranking. Currently, there is no unified indicator system for heat evaluation in academia. Some scholars have conducted extensive research on constructing indicator systems for network public opinion heat and Weibo heat. When constructing indicator systems for evaluating public opinion or Weibo heat, they are typically built from three aspects: user characteristics, information dissemination characteristics, and content characteristics. The existence of Weibo opinion leaders reflects the influence of user characteristics on heat. For dissemination characteristics, propagation features such as like rates, comment rates, and repost rates of public opinion texts or Weibo posts can most intuitively reflect the attention heat generated by a text. The influence of content characteristics on heat is reflected in features such as the emotional polarity expressed in the text and the number of texts related to the topic. When determining indicators, specific measurement schemes for each indicator must also be determined. After completing the indicator system construction, weights need to be assigned to each indicator. The weighting method is usually based on group decision-making combined with the Delphi method, using the Analytic Hierarchy Process (AHP) to determine weights. Additionally, when certain evaluation indicators of the evaluation object are relatively fuzzy, making it impossible to draw clear conclusions, fuzzy comprehensive evaluation methods are generally used to calculate these indicators. This method, based on fuzzy mathematics and applying fuzzy relation synthesis principles, conducts comprehensive evaluations of evaluation objects' membership levels from multiple factors. It can better solve problems with fuzzy evaluation indicators and standards, reduce the impact of subjective assumptions, and enhance the accuracy and objectivity of evaluation results.

Model Design

2.1 Data Format

The professional field hotspot detection model for publishing topic selection decisions requires a large amount of the latest multi-source information in specific professional fields as basic data. For scientific literature information, basic data collection involves retrieving scientific papers on professional field-related subject terms from platforms such as CNKI and Wanfang Data, recording the titles, abstracts, and keywords of retrieved papers related to the professional field in a database, along with publication time, authors, author institutions, and citation information. Additionally, information on national science and technology project dynamics needs to be entered in a timely manner, collecting project names, project levels, proposal titles, applicants, application units, proposal abstracts, and funding amounts. For government bulletins and industry news information and professional institution dynamics, professional planning editors need to pre-specify specific information collection sources, such as official websites, Weibo accounts, and WeChat public accounts of government departments in charge of professional fields, authoritative industry news agencies, research groups within professional fields, and large enterprises. Then, titles and content of government bulletins, industry news, and professional institution dynamics are crawled from these specified sources. For Weibo and WeChat public account articles, data reflecting dissemination breadth such as repost volume, comment volume, and like volume are also collected. The specific basic data format required by the model is shown in Table 1 .

2.2 Heat Evaluation Indicators

The professional field hotspot detection model proposed in this study utilizes the Analytic Hierarchy Process (AHP) methodology. Through expert interviews and surveys, the evaluation indicators for professional field hotspot heat were decomposed into multiple levels, establishing evaluation domains at each level, ultimately forming a comprehensive heat evaluation indicator system. Based on two evaluation dimensions—content and dissemination—of professional field hotspot heat, two evaluation criteria were determined. By analyzing specific indicators obtainable in different evaluation dimensions, eight specific indicators were proposed. All proposed indicators are positively correlated with the heat of professional field hotspots. The specific evaluation indicator system and indicator symbols are shown in Table 2 .

The detailed introduction of each indicator is as follows:

(1) Content Dimension (A). The content dimension refers to analyzing the heat of professional field hotspots from the content of literature, information, news, and dynamics related to professional fields that can be obtained on the internet. It is mainly evaluated based on the total amount of basic data collected in the current period, the amount of data containing relevant hotspots, and data obtained from comparisons with previous periods, including burst index, scaling

speed, and scaling acceleration.

- **A1 Relative Document Quantity (r):** The percentage of data related to a certain candidate professional field hotspot among all basic data collected in the current time period.
- **A2 Feature Word Frequency Proportion (f):** For each piece of basic data collected in the current period, feature words are extracted to obtain its feature word list. The percentage of data where a certain candidate hotspot theme appears as a feature word among the total data where the candidate hotspot appears as a feature word.
- **A3 Burst Index Heat (b):** Using Kleinberg's burst detection algorithm, the burst index is calculated from the cost-benefit between burst and non-burst states of candidate professional field hotspots across multiple periods including previous time periods.
- **A4 Scaling Speed (v):** The ratio of the number of documents related to a certain candidate professional field hotspot in the current period to that in the previous time period.
- **A5 Scaling Acceleration (α):** The ratio of the proportion of documents related to a certain candidate professional field hotspot in the current period to that in the previous time period.

(2) Dissemination Dimension (B). In the dissemination dimension, evaluation is mainly based on the total volume of Weibo and WeChat public account posts by professional institutions, government departments, or industry news agencies related to the professional field during the current period, the total volume of related reposts, comments, and likes, and the quantity of government bulletins and industry news.

- **B1 Weibo/WeChat Post Relative Quantity (q):** The percentage of data related to a certain candidate professional field hotspot among all Weibo and WeChat public account articles collected in the current time period.
- **B2 Government Bulletin and Industry News Relative Quantity (t):** The percentage of data related to a certain candidate professional field hotspot among all government bulletins and industry news collected in the current time period.
- **B3 Repost/Comment/Like Relative Quantity (Relative Dissemination Breadth) (e):** The sum of reposts, comments, and likes of Weibo and WeChat public account articles related to a certain candidate professional field hotspot collected in the current time period is defined as the dissemination breadth of that hotspot. The percentage of this hotspot's dissemination breadth among the total dissemination breadth of all candidate hotspots is defined as the relative dissemination breadth.

2.3 Model Structure and Detection Process

Combining relevant ideas and solutions from the literature review, the professional field hotspot detection model is ultimately designed to be divided into two major processes: professional field hotspot discovery and hotspot theme heat calculation. The overall structure of the model is shown in Figure 1 [Figure 1: see original paper].

After basic data collection is completed, the original input data fields in the basic data need to be concatenated. The concatenated strings are processed through Chinese word segmentation and stop word removal, and the resulting word lists serve as the raw input dataset for the model. Let D represent the raw input data collection containing n pieces of data collected in the current period T=t. Then D = {D_1, D_2, …, D }, where D represents the i-th piece of collected data, and each piece of data contains several words {T $_1$, T $_2$, …, T , …}.

The specific process of the professional field hotspot detection model is briefly introduced below in combination with the model's heat evaluation indicator system and overall structure:

(1) Feature Word Extraction. The first step of the model is to extract feature words G from the word lists of each piece of data D in the raw input dataset D. First, each data's word list is filtered through word segmentation and stop word removal to obtain a feature word list for each piece of data. Then, the TF-IDF algorithm and word growth rate algorithm are used to comprehensively calculate word weights. The TF-IDF algorithm can comprehensively consider the frequency of each feature word in the current piece of data and its frequency in the entire dataset for comprehensive word weight calculation, which can eliminate the impact of common words on subsequent hotspot identification algorithms and highlight important words. The word growth rate algorithm can pay more attention to recent and sudden potential important hotspots, making such words noticeable.

Based on relevant literature [7], the formula for the word growth rate algorithm is shown in Formula (1). Where G , represents the word growth rate of word k in the word list of a piece of data in the raw input dataset D collected in the current period T=t, F , represents the word frequency of vocabulary k in time window T=t, t-u+1 is the size of the backtracking window (i.e., the backtracking window calculated is from time period T=u to time period T=t), mean(F ,) represents the average frequency of word k in the backtracking window, and sp is a smoothing coefficient given by Formula (2).

In Formula (2), length(D) represents the number of words in the raw input dataset collected in time window T=u, while |V | represents the number of word items in the raw input dataset collected in time window T=u (i.e., how many different words it contains).

(2) Candidate Hotspot Identification. The second step of the model is to identify candidate hotspots from the collection of feature word lists. The

specific steps are as follows: merge the feature word lists of all data in collection D, and take the top several words with the highest word frequency as candidate popular hotspots. For potential hotspot identification, it is necessary to weight each piece of data in collection D based on relevant data fields such as institution names and authors recorded in the basic data (the more authoritative the institution and author, and the more widely disseminated the data, the higher the weight, taking natural numbers). Then, repeat each piece of data in the feature word list collection according to its weight times, merge them, and the top several words with the highest word frequency in the newly formed collection are the candidate potential hotspots.

(3) Content Dimension Parameter Calculation. The third step of the model, after obtaining the candidate hotspot theme list, is to complete the content dimension parameter calculation for these candidate hotspot themes. The Burst algorithm used in the model for burst index heat calculation is highlighted here, as it is commonly used to calculate the burst index of words with suddenly increased relative growth rates across several periods in a text data stream.

For the burst index, the Burst algorithm is first used to calculate the benefit C of candidate hotspot T in the current period, as shown in Formula (3). Where β is the empirical value of the algorithm, r' represents the number of data pieces in collection D containing word T , N represents the number of data pieces in collection D' summarizing all data across several periods, and R represents the number of data pieces in collection D' containing word T . The burst index heat is calculated by accumulating the benefits of candidate hotspots across several periods.

Scaling speed is the ratio of the number of data pieces containing candidate hotspot T in the current period to that in the previous period. Scaling acceleration is the ratio of the proportion of data pieces containing candidate hotspot T in the current period to that in the previous period, to eliminate the impact of different total data quantities across periods.

(4) Dissemination Dimension Parameter Calculation. The fourth step of the model considers factors of mass dissemination in addition to quantifying hotspot heat based on the content dimension. In the dissemination dimension, it mainly calculates the proportion of several types of sources related to candidate hotspot theme words in the current period, including Weibo and WeChat public account articles, government bulletins, and industry news, in the complete dataset of each type of data. Since data in the government bulletin and industry news category are more rigorous and authoritative, the quantity proportion of government bulletins and industry news is calculated as one category, while the quantity of Weibo and WeChat public account articles is calculated as another category. Additionally, for Weibo and WeChat public account articles, from the perspective of dissemination breadth, the relative proportion of reposts, comments, and likes of Weibo and WeChat public account articles related to a certain candidate hotspot theme word should also be calculated.

(5) Heat Evaluation and Ranking. The fifth step of the model requires combining the heat parameter indicators of the two dimensions obtained in the third and fourth steps of the model. Through the fuzzy analytic hierarchy process, the specific evaluation matrix is determined, and comprehensive evaluations of the heat of the two types of candidate hotspots are conducted to complete the ranking of hotspots in the candidate hotspot theme list. Since there are some factors in the content dimension parameters and dissemination dimension parameters that are not easily evaluated clearly, the characteristics of fuzzy language variables and fuzzy numbers that can quantify fuzzy information can be used to apply the analytic hierarchy process and fuzzy comprehensive evaluation method. From a multi-factor perspective, the evaluation matrix is constructed to conduct comprehensive quantitative evaluation of the heat of candidate hotspot themes. Comprehensive evaluations are conducted on the two types of hotspot theme candidate lists separately to obtain the final hotspot heat, and ranking is performed based on the final quantitative evaluation results to facilitate users' direct judgment of professional field hotspots. At the same time, by analyzing the heat changes of different hotspots across several periods, professional field hotspot detection and analysis can be completed.

Experiments

3.1 Data Preparation

When conducting hotspot detection experiments on the ICT field, it is necessary to determine the weights between the criteria layer and specific indicator items within each criterion in the hotspot heat evaluation indicator system, so as to ultimately obtain quantifiable heat calculation results. The author conducted multi-round telephone surveys and field interviews with seven professional publishing practitioners from representative publishing institutions in the ICT field—Electronic Industry Press, Posts and Telecom Press, and Hubei Science and Technology Press—to fully understand the overall business process of professional field publishing and the specific influencing factors of professional field topic selection decisions. The Delphi method was used to determine the weights of the heat evaluation indicator system.

To ensure that the professional field hotspot detection model can detect hotspots in professional fields in a timely and accurate manner, the model's multi-source basic data and evaluation matrices determined by the fuzzy analytic hierarchy process need to be updated regularly. The author selected the Information, Communications, and Technology (ICT) field and conducted hotspot detection experiments on topic selection hotspots in March and April 2018. Since the detection of potential hotspots requires professional planning editors to conduct a large amount of weight presetting work based on relevant data fields recorded during basic data collection, the author only conducted experiments on popular hotspots.

According to the model's basic data collection specifications, the Octopus data

collector was used to crawl government bulletin data from websites such as the Ministry of Industry and Information Technology and the Ministry of Science and Technology, news data from portals such as Sina and Sohu, and Weibo data related to topics such as “communication” and “technology” was obtained through the Weibo API. At the same time, scientific paper data was collected by searching CNKI with themes such as “internet” and “information technology.” Additionally, to calculate indicators such as cross-period burst index heat, scaling speed, and scaling acceleration, relevant data from January and February 2018 were crawled to form the initial dataset for this study.

To accurately detect popular hotspots in the ICT field, the obtained initial dataset was preprocessed to filter out texts with no content or few words, forming an experimental dataset containing 36,550 texts. The distribution of experimental data is shown in Table 3 . For this dataset, Chinese word segmentation was performed using the IKAnalyzer toolkit, and punctuation, symbols, pronouns, prepositions, and conjunctions were further removed by constructing a stop word table. At the same time, since news content usually contains its source information, phrases such as “Sohu Technology” and “Sina Technology” were also added to the stop word table to make the segmentation results more accurate. Additionally, in the feature word extraction step, words without specific references such as “company” and “news” were filtered by constructing a non-feature word table, thereby ensuring that the final identified candidate hotspots have higher comprehensibility.

3.2 Experimental Results

When actually conducting popular hotspot detection experiments on the ICT field, it is necessary to determine the weights of the criteria layer and specific indicators within each criterion in the hotspot heat evaluation indicator system to ultimately obtain quantifiable heat calculation results. The author conducted multi-round telephone surveys and field interviews with seven professional publishing practitioners from representative publishing institutions in the ICT field—Electronic Industry Press, Posts and Telecom Press, and Hubei Science and Technology Press—to fully understand the overall business process of professional field publishing and the specific influencing factors of professional field topic selection decisions. The Delphi method was used to determine the weights of the heat evaluation indicator system.

Multiple experts in the professional publishing field conducted multi-round comparisons of indicator importance until the constructed membership matrix passed the consistency test. Afterward, the author transformed the membership matrix into a weight vector and conducted comprehensive evaluation of the final evaluation target based on the values of each indicator. According to the opinions of experts in the ICT field and relevant literature [1,12,25], the weight vector for the professional field hotspot heat evaluation matrix was ultimately determined as:

- Criteria layer weight vector: (0.5, 0.5)
- Indicator layer (A) weight vector: (0.1, 0.5, 0.3, 0.05, 0.05)
- Indicator layer (B) weight vector: (0.2, 0.5, 0.3)

After obtaining the weight values, the content dimension and dissemination dimension parameters of professional field hotspots were calculated according to the proposed model, and heat calculation and ranking were completed. Arranged in descending order according to the values of each parameter indicator and the final heat calculation results, the top 20 popular hotspots and their rankings in the ICT field for March and April 2018 are shown in Table 4 and Table 5 .

Through Tables 4 and 5, the trend of popular hotspot changes in China's ICT field between March and April 2018 can be further detected and analyzed, as shown in Figure 2 [Figure 2: see original paper].

By analyzing the experimental results, it can be found that the proposed model effectively achieves the detection of popular hotspots: In March 2018, the release of the new "R15" phone with AI photography features brought widespread attention from Chinese users to the mobile phone company OPPO. On the other hand, blockchain technology, developing rapidly in the financial industry, was identified by China's internet service companies as a key focus for future technology development and innovation, alongside big data and artificial intelligence. In April 2018, the U.S. sanctions against ZTE in the technology field made China realize the need to systematically encourage universities, research institutions, and the information technology industry to achieve independent design and development of smart chips, promoting innovation-driven development through technological innovation.

Summary and Discussion

The professional field hotspot detection model proposed by the author for publishing topic selection decisions integrates multi-source information dynamics publicly available on the internet, including government bulletins and industry news, professional institution dynamics, and scientific literature. Combining elements from both content and dissemination dimensions, the model completes professional field hotspot detection and heat calculation through steps such as feature word extraction, candidate hotspot identification, content and dissemination dimension parameter calculation, and heat evaluation and ranking, achieving detection of professional field hotspot change trends. The proposed model helps relevant personnel engaged in professional field publishing to comprehensively and multi-dimensionally detect and analyze hotspot topic changes and mass communication trends in professional fields automatically, thereby achieving an upgrade from previous experience-led topic selection decisions to data-driven scientific topic selection decisions. At the same time, this study also provides practical data analysis support for the topic planning stage in the publishing workflow.

The author focuses on the topic selection decision-making stage in professional field publishing and systematically designs a professional field hotspot detection model constructed based on publicly available multi-source information dynamics, conducting intelligence analysis application practice. The model combines TF-IDF and word growth rate algorithms to calculate word weights for candidate topic hotspots. Simultaneously, it objectively and multi-dimensionally designs and selects indicator parameters such as burst heat index, scaling speed and acceleration, and dissemination breadth from two dimensions, constructs a hotspot heat evaluation indicator system oriented toward publishing topic selection decisions, and finally completes hotspot heat evaluation and ranking through the fuzzy analytic hierarchy process. Using 36,550 pieces of multi-source Chinese scientific and technological information in the ICT field from January to April 2018, hotspot detection experiments verified the model's effectiveness in detecting popular hotspots in Chinese professional fields.

When publishing institutions actually conduct professional field publishing, they undergo a relatively complex topic selection decision-making process. According to expert research, after planning editors submit topic proposals, small and medium-sized publishers mostly adopt the form of editorial department meetings for group decision-making before reporting topics. Large publishers typically use management information systems to manage the business processes of multi-level participation in topic selection decision-making, improving the circulation efficiency of the topic selection decision-making process. Afterward, publishing institutions need to report the completed topic plans to the competent administrative departments, and the overall topic selection decision-making process ends after obtaining administrative approval. The model proposed in this study can be used for automated collection, processing, and analysis of Chinese multi-source information in subsequent applications, achieving automatic detection of topic hotspots, thereby reducing the workload of professional field publishing practitioners in collecting, processing, and analyzing relevant topic materials and information. On the other hand, the model proposed in this study also lays a good service foundation for conducting subsequent industrial trend analysis and hotspot tracking consulting in professional fields.

This study also has some limitations. For example, the fuzzy analytic hierarchy process is an evaluation method that relies on group decision-making and expert scoring, inevitably having some subjective shortcomings. At the same time, this study does not conduct further cluster analysis and association analysis on the detected professional field hotspots, requiring certain manual intervention when professional field publishing practitioners conduct actual topic selection decision-making. In future research, machine learning algorithms will be attempted to use historical market data to train regression models to scientifically and automatically determine the weights of each evaluation indicator, making the quantitative heat calculation results more objective and accurate. Additionally, cluster and association analysis models will be used to conduct in-depth mining of topic hotspot detection results, further providing more effective data support for the professional field publishing topic selection decision-making process.

References

[1] Zeng W, Xu HJ, Che Y, et al. Research on topic selection decision analysis model based on big data of book publishing industry [J]. Journal of the China Society for Scientific and Technical Information, 2018, 37(8): 813-821. [2] Huang Z. Digital navigation leads traditional publishing into a new era [J]. Publishing Wide Angle, 2018(17): 35-37. [3] Wang HW, Gao S, Lu B. Research on online news hotspot identification based on LDA and SNA [J]. Journal of the China Society for Scientific and Technical Information, 2016, 35(10): 1022-1037. [4] ASATANIK, MORIJ, OCHIM, et al. Detecting trends in academic research from citation network using network representation learning [J]. Plos One, 2018, 13(5): e0197260. [5] SALMERON-MANZANO E, MANZANO-AGUGLIARO F, ENERGIES, et al. The electric bicycle: worldwide research trends [J]. Energies, 2018, 11(7): 1894. [6] Cheng QK, Wang XG. A framework for analyzing the evolution of research topics based on co-word network communities [J]. Library and Information Service, 2013, 57(8): 91-96. [7] Zhuang TT, Wang P, Cheng QK. A time-context-dependent method for Weibo topic extraction [J]. Journal of Information Resources Management, 2013(3): 40-46. [8] Chen W, Lu W, Han SG. Design and implementation of expert retrieval and hotspot detection system [J]. Journal of Intelligence, 2009, 28(12): 113-117. [9] WANG H, LIU C, ZHAO Z, et al. Efficiency evaluation of an Internet Plus University Student Affairs System based on fuzzy theory and the analytic hierarchy process [J]. Journal of Intelligent & Fuzzy Systems, 2016, 31(6), 3121-3130. [10] Yang CJ, Cheng G. Research on the evaluation system of knowledge service capability of scientific and technical intelligence agencies [J]. Information Studies: Theory & Application, 2017, 40(7): 43-49. [11] Li SQ, Bai Y. Analysis of research trends in library and information science based on time-series keyword hotspot identification method (2000-2009) [J]. New Technology of Library and Information Service, 2011, 27(5): 69-76. [12] He Y, Cai BC. Weibo heat evaluation model based on factor analysis [J]. Statistics & Decision, 2016(18): 52-54. [13] Li X, Li XH, Lu W, et al. Mining hot fields in library and information science driven by big data: an empirical perspective on WOS bibliographic data [J]. Library Tribune, 2017, 37(4): 49-57. [14] Lu W, Peng Y, Chen W. Domain hotspot topic detection based on SOM [J]. New Technology of Library and Information Service, 2011, 27(1): 63-68. [15] Zheng K, Shu XM, Yuan HY. Automatic discovery method for hotspot information in network public opinion [J]. Computer Engineering, 2010, 36(3): 4-6. [16] Chen XM, Gao C, Guan XH. LDA topic model method for network public opinion viewpoint extraction [J]. Library and Information Service, 2015, 59(21): 21-26. [17] Yang YF, Yu WP, Tian P. Research on classification prediction of brand scandal Weibo communication based on SOM neural network [J]. Journal of Intelligence, 2013(10): 23-28. [18] Wu XJ. Analysis of network public opinion theme evolution based on Weibo text: a case study of the "Blue Qianjiang Arson Case" [D]. Nanjing: Nanjing University, 2018. [19] Liu HF, Yu LJ, Liu SS. A text feature selection model based on category distribution information [J]. Library and Information Service, 2013, 57(15): 137-141. [20] YU B, YU YH. Auto-Tracking controversial topics in social-media-

based customer dialog: a case study on starbucks [C]//GOBINDA C, JULIE M. Lecture notes in computer science, volume 10766. Berlin: Springer, 2018: 87-96. [21] LIN, WU D. Using text mining and sentiment analysis for online forum hotspot detection and forecast [J]. Decision support systems, 2010, 48(2), 354-368. [22] KLEINBERG J. Bursty and hierarchical structure in streams [J]. Data mining & knowledge discovery, 2003, 7(4): 373-397. [23] Gao YB, Yang GP, Zhang D, et al. Official microblog event detection method based on burst word blog clustering [J]. Data Analysis and Knowledge Discovery, 2017, 1(9): 57-64. [24] Yang XH, Cai ZQ. Emerging trend detection of linked data based on burst detection and co-word analysis [J]. Information Science, 2018, 36(11): 164-168. [25] Sun FX, Cheng SH, Jin XT, et al. Quantitative evaluation method for government negative network public opinion heat: a case study of Sina Weibo [J]. Journal of Intelligence, 2015(8): 137-143.

Author Contributions

Wang Xiaoguang: Proposed research ideas, revised and polished the paper.
Wang Hongyu: Designed research plan, designed and conducted experiments, drafted the paper.
Huang Han: Collected experimental data and conducted experiments.

English Title and Abstract

Towards Professional Publishing: Research on Hotspot Detection Model Based on Multi-source Data

Wang Xiaoguang[1], Wang Hongyu[1], Huang Han[2]
[1]Center for Studies of Information Resources, Wuhan University, Wuhan 430072
[2]School of Information and Safety Engineering, Zhongnan University of Economics and Law, Wuhan 430072

Abstract: [Purpose/significance] In order to solve the problem of topic selection for professional fields in publishing industry, this paper integrates multi-source dynamic information publicly available on the internet, detects hotspots in professional fields through multi-dimensional intelligence analysis, realizes data-driven publishing topic selection decisions, and lays a solid foundation for the digital transformation and development of publishing industry. [Method/process] An intelligence analysis model was designed for hotspot detection in professional fields oriented toward publishing topic selection decisions. The model was divided into two steps: the hotspot discovery and the hotness evaluation. The hotspot discovery in this model identified hotspots in professional fields through word frequency statistics and the algorithm of word growth rate. Then, in the step of hotness evaluation, a series of indices in the dimension of content and spread were designed to calculate and evaluate the hotness of the hotspots identified in the last step. [Result/conclusion] A hotspot detecting experiment was conducted with 36,550 pieces of Chinese multi-source dynamic information in the area of ICT collected from January

to April of 2018, which verified the effectiveness of the proposed model. This model can be used in publishing industry to complete the step of topic selection scientifically.

Keywords: topic selection; hotspot tracking; hotspot detection; hotness calculate; hotness evaluation

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