ChinaRxiv

AI translation · View original & related papers at
chinarxiv.org/items/chinaxiv-202304.00220

Multi-Metric Bursty Term Detection and Validation: Postprint

Authors: Feng Guohe, Wu Jiajia, Mo Xingqing

Date: 2023-04-01T16:15:55+00:00

Abstract

[Purpose/Significance] To effectively detect potential research hotspots in scientific literature, investigating the characteristic conditions of keyword bursts and constructing a burst term recognition model holds significant importance for enabling researchers to accurately grasp research directions. [Method/Process] Keywords and their frequencies are obtained for each year to construct a keyword-year matrix. The analysis timeframe is divided into a standard window, observation window, and performance window. Within the observation window, a multi-measure burst term detection model is employed to identify keywords exhibiting burst characteristics; within the performance window, LDA is utilized to mine thematic vocabulary as a hotspot term set. A burst term coverage rate metric is designed, and with the assistance of the sliding time window method, the coverage rates of burst term sets and hotspot term sets across different time windows are calculated to verify the accuracy of model identification. [Results/Conclusion] Through three iterations of sliding time windows, the calculated burst term coverage rates were all above 70% in each case; in a comparative experiment with Citespace burst terms, the coverage rates of this model exceeded those of the former in all three instances, demonstrating that the designed burst term detection model exhibits good performance.

Full Text

Research on Detection and Verification of Burst Words with Multiple Measures

Authors: Feng Guohe, Wu Jiajia, Mo Xingqing
Affiliation: Department of Information Management, School of Economics and Management, South China Normal University, Guangzhou 510006

Abstract: [Purpose/Significance] Effectively detecting potential research hotspots in scientific literature, studying the characteristic conditions of keyword bursts, and constructing a burst word recognition model is of great significance for promoting researchers to accurately grasp research directions. [Method/Process] This study obtained keywords and their frequencies for each year, constructed a keyword-year matrix, divided the analysis period into standard window, observation window, and performance window, used a multi-measure burst word detection model to identify keywords with burst characteristics in the observation window, and employed LDA to mine topic vocabulary as a hotspot word set in the performance window. A burst word coverage index was designed, and the sliding time window method was used to calculate the coverage of burst word sets and hotspot word sets in different time windows to verify the accuracy of model recognition. [Result/Conclusion] Using three sliding time windows, the calculated burst word coverage rates were all above 70%. In the control experiment with CiteSpace burst words, the coverage rates of this model were greater than the former in all three cases, indicating that the designed burst word detection model performs well.

Keywords: burst word detection; sliding time window; multiple measures; LDA topic mining
Classification: G250
DOI: 10.13266/j.issn.0252-3116.2020.11.008

1 Introduction

Burst words refer to keywords with relatively low frequency but increasingly strong growth momentum, indicating that the keyword is receiving increasing attention from scholars in the discipline and has a high probability of developing into a research hotspot in the future. The development of things follows the basic life cycle theory, and keywords are no exception. In the process of scientific communication, keyword development can be roughly divided into four stages: germination, development, maturity, and decline [1]. As a concentrated embodiment of the themes and core concepts of journal articles, keywords reveal to a certain extent the research content and topics of papers. Using keywords as the analysis object for burst word detection in disciplinary fields and identifying keywords with burst characteristics in advance during the germination period can help scholars grasp disciplinary research trends and determine future research hotspots. Burst word detection is an important issue in the field of informetrics research both domestically and internationally.

2 Literature Review

Current domestic and international research on burst word detection methods can be broadly divided into three categories:

(1) Burst word identification based on word frequency growth rate. The typical representative is the burst detection algorithm (BDA) proposed by

J. Kleinberg [2]. This algorithm argues that the importance of a word lies not in how long it appears but in its density when it appears; that is, words with suddenly increased relative frequency growth are burst words [3]. Scholars both domestically and internationally have conducted extensive research based on BDA and achieved phased results. C. M. Chen [4] developed CiteSpace based on BDA for visual analysis of burst word detection, providing researchers with a simple and easy-to-use tool for topic detection and evolution analysis [2,5]. Tang Xiaobin et al. argued that Kleinberg' s use of the Viterbi algorithm to judge abnormal events based on only 10 pieces of information about whether they are in an abnormal state is unreasonable, as BDA would misjudge states with slowly changing information frequency over time as burst anomalies. They proposed an improved BDA algorithm to address these defects and successfully detected microblog emergencies [6]. With rich research achievements, burst topic detection performs particularly prominently in social media. Unlike previous research results on burst word detection, this study designs a burst word detection model based on multi-dimensional characteristics of burst words in scientific literature, uses sliding time windows to verify results, and compares them with CiteSpace burst word detection results.

(2) Burst word identification based on multi-feature fusion. The typical representative is Chen Guolan' s use of three indicators—relative word frequency, word frequency growth rate, and burst word weight—to identify burst words in microblog text, using co-word analysis to cluster burst word-related events and successfully extract microblog emergencies [8]. Lu Wanhui et al. argued that single words cannot express complete semantic information and that it is necessary to explore the evolution of knowledge in a field from the perspective of domain terminology. After constructing a terminology feature lexicon, they successfully identified burst words in nickel-cobalt industry patent texts using three indicators: frequency, rate, and word frequency-document ratio [9]. Jie Fei et al. argued that using only text features (keywords) or social behavior (comments, likes, forwards) features would cause missed detection of implicit burst events in social networks. By correlating burst results obtained from keyword features and behavior features, they effectively identified implicit burst events in comparative experiments [10]. W. Xie et al. used three indicators—total number of tweets, word frequency, and word pair frequency—to identify burst topics in Twitter, using acceleration calculation to timely reflect bursts, but this model might ignore topics that do not show bursts in the short term [11].

(3) Improving burst word detection methods by borrowing theories from other disciplines. Wang Liya combined the principle of information entropy change, judging the burst degree of data by observing entropy changes before and after adding data to the dataset, successfully solving the defect that the evolution stage of topic development divided into 2-year, 5-year, or 10-year units is subjective and unreasonable [12]. Wang Zheng et al. argued that keywords are the smallest units carrying various scientific concepts in scientific journals and proposed the SRHM model based on power spectral density theory

and grey correlation theory. Its simulation experimental effect was better than CiteSpace burst word detection, but it did not show the burst word identification results [13]. Zhang Jinzhu et al. argued that calculating similarity or correlation of topics in adjacent time periods is the core of topic evolution and mutation identification, but point similarity and relationship similarity ignore the overall network structure and are not suitable for actual networks. Therefore, by comprehensively considering node quantity and importance and combining strategic coordinate diagrams, they successfully detected topic evolution processes and mutation topics in the gene editing field in the WoS dataset [14]. Jiang Xin et al. argued that keyword frequency in small sample data is low and volatile, and it is not appropriate to reflect change trends by calculating Z-scores and moving averages of word frequency. Therefore, they used log-likelihood values to reflect the significance of keyword frequency changes. By eliminating the impact of fluctuations in scientific research output in different periods on keyword change trends, this method successfully identified the theme evolution process based on burst words in the field of scientific data [15]. With the widespread application of deep learning technology, some scholars have begun to detect burst words through deep neural networks. For example, L. Shi et al. proposed a sparse topic model (STRM) for social network data such as Weibo and Facebook, using RNN to learn the intrinsic relationship between words and IDF to measure high-frequency words. The model distinguishes burst topics from public topics based on vocabulary diversity [16].

Existing research results have the following problems: Burst word detection methods have their own shortcomings. The first category of methods has good identification effects for rapidly circulating data streams such as Weibo but is not applicable to journal literature with slower circulation speeds.

3 Methodology

This study proposes a multi-measure burst word detection and verification model. The specific steps are as follows:

Step 1: Data Preprocessing

Collect keywords annually, count their frequencies, and perform processing such as merging synonyms and near-synonyms and removing function words. If a word ranks high for consecutive years, it is considered a professional basic vocabulary or already a hotspot vocabulary and is not included in burst word analysis. The annual frequency distribution of keywords conforms to a long-tail distribution [19]. Keywords with frequencies of 1 or 2 are at the tail. If a vocabulary only appears once or twice in many years, it is considered that the low-frequency word does not have burst characteristics and is also not included in burst word analysis. Further screening is performed on middle-ranking keywords. According to the Pareto principle [20], the top 20% of middle-frequency words are more analytically meaningful than the remaining 80%, so these 20%

of middle-frequency words are used as the analysis objects for burst words in this study to construct a keyword-year matrix F_{m×n}.

$$F_{m\times n} = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{pmatrix}$$

where m represents the number of keywords, n represents the total number of years, and a_{uv} (u = 1,⋯,m; v = 1,⋯,n) represents the frequency of the uth keyword in the vth year.

Step 2: Time Window Division

In the keyword-year frequency matrix, to verify the stability of the model in identifying burst words in different sample matrices, the analysis period is divided into K sample matrices based on the time dimension, defined as A_{m×i}, B_{m×(i+1)}, C_{m×(i+2)} (i+2 < n). To reduce the impact of time window length on burst word identification, each sample matrix is set to the same time window length. At the same time, to ensure data diversity in sample matrices, a window sliding threshold T is set (i.e., sliding T length time units each time). This parameter can be adjusted according to actual needs. For example, to observe burst word changes within one unit year, T can be set to 1; to observe burst word changes within two or more unit years, T can be set to 2 or a number greater than 2. To ensure temporal continuity of burst word changes, this study sets T to 1, meaning that starting from sample matrix A_{m×i}, sliding one unit year sequentially yields multiple sample matrices B_{m×(i+1)}, C_{m×(i+2)}⋯

Each sample matrix is divided into three window data matrices. Taking sample matrix A_{m×i} as an example, the i unit years are divided into three time windows. The earliest time window is the standard window, the middle window is the burst word detection window (called the observation window), and the latest window is the hotspot topic detection window (called the performance window). These three windows are defined as AT1, AT2, and AT3 respectively. To meet the needs of keyword burst quantitative changes and computability, the window lengths of AT1, AT2, and AT3 are set to 3 unit years each. The standard window data serves as the comparison standard for keyword burst changes. Data in the observation window is judged based on burst word characteristic conditions, and keywords meeting the conditions are included in the burst word set. All frequency words in the performance window are mined for hotspot topics through LDA [21], and thresholds are set to select TopN keywords with probability values in each topic into the hotspot word set.

By sliding a fixed window size by one unit year, the coverage rates of three sample matrices A_{m×i}, B_{m×(i+1)}, C_{m×(i+2)} (i+2 < n) are obtained, expressed as p_A, p_B, p_C respectively. The sliding window and analysis

matrix designed in this study are shown in Figure 1 [Figure 1: see original paper].

Step 3: Burst Word Identification Metrics

Comprehensively considering burst word characteristic indicators in existing literature, this study excludes topic tag indicators for microblog short texts [22] and TF-IDF weight indicators [23], and designs and agrees that burst words should be in the observation window and meet the following basic conditions: Quantitative condition: The total word frequency of keywords should reach a certain amount, causing qualitative changes and subsequent bursts; Trend condition: The word frequency of keywords increases year by year, showing an upward trend; Volatility condition: The word frequency of keywords fluctuates greatly, with strong discriminability.

Since the update cycle of journal literature is long and the above condition cannot be met within a unit year, the sum of keyword frequencies in the three time windows is calculated separately. Based on the above basic conditions, the following heuristic descriptive quantities are set:

(1) Relative word frequency. Calculate the ratio of keyword frequency to the maximum keyword frequency in the current window, as shown in Formula (1):

$$X = \frac{\sum a_{mn}}{\max(\sum a_{mn})} \quad (1)$$

where X represents the relative word frequency of keyword M in n years, and max(a_{mn}) represents the maximum word frequency value in n years. X examines the vertical change trend of keywords. The larger the X value, the greater the popularity of the keyword in the time window and the more likely it is to become a hotspot vocabulary in the future. Taking sample matrix A_{m×i} as an example, in the corresponding AT1, AT2, and AT3 time windows, corresponding relative word frequencies X_{AT1}, X_{AT2}, and X_{AT3} are obtained through Formula (1).

(2) Word frequency growth rate. Calculate the ratio of word frequency growth in the current window relative to the previous window, as shown in Formula (2):

$$Z = \frac{\sum a_{mn} - \sum a_{m(n-1)}}{1 + \sum a_{m(n-1)}} \quad (2)$$

where Z represents the growth rate of keyword M' s word frequency in n years relative to n-1 years. The 1+ a_{m(n-1)} in the denominator avoids division by zero when the keyword did not appear in the previous time period. Z examines the horizontal change trend of keywords. The larger the Z value, the more

obvious the growth trend of the keyword, and the more likely it is to become a hotspot vocabulary.

(3) Word frequency heat weight. Calculate the ratio of keywords appearing in scientific literature titles, as shown in Formula (3):

$$H = \frac{\sum t\{a_{mn}\}}{\sum title} \quad (3)$$

where H represents the ratio of the number of times keyword M appears in titles to the total number of titles in the current period, t{a_{mn}} represents the number of titles containing the keyword, and title represents the total number of literature titles. The larger the H value, the more times the keyword appears in titles, the greater its heat in the current time window, and the more likely it is to become a hotspot word in the future.

(4) Keywords screened based on descriptive quantities X, Z, and H are included in burst word candidate sets X_{ht}, Z_{ht}, and H_{ht} respectively. A threshold s is set, and keywords ranking greater than s in each set are included in the burst word set T. Mathematically:

$$T = \{X_{ht} \cap Z_{ht} \cap H_{ht}\} \quad (\text{descriptive quantity} > s) \quad (4)$$

Step 4: Hotspot Word Extraction via LDA

Hotspot words are keywords with high and stable frequency in the performance window. The hotspot word acquisition range should be larger than the burst word analysis range to ensure the possibility of burst words becoming hotspot words in subsequent time windows. The LDA language model is a three-layer Bayesian probability model [24] containing word, topic, and document structures. Compared with co-word analysis for hotspot mining, LDA has three advantages [1]: No need to determine the boundary between high-frequency and low-frequency keywords; LDA can reflect deep semantic relationships between topics; It avoids the subjectivity of keyword selection in co-word analysis.

Using this model to mine hotspot topics and keywords contained in each topic in the performance window yields the hotspot words needed in this study. The document collection is defined as D = {d_1,⋯,d_p}, where d_p represents the pth document, d_p = {x_1,⋯,x_j}, and x_j represents the jth vocabulary in the pth document. Topic symbols are defined as E = {k_1,⋯,k_o}, where k_o represents the oth keyword in the topic. The hotspot keyword calculation formula is:

$$P(k|d) = P(k|e) * P(e|d) \quad (5)$$

where k, d, and e represent keywords, documents, and topics respectively. Based on Formula (5), topics in the document collection and keywords contained in

each topic are obtained. The LDA hyperparameter q is set to adjust the number of topics generated by LDA. Topic vocabulary with probability values less than q is included in the hotspot word set, defined as R.

Step 5: Coverage Rate Calculation for Model Validation

To verify the effectiveness of the model in identifying burst words, a coverage discrimination index is proposed, which selects common vocabulary between the burst word set and hotspot word set and calculates the ratio of identical vocabulary to the burst word set, defined as follows:

$$P = \frac{|T \cap R|}{|T|} \quad (6)$$

where P is the coverage rate. The larger the coverage rate, the higher the correspondence between burst words obtained in the observation window and hotspot words obtained in the performance window, and the better the model performance.

To ensure the applicability of the model in different time period samples, the sliding window method is used to move the standard window, observation window, and performance window backward by one unit (i.e., moving 1 year) while keeping the three window lengths unchanged. The above steps are repeated to calculate coverage rates, sequentially obtaining p_A, p_B, and p_C. The stability of the burst word detection model is judged based on coverage rates of different samples.

4 Experiments

The model concept is applied to burst word detection in library and information science literature. Literature information from 18 CSSCI core journals in the library and information field between 2007-2017 was collected from CNKI. Each data structure is as follows: {author, title, keywords, year}. According to the analysis time window division method in Figure 1 [Figure 1: see original paper], the length of n is 11, the sliding window interval is 1, and the time length of three sample matrices A_{m×i}, B_{m×(i+1)}, C_{m×(i+2)} (i+2 < n) is 9 years. The length of each sample matrix' s standard window, observation window, and performance window is 3 years. The division of analysis time windows from 2007-2017 is shown in Table 2 .

4.1 Data Collection and Preprocessing

The initial data structure {title, author, keywords, year} consists of 53,221 records with four elements. Records requiring processing are shown in Table 3 , mainly including three categories: Missing data: records without keywords, titles, authors, etc.; Non-journal papers: call for papers, announcements with words like "held," "organized," "committee," "speech," "address" in titles; Special

characters: such as ";;" and ";" . Python' s pandas and numpy tools were used for data analysis, and excel and sqlite were used for data storage.

Based on the above situation, category and errors were deleted, and category data were cleaned. At the same time, a synonym table and stop word table were established to merge keywords with similar meanings, English case variations, and Chinese-English synonyms. Words without research significance such as "Mr.," "characteristics," and "article" were included in the stop word list and removed from keywords.

4.2 Matrix Construction

According to the model, F_{m×n} was first constructed with keyword columns as the unique index, years as column names, and keyword frequencies as matrix element values. To meet the foundation of keyword quantitative changes, keywords with frequencies below 3 were eliminated, while basic disciplinary vocabulary that consistently ranked at the top of word frequency over 11 years, such as "library," "university library," and "public library," were also eliminated. On this basis, according to the Pareto principle, an F_{m×n} with dimensions 1904 × 11 was constructed. The selected keywords accounted for 24% of the total vocabulary, conforming to the Pareto principle. The cleaned F_{m×n} is shown in Table 4 (Note: 0717 total frequency represents the sum of a vocabulary' s frequency from 2007 to 2017).

4.3 Burst Word Detection

According to Section 3.2, the F_{m×n} matrix in Table 4 was cut into three sample matrices according to the divided time windows, namely A_{m×i}, B_{m×(i+1)}, C_{m×(i+2)} (i+2 < n) mentioned above. Since the time windows and calculation methods of each sample are consistent, this paper takes sample matrix A_{m×i} as an example for burst word detection and verification.

According to Section 3.3, the sum of word frequencies for every 3 years in sample matrix A_{m×i} was merged, i.e., a_{mn}. The relative word frequency, word frequency growth rate, and word frequency heat weight of AT2 were calculated according to Formulas (1), (2), and (3) respectively. Through experiments, when s = 200, the model performed well, i.e., the top 200 keywords in each indicator in the AT2 window were included in the burst word candidate set. According to Formula (4), the intersection of the three sets was calculated, yielding 13 burst words. The burst word results are shown in Table 5 .

4.4 Hotspot Word Extraction via LDA

According to Step 4 of the model, the keyword entry column in the original data {title, author, keywords, year} was used as LDA mining corpus. The keyword entry column is the D document collection in Step 4, and each keyword in the entry is a keyword in each document d_p = {x_1,⋯,x_j}. Keywords in the

AT3 window were segmented, and stop words without practical meaning in the document collection were removed as input for the LDA model. The gensim text analysis tool was used to train the text set. After multiple experiments, it was found that when q = 10, the model performed well, i.e., the number of topics was set to 10, and each topic contained the top 10 keywords with probability values. The topic vocabulary in the AT3 window is shown in Table 6 .

It can be seen from Table 6 that each topic consists of 10 keywords. Among them, Topic 1 describes the function of cloud computing and big data in providing resource sharing and knowledge services in the library and information field; Topic 2 describes the knowledge organization of public libraries and the career development of librarians and libraries; Topic 3 describes library services, including document delivery, scientific novelty search, and service quality; Topic 4 describes the construction of information resources, including e-government, institutional repositories, and open access to resources; Topic 5 describes information services, including library services, knowledge graphs, user behavior, and user needs; Topic 6 describes analysis methods in library and information science, including data mining, text mining, clustering analysis, and co-word analysis; Topic 7 describes knowledge-sharing communities, including knowledge management, knowledge sharing, virtual communities, and knowledge innovation; Topic 8 describes library reading promotion activities to improve readers' information literacy; Topic 9 describes the development of network public opinion, including the communication tool for public opinion—Weibo—and library services; Topic 10 describes disciplinary services in the library and information field, including the improvement of subject librarians' quality, improvement of service models, and academic influence.

4.5 Coverage Rate Calculation

According to Step 5 of the model, the burst word set screened in Section 4.3 and the hotspot topic words screened in Section 4.4 were used to calculate burst word coverage. According to Formula (6), P = |T R|/|T| = 12/13 = 0.92. This result indicates that in sample matrix A_{m×i}, 92% of burst words identified by the burst word detection model in the AT2 window were accurately represented in the AT3 window. Meanwhile, the burst word set = {Weibo, linked data, cloud computing, emergency events, knowledge graph, disciplinary service, network public opinion, reading promotion, research hotspots, information behavior, service system, virtual community, bibliometrics} comprehensively reflects hotspot topic words (elements contained in the intersection of set T and set R, i.e., the bold italic words in Table 6).

4.6 Sliding Window Validation

According to the analysis time window settings, the calculation steps in Sections 4.1-4.5 were repeated to obtain the coverage rates p_B and p_C for sample matrix B_{m×(i+1)} and sample matrix C_{m×(i+2)} respectively.

(1) In sample matrix B_{m×(i+1)}, the BT2 window burst word set = {reading promotion, mobile library, Weibo, social network analysis, free opening, visual analysis, scientific data, linked data, big data}, totaling 9 burst words. The burst word results for the BT2 window in sample matrix B_{m×(i+1)} are shown in Table 7 , and the topic vocabulary for the BT3 window in sample matrix B_{m×(i+1)} is shown in Table 8 .

It can be seen from Table 8 that Topic 1 describes analysis methods in the information science discipline; Topic 2 describes common forms of information resource open access and knowledge organization, including knowledge bases, ontologies, and metadata; Topic 3 describes library service content and methods, including collection services, recommendation services, and electronic resource management; Topic 4 describes the application of big data technology in modern information networks; Topic 5 describes library database management systems and library and information industry associations; Topic 6 describes information services of university libraries and readers' information literacy; Topic 7 describes librarians and disciplinary service topics; Topic 8 describes the development of network information resources, including network public opinion, Weibo, information science, emergency events, and open access to data resources; Topic 9 describes data sharing topics, including scientific data management and enterprise data management; Topic 10 describes the interdisciplinary nature of library science with other disciplines, including electronics, light industry, and other science and engineering disciplines. According to Formula (6), P = 8/9 = 0.89. This result indicates that in sample matrix B_{m×(i+1)}, 89% of burst words identified by the burst word detection model in the BT2 window were accurately represented in the BT3 window.

(2) In sample matrix C_{m×(i+2)}, the CT2 window burst word set = {cloud service, reading promotion, mobile library, Weibo, WeChat, data management, collection resources, scientific data, linked data, big data}, totaling 10 burst words. The burst word results for the CT2 window in sample matrix C_{m×(i+2)} are shown in Table 9 , and the topic vocabulary for the CT3 window in sample matrix C_{m×(i+2)} is shown in Table 10 .

It can be seen from Table 10 that Topic 1 describes information organization and information analysis; Topic 2 describes the application of cloud computing technology in the library and information field and data visualization; Topic 3 describes the improvement of university librarians' information literacy and reading promotion activities, and the emerging online learning space—MOOCs; Topic 4 describes the use of linked data, digitalization, and other technologies to build collection resources; Topic 5 describes new methods in the library and information field, including knowledge graphs and social network analysis; Topic 6 describes the data open access movement, emphasizing information intellectualization and knowledge sharing, including knowledge organization, intellectual property, and institutional repositories; Topic 7 describes the basic work of library and information science, including data governance technology, data organization technology, and data representation technology; Topic 8 is not ob-

vious in content, involving multiple topic vocabulary including patent analysis and mobile libraries; Topic 9 describes the development of network public opinion, including network emergency events and data mining technology; Topic 10 describes library reading promotion activities, digital reading, universal reading services, and data sharing and opening. According to Formula (6), P = 7/10 = 0.70. This result indicates that in sample matrix C_{m×(i+2)}, 70% of burst words identified by the burst word detection model in the CT2 window were accurately represented in the CT3 window.

4.7 Comparison with CiteSpace

To verify the performance of the new model, a control experiment was conducted using the mainstream burst word detection tool CiteSpace. The data source and burst word detection time period were the same as above. In CiteSpace software, Burstness detection was selected with parameters set as follows: words with annual frequency greater than 50 were selected as candidate burst word sets (SelectTop = 50); in the Burstness panel, the minimum burst duration was set to 1 year (MinimumDuration = 1). Burst results for different data samples were obtained according to burst intensity values. The years 2010-2012, 2011-2013, and 2012-2014 correspond to Table 11 , Table 12 , and Table 13 respectively. Observation of time change trends reveals that burst words detected by CiteSpace include two types: dying trends (e.g., information literacy, personalized service) and rising trends (e.g., bibliometrics, co-word analysis). This study argues that rising burst words are more likely to become research hotspots in the future and have more guiding significance for disciplinary research directions. Therefore, the new model focuses more on rising burst vocabulary.

According to Formula (6), burst words detected by the two methods were calculated for coverage with hotspot words respectively. The results are shown in Table 14 . Observation results show that the burst word coverage rates of the new model on three data samples are all greater than those of CiteSpace analysis results, thus indicating that the new model performs better than CiteSpace.

5 Conclusion

This study proposes a multi-measure burst word detection and verification model. Using literature information from 18 core journals in the library and information field from 2007-2017 as the data source, a 9-year analysis time window was fixed, sliding three times, with each window further divided into standard window, observation window, and performance window. Burst words in the observation window were identified based on relative word frequency, word frequency growth rate, and word frequency heat weight. Hotspot topic words in the performance window were mined through LDA, and burst word coverage rates were calculated. The results show that coverage rates in three time windows were all greater than 70%, indicating that the designed model can effectively capture burst words and discover research hotspots. In the control

experiment with CiteSpace burst word detection tools, the burst word coverage rate was better than the latter, demonstrating the value of this research.

This study still has some limitations and future research priorities: Improvement of burst word identification conditions to enhance identification accuracy; Improvement of model verification methods—the current relationship between burst words and hotspot words is one-to-many, and future research will change it to one-to-one; Use of other methods such as LDA2Vec, Word2Vec, Coder-autoencoder and other deep learning methods for multi-hotspot comparative analysis to find the best application.

References

[1] Guan Peng, Wang Yuefen. Scientific Literature Theme 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.

[2] KLEINBERG J. Bursty and hierarchical structure in streams [J]. Data mining & knowledge discovery, 2003, 7(4): 373-397.

[3] Zheng Ledan. Analysis of Research Frontiers and Evolution of Digital Library in China Based on Burst Detection [J]. Library Tribune, 2013, 33(1): 47-51.

[4] CHEN C M. CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature [J]. Journal of the Association for Information Science & Technology, 2006, 57(3): 359-377.

[5] Yang Xuanhui, Cai Zhiqiang. Detection of Emerging Trends in Linked Data Based on Mutation Detection and Co-word Analysis [J]. Information Science, 2018, 36(11): 164-168.

[6] Tang Xiaobin, Zhou Zhimin, Dong Li. Dynamic Monitoring of Network Emergencies in the Context of Big Data [J]. Statistical Research, 2017, 34(2): 46-56.

[7] Zhuo Keqiu, Yu Wei, Su Xinning. MapReduce Parallel Implementation of Emergency Event Detection [J]. New Technology of Library and Information Service, 2015(2): 46-54.

[8] Chen Guolan. Research on Microblog Emergency Monitoring Method Based on Burst Word Identification [J]. Journal of Intelligence, 2014, 33(9): 123-128.

[9] Lu Wanhui, Ma Jianxia. Research and Implementation of Domain Burst Word Recognition Based on CRFs [J]. Information Science, 2014, 32(1): 89-93.

[10] Jie Fei, Xie Fei, Li Lei, et al. Implicit Event Burst Detection in Social Networks [J]. Acta Automatica Sinica, 2018, 44(4): 730-742.

[11] XIE W, ZHU F, JIANG J, et al. TopicSketch: Real-time bursty topic detection from Twitter [J]. IEEE transactions on knowledge and data engineering, 2016, 28(8): 2216-2229.

[12] Wang Liya. Research on Topic Mutation Based on Keyword Mutation [J]. Information Studies: Theory & Application, 2013, 36(11): 45-48.

[13] Wang Zheng, Yi Li, Zhao Lei. Research on Scientific Research Hotspot Discovery Service Model Based on Burst Word Detection [J]. Information Studies: Theory & Application, 2018, 41(3): 129-135.

[14] Zhang Jinzhu, Lv Pin. Topic Evolution and Mutation Analysis Based on Improved Topic Correlation [J]. Journal of Intelligence, 2015, 34(12): 176-180.

[15] Jiang Xin, Wang Dezhuang, Ma Haiqun. Theme Evolution of "Scientific Data" Field in China from the Perspective of Keyword Frequency Change [J]. Modern Information, 2018, 38(1): 141-146, 161.

[16] SHI L, DU J P, LIANG M Y. STRM: a sparse rnn-topic model for discovering bursty topics in big data of social networks [J]. Journal of information science and engineering, 2019, 35(4): 749-767.

[17] Fu Zhu, Wang Yuefen. Research on Several Issues in the Term Collection Stage of Co-word Analysis [J]. Journal of the China Society for Scientific and Technical Information, 2016, 35(7): 704-713.

[18] Liu Minjuan, Zhang Xuefu, Yan Yun. Research on Selection Methods of Co-word Analysis Vocabulary Range Based on Word Frequency, Word Volume, and Cumulative Word Frequency Ratio [J]. Library and Information Service, 2016, 60(23): 135-142.

[19] Wikipedia. Long tail [EB/OL]. [2019-09-08]. https://en.wikipedia.org/wiki/Long_tail.

[20] Xu Jian, Huang Qiuyue. Application of "Pareto Principle" in Library Management [J]. Journal of Library Science in China, 2007(5): 106-108.

[21] BLEI D M, NG A Y, JORDAN M I. Latent dirichlet allocation [J]. Journal of machine learning research, 2003, 3(4/5): 993-1022.

[22] Wang Jian. Research on Microblog Emergency Detection Method Based on Multi-feature Fusion [D]. Beijing: Beijing Information Science & Technology University, 2018.

[23] Ma Wenjian. Chinese Patent Early Warning System Based on Burst Word Detection [D]. Beijing: Beijing University of Technology, 2016.

[24] An Lu, Du Tingyao, Li Gang, et al. Focus and Evolution Patterns of Stakeholders in Public Health Emergencies in Social Media [J]. Journal of the China Society for Scientific and Technical Information, 2018, 37(4): 394-405.

Author Contributions:

Feng Guohe: Proposed the main ideas and revision suggestions, finalized the paper;
Wu Jiajia: Data screening and processing, experiments, initial draft writing;
Mo Xingqing: Proofreading and data verification.

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

ChinaRxiv 中X

Source: ChinaXiv —Machine translation. Verify with original.