Construction of Evaluation Metrics for Hot Words in Scientific Journals and Knowledge Services

Authors: Chang Zongqiang, Liu Wei, Hou Chunmei, Ye Xiyan, Zhang Jinghui, Tao Hua, Ye Xiyan

Date: 2024-02-23T20:32:05+00:00

Abstract

【Objective】To construct evaluation indicators for journal hot terms and, through a hot term ranking service, help users quickly and intuitively understand the frontier fields and research directions of journals. 【Methods】 Utilizing the web literature survey method to analyze characteristics of entries extracted from current intelligent segmentation lexicons; employing parameterization, standardized unique marking, and other methods to extract computational parameters associated with hot terms, conduct statistical analysis, construct mathematical models, and perform multi-dimensional ranking. 【Results】 Constructed a model for extracting effective entries from journals, performed screening and standardized marking of effective entries, built a mathematical model for hot term evaluation indicators through logical calculation of entries bearing parameter information, and presented an application example of hot term indicator knowledge service in ranking format. 【Conclusion】 The standardized unique marking method can enhance the entry recognition capability of segmentation lexicons, rendering segmentation results more professional and reliable; the journal hot term ranking service can facilitate rapid and intuitive understanding of frontier fields and research directions in journal publications.

Full Text

Construction and Application of Evaluation Indicators for Hot Keywords in Journals Based on Term Banks

CHANG Zongqiang[1,2], LIU Wei[1,2], HOU Chunmei[1,2], YE Xiyan[1,2]*, ZHANG Jinghui[1,2], TAO Hua[1,2]

[1] Northwest Institute of Eco-Environment and Resources, Chinese Academy of Sciences, Lanzhou, Gansu 730000, China

[2] Key Laboratory of Knowledge Computing and Intelligent Decision, Gansu Province, Lanzhou, Gansu 730000, China

Abstract

[**Purpose**] To construct evaluation indicators for hot terms in scientific journals and, through a hot term ranking service, enable users to quickly and intuitively understand the cutting-edge fields and research directions of journals. [**Methods**] We employed web-based literature research to analyze the characteristics of entries extracted from current intelligent segmentation lexicons, utilized parameterization and standardized unique labeling methods to extract computational parameters associated with hot terms, performed statistical analysis, constructed mathematical models, and conducted multi-dimensional ranking. [**Findings**] We constructed a journal effective term extraction model to screen and standardize effective entries, developed a mathematical model for hot term evaluation indicators through logical calculation of effective entries with parameter information, and presented an application example of hot term indicator knowledge services in the form of ranking lists. [**Conclusion**] The standardized unique labeling method can enhance the entry recognition capability of segmentation lexicons, making segmentation results more professional and reliable. The journal hot term ranking service can help users quickly and intuitively understand the cutting-edge fields and research directions of journals.

Keywords: scientific journals; evaluation indicators; knowledge services; intelligent segmentation

In the era of big data, journal development relies heavily on database support, and the digital publishing model of journal-database integration has become one of the major trends in journal publishing. As digital publishing technology advances and users' efficiency in knowledge acquisition continuously improves, their demands for journal knowledge services have inevitably increased. Hot terms in journals, as textual elements reflecting journal development dynamics, can help users rapidly obtain information about current hotspots at the research frontier. Therefore, whether hot terms can be effectively extracted and applied in knowledge services is a topic worthy of investigation.

Current research on hot term extraction and analysis primarily focuses on social network information, mainly involving social hot topics [1], value analysis [2-3], and ethical research [4-5]. Studies in specific industry domains also predominantly target general web information, such as automatic hot term extraction methods based on agricultural web information classification [6], while research on hot term extraction in the field of scientific journals remains relatively scarce. It is worth noting that although various indicators exist for journal evaluation, such as impact factor, total citation frequency, source literature volume, and funded paper ratio [7], evaluation indicators for journal hot terms are rarely discussed. Furthermore, regarding digital service technologies for journal hot terms, Ouyang et al. [8] adopted a hybrid decision model-based compound con-

cept extraction method to achieve multi-compound concept extraction through text segmentation, term denoising, and synonym merging, thereby improving the accuracy and efficiency of compound term extraction. Fu et al. [9] proposed a lexicon-based segmentation algorithm incorporating word frequency, part of speech, Chinese grammar rules, and unknown word recognition rules, which can largely eliminate ambiguous segmentation and improve unknown word recognition probability. Chinese patents [10] have reported an intelligent template model establishment method for publishing, enabling one-time data processing for multiple publications across different typesetting software based on template technology. Another Chinese patent [11] reported a semantic enhancement description system and method for digital publishing resources, which can identify basic copyright points and semantic expression points in digital publishing resources. However, these methods remain at the level of publishing function implementation and intelligent segmentation/text recognition, without considering knowledge service functions from the user's perspective.

Knowledge services are broadly defined as behaviors that, based on knowledge resources or products and according to user needs and usage scenarios, integrate into users' problem-solving processes to effectively support knowledge application and innovation [12,13]. This represents secondary processing and multiple derivations of existing knowledge. Currently, knowledge services constitute the main direction for the transformation and development of scientific journals [14]. Therefore, further exploring the knowledge service functions based on journal hot term evaluation indicators can better leverage the knowledge service capabilities of scientific journals and expand and enhance journal brand influence.

Based on the above analysis, this paper focuses on constructing evaluation indicators for journal hot terms. According to the extraction characteristics of effective terms in journals, we extract information variables affecting term heat, perform parameterization and standardized labeling, design underlying computational logic to build mathematical models, and propose a multi-dimensional indicator for measuring journal hot terms. This provides a new approach for evaluating journal hot terms, further explores their knowledge service functions, and helps users quickly and intuitively understand the cutting-edge fields and research directions of journals. Additionally, the application of hot term indicators can provide editorial staff with a reference path for enhancing journal knowledge services and offer publishing industry technicians a model reference for improving the professionalism and reliability of hot terms.

1 Research Methods

We primarily employed web-based literature research to analyze the characteristics of entries extracted from current intelligent segmentation lexicons, such as term validity, professionalism, and completeness. Based on existing problems in term extraction, we extracted effective terms through comparison and matching with a regularly maintained and updated professional basic term bank. According to investigations of heat parameters, we utilized parameterization and stan-

dardized unique labeling methods to extract computational parameters related to hot term heat, and performed unified standardized processing and statistical analysis on the data. On this basis, by analyzing the correlation between various parameters and journal hot term heat, we constructed a mathematical model for journal hot term evaluation indicators using indicator construction methods. Finally, according to the characteristics of each parameter and indicator, we performed value-based ranking of journal hot terms from different dimensions and presented an application example of hot term indicator knowledge services in the form of ranking lists.

2.1 Journal Effective Term Extraction Model

Currently, intelligent segmentation technology has matured considerably, with increasingly better segmentation effects. However, specialized segmentation functions for disciplinary terminology and neologisms still need further improvement. As this paper focuses on specialization and academic orientation for scientific journals, it is necessary to screen the terms extracted by intelligent segmentation technology and extract effective terms. To meet this requirement, we constructed a journal hot term extraction model comprising a basic professional term bank, term screening channel, middle-layer intelligent segmentation lexicon, effective term output channel, and top-layer extracted term bank, as illustrated in Figure 1. The term screening channel connects the basic professional term bank and the middle-layer intelligent segmentation lexicon, while the effective term output channel connects the middle-layer intelligent segmentation lexicon and the top-layer extracted term bank.

In this model, the basic professional term bank refers to a professional lexicon built into the journal platform that is highly relevant to the journal's scope. Journals can update this lexicon through modification, deletion, and addition operations, making it the platform's foundational term bank. Notably, journals need to actively and timely maintain this basic term bank by adding new terms extracted from recent articles, new concepts and terminology emerging in academia, etc., to ensure the timeliness of the basic term bank. The middle-layer intelligent segmentation lexicon refers to the temporary storage of semantic segmentation results obtained by applying existing intelligent semantic segmentation technology to journal papers. The top-layer extracted term bank stores effective terms with standardized unique labels.

2.2 Screening and Standardized Labeling of Effective Terms

Screening effective terms aims to select terms that meet journal development needs from numerous candidates. This is primarily achieved through comparison and matching: first, semantic segmentation results temporarily stored in the middle-layer intelligent segmentation lexicon are compared with entries in the basic professional term bank; then different labeling operations are performed based on different comparison results. If a compared term A exists in the basic

professional term bank, it receives a standardized unique label; if not, it remains unlabeled, thereby cleaning invalid terms.

After screening effective terms, the middle-layer intelligent segmentation lexicon performs standardized unique labeling on them, and these labeled effective terms are output through the effective term output channel to the top-layer extracted term bank. The standardized unique labeling method used by the middle-layer intelligent segmentation lexicon employs multiple parameters with quantitative characteristics to label the screened effective terms. For example:

Taking term A as an example, let Y denote the publication year of the journal article from which it was extracted, and F denote its frequency in that article. If term A is being compared for the first time with the basic professional term bank, it is labeled by year as YAF1, by frequency as F1A1, and the article count is recorded as 1(YA). For the second comparison, the labels become YA F2, F2A2, and 2(YA), respectively. By extension, for the nth comparison, its year, frequency, and article count are labeled as YA Fn, FnAn, and n(YA), respectively.

2.3 Logical Calculation of Effective Term Parameter Information

As described above, the parameters used to label effective term A include year, frequency, and article count, with the single comparison labeling method for a given publication year already provided. However, term A may be extracted multiple times in the same publication year or across different years, requiring separate labeling. If the top-layer extracted term bank contains multiple year-labeled entries for term A in the same publication year Y, i.e., YA F1, YA F2, YA F3, …, YAFn, then the article count for term A in year Y is n(YA), and the total frequency is F(Y) = F1 + F2 + F3 + … + Fn. Term A with year, frequency, and article count information is then labeled as Y F(Y)An, as shown in Figure 2. For instance, if term A appears in 8 papers in 2020, with frequencies in these 8 papers recorded as F1, F2, F3, …, F8, then the article count for term A in 2020 is 8, and the total frequency is F(Y) = F1 + F2 + F3 + … + F8, recorded as F(2020)A8.

The total article count for term A across all years is recorded as NA, and the total frequency as FA, where NA = n1 + n2 + n3 + … + ni (with n1, n2, n3, …, ni representing article counts for term A in different years) and FA = F(Y1) + F(Y2) + F(Y3) + … + F(Yi) (with F(Y1), F(Y2), F(Y3), …, F(Yi) representing total frequencies of term A in different years).

2.4 Mathematical Model for Term Heat Evaluation Indicators

(1) Weight Score Evaluation Indicator

The weight score evaluation indicator SA primarily performs weighted assignment calculations based on the total article count NA and total frequency FA of term A across all years. Let the weight combination be (x, y), such that:

$$SA = xNA + yFA$$

The weight combination (x, y) can be flexibly adjusted according to disciplinary characteristics. In this paper, we temporarily adopt the Pareto principle with x = 0.8 and y = 0.2, yielding:

$$SA = 0.8NA + 0.2FA$$

where NA = n1 + n2 + n3 + … + ni (with n1, n2, n3, …, ni representing article counts for term A in different years) and FA = F(Y1) + F(Y2) + F(Y3) + … + F(Yi) (with F(Y1), F(Y2), F(Y3), …, F(Yi) representing total frequencies of term A in different years).

(2) Heat Evaluation Indicator

Considering the cooling of hot terms over time, we further construct a heat evaluation indicator by incorporating a cooling coefficient R on the basis of the weight score. The heat of a hot term is primarily related to the weight score SA and the cooling coefficient R. Based on this relationship, the initial mathematical model for heat is constructed as:

$$HA = \frac{\alpha SA}{R}$$

where α is a debugging parameter for heat applicability, empirically calculated as $\alpha = 8.76$ in this paper. The cooling coefficient R is determined by the following logic: the heat of hot terms decays over time. We define the time elapsed as ΔT = i*b - Yi, where b is the current year, Yi is the sum of all years in which the term appears (i.e., Yi = Y1 + Y2 + Y3 + … + Yn), and i = 1, 2, 3, …, n. Then:

$$R = (i * b - Yi + C)^{\beta}$$

where C is a constant ensuring the heat HA remains valid when ΔT = 0 (C = 2 in this paper), and parameter β represents the decay rate of hot terms over time, adjustable according to specific circumstances (β = 1.2 in this paper). The final constructed heat evaluation indicator mathematical model is:

$$HA = \frac{8.76SA}{(i * b - Yi + 2)^{1.2}}$$

This paper primarily uses the year as the time scale to provide a reference approach for constructing journal hot term indicators. Although effective terms

can be updated in real-time dynamically, calculating by annual scale may not present neologisms in a timely manner. Therefore, a journal neologism module can be added in addition to hot terms. As mentioned earlier, the underlying basic professional term bank requires timely maintenance and updates. Moreover, the time scale for hot terms can also be measured by month or quarter; if statistical significance is lacking, newly added terms can be presented in the journal neologism module.

3 Knowledge Services Based on Journal Hot Term Evaluation Indicators

After obtaining hot term parameter information and evaluation indicators, we further explore their application value primarily through ranking methods, presenting application examples of hot term indicator knowledge services in the form of ranking lists. The ranking methods refer to any of the following: sorting by year, frequency, and article count; sorting by weight score; or sorting by term heat. These three sorting methods coexist in the top-layer extracted term bank, and while the content items for ranking can be presented simultaneously, only one sorting method can be selected at a time. For example, if term A is sorted by frequency, it cannot be simultaneously sorted by weight score or heat, though the values of weight score or heat can be displayed concurrently.

3.1 Basic Parameter Ranking of Effective Terms

Effective terms can be sorted by year, frequency, and article count, with the year range (a, b) freely selectable, where b is the current year and a is a year prior to b ($a \leq b$). This ranking allows querying the frequency F(Y) and article count n of different terms in the same year (Table 1) or within a certain year range (Table 2). On the basis of a selected year range, further sorting by term frequency or article count can be performed.

Alternatively, without setting a year range, sorting can be directly performed based on the total frequency F and total article count N of all terms in the top-layer extracted term bank, enabling queries of TOP10, TOP20, TOP50 hot term rankings from the dimensions of journal frequency and article count, as exemplified in Table 3.

Table 1 Example of sorting by annual frequency and article count for different terms in the same year
Table 2 Example of sorting by annual frequency and article count for the same term across different years
Table 3 Example of sorting by total frequency and total article count for different terms

3.2 Hot Term Weight Score and Heat Ranking

This primarily displays the weight score S and heat H of different terms, which can be sorted by S and H respectively, enabling queries of TOP10, TOP20, TOP50 hot term rankings from the dimensions of weight score and heat, as exemplified in Table 4.

Table 4 Example of weight score and heat ranking for different terms

In summary, this paper presents four forms of ranking lists, displaying the status of hot terms under specific dimensions from multiple perspectives. These dimensions can complement each other as references to ensure objectivity and effectiveness. Among them, the heat indicator comprehensively considers frequency, article count, weight, and time decay, serving as an important metric reflecting term heat. However, it requires further correction due to limited empirical analysis. Additionally, besides ranking lists, hot terms can be presented graphically based on the underlying ranking data.

4 Advantage Analysis

(1) This paper performs relevant basic calculations on labeled terms in the top-layer extracted term bank to obtain the frequency and article count of professional terms in journal papers. Based on total term frequency and total article count, weight scores are assigned to terms according to preset formulas. Term heat is then calculated based on weight score and year, and through ranking lists of different dimensions (frequency, article count, weight score, and heat), users can quickly and intuitively understand journal research dynamics.

(2) By constructing a journal hot term extraction model that incorporates a basic professional term bank and a middle-layer intelligent segmentation lexicon, semantic segmentation results temporarily stored in the middle-layer lexicon are compared with entries in the basic professional term bank. Different labeling operations are performed based on different comparison results to clean invalid terms and screen professional terms related to the journal's scope, thereby improving the professionalism and effectiveness of middle-layer intelligent segmentation.

(3) This paper performs quantitative labeling on terms extracted from the middle-layer intelligent segmentation lexicon, using multiple parameters with quantitative characteristics to label screened effective terms. This endows screened effective terms with unique identifier features, enhancing the term recognition capability of the middle-layer intelligent segmentation lexicon and making segmentation results more reliable.

(4) Journal hot term ranking lists primarily include term rankings by frequency, article count, weight score, and heat, presenting the changing dynamics of journal hot terms from multiple dimensions to help users quickly

grasp journal research directions, thereby enhancing journals' knowledge service capabilities for users.

5 Conclusion

This paper primarily employs a parameterization scheme to extract statistically significant parameter variables based on an embedded professional term bank. After correlational analysis of these parameter variables, underlying computational logic is designed, basic mathematical models are constructed, and constant coefficients are debugged for model optimization to achieve the desired effect. The paper further develops the transformation and application of hot term evaluation indicators, launching a hot term ranking knowledge service product. Reviewing the construction process of the hot term heat evaluation indicator, the main difficulties lie in determining the parameterization scheme and designing the underlying logic. Due to the lack of empirical validation with specific cases and potential disciplinary differences, the next step involves conducting empirical research on the hot term evaluation indicator and performing comparative analysis across different disciplines to further adjust parameters, make corrections, and improve indicator reliability, as well as investigate its universality. On this basis, we will further analyze the correlation between hot terms in the temporal dimension, deeply explore the shifting trends in research heat among correlated hot terms, and effectively enhance journals' knowledge service capabilities.

References

[1] HUANG Ju. Hot word extraction and sentiment analysis for Weibo hot topics[D]. Hefei: Anhui University of Science and Technology, 2022.
[2] QIU Hongxia, WANG Xisheng. Value analysis of "network hot words" in recent years[J]. Journal of Mudanjiang Normal University (Philosophy and Social Sciences Edition), 2017(05): 5-10.
[3] FANG Ting. Network contradictory hot words and audience social identity[D]. Hefei: Anhui University, 2019.
[4] ZHOU Siyuan. Ethical research on Chinese network hot words[D]. Shijiazhuang: Hebei University of Economics and Business, 2021.
[5] HU Qingqing. Ethical research on network hot words[D]. Changsha: Hunan Normal University, 2015.
[6] DUAN Qingling, ZHANG Lu, LIU Yiran, et al. Automatic hot word extraction method based on agricultural network information classification[J]. Transactions of the Chinese Society for Agricultural Machinery, 2018, 49(7): 160-167.
[7] SUN Xiaohong, YAN Lijuan, ZHANG Gaixia, et al. Analysis of the changing trends of major journal evaluation indicators of Acta Geoscientica Sinica from 2005 to 2018[J]. Acta Geoscientica Sinica, 2021, 42(1): 124-128.
[8] OUYANG Liubo, ZOU Beiji, LIU Lijie. A compound concept extraction method based on hybrid decision model[J]. Acta Electronica Sinica, 2013, 41(3): 488-495.

[9] FU Shiguang, LIN Youfang, WAN Huaiyu, et al. A rule-based Chinese word segmentation algorithm[C]//Chinese Information Processing Society of China, Chinese and Oriental Language Information Processing Society of Singapore, Language and Information Research Center of Wuhan University. Research on Chinese Computational Technology and Language Problems—Proceedings of the 7th International Conference on Chinese Information Processing. Publishing House of Electronics Industry, 2007: 52-56.
[10] WANG Rong, LI Pingli, GONG Jian. A method for establishing an intelligent template model for publishing[P]. Beijing: CN100392654C, 2008-06-04.
[11] CHEN Lin, XIE Bing, LU Peng, et al. A digital publishing resource semantic enhancement description system and method[P]. Beijing: CN102999487B, 2015-06-24.
[12] Press and Publication Knowledge Services—Basic Terminology for Knowledge Resource Construction and Services: GB/T 38377—2019[S]. Beijing: Standards Press of China, 2019.
[13] Press and Publication Knowledge Services—Guidelines for Knowledge Resource Construction and Services: GB/T 38382—2019[S]. Beijing: Standards Press of China, 2019.
[14] GUO Yumei, JING Yong, GUO Xiaoliang, et al. Analysis of the operation model of scientific journal knowledge service platforms under open science[J]. Acta Editologica, 2023, 35(3): 273-278.

Author Contributions:
CHANG Zongqiang and YE Xiyan: Proposed the research direction, conducted literature review, and wrote the paper;
LIU Wei and HOU Chunmei: Guided the writing approach and reviewed/revised the paper;
ZHANG Jinghui and TAO Hua: Revised the paper.

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