A New Method for Domain Hotspot and Trend Analysis Based on Weighted Keywords (Post-print)

Authors: Feng Guohe, Kong Yongxin, Xiao Jieqiong

Date: 2023-08-27T00:00:00+00:00

Abstract

[Purpose/Significance] To overcome the limitations of absolute keyword frequency analysis, and to explore domain hotspots and trends through multi-factor weighting of keywords and score-based ranking.

[Method/Process] Construct a year-keyword frequency matrix, process keyword frequencies using horizontal weighting and vertical weighting, design a relative term frequency model, and calculate weighted composite scores for keywords to obtain more effective keyword ranking.

[Results/Conclusion] Based on weighted keyword ranking, it is possible to identify high-frequency-high-quality, low-frequency-high-quality, and burst-type keywords, thereby facilitating the mining of research hotspots and trend analysis.

Full Text

A New Model for Domain Hotspot and Trend Analysis Based on Weighted Keywords

Feng Guohe[1], **Kong Yongxin**[2], **Xiao Jieqiong**[1]
[1] Department of Information Management, School of Economics and Management, South China Normal University, Guangzhou 510006
[2] Department of Information Resources Management, Business School, Nankai University, Tianjin 300071

Abstract

[Purpose/Significance] To overcome the limitations of absolute keyword frequency analysis, this study proposes a method that employs multi-factor weighting and scoring of keywords to explore domain hotspots and trends.

[**Method/Process**] The method constructs an annual-keyword frequency matrix, applies horizontal and vertical weighting to process keyword frequencies, designs a relative frequency model, and calculates weighted comprehensive scores to obtain more effective keyword rankings. [**Result/Conclusion**] Based on weighted keyword ranking, three types of keywords can be identified: high-quantity-high-quality, low-quantity-high-quality, and burst types, which facilitates hotspot identification and trend analysis.

Keywords: keyword frequency analysis; weighted keywords; hotspot research; trend analysis
Classification Number: G250
DOI: 10.13266/j.issn.0252-3116.2018.18.011

Keywords are natural language vocabulary that express the thematic concepts of literature. The collection of keywords from a large volume of academic research in a field over an extended period can reveal the content characteristics of research outcomes and illuminate the developmental trajectory and direction of academic inquiry. Therefore, statistical analysis of keyword frequencies in specific academic literature can identify research hotspots and analyze development trends in that academic domain. Keyword frequency analysis is based on statistical data, offering objectivity and accuracy that, to a certain extent, eliminates the subjectivity inherent in qualitative methods and enhances credibility. Consequently, it has been widely applied to reveal research hotspots and development dynamics across various disciplines.

As keyword frequency analysis has become increasingly prevalent across disciplines, the volume of related literature has grown substantially, yet serious methodological abuse and templating have emerged. Some studies are limited to simple statistical counting and crude analysis of word frequencies, failing to reveal the inherent patterns of knowledge within disciplinary fields through their data results. While keyword frequency analysis offers broad applicability, its application suffers from certain drawbacks that necessitate methodological refinement. Although hotspot analysis articles commonly employ various bibliometric methods such as keyword frequency analysis, citation analysis, and literature growth rate analysis, studies using keyword frequency analysis account for 61% of all hotspot analysis literature employing bibliometric methods, making it the most frequently used approach. Moreover, among literature applying keyword frequency analysis, studies using keywords as the statistical element occupy an absolute proportion, attributable to the direct availability of keywords and the absence of segmentation requirements.

Most research outcomes use the natural frequency of keywords as their foundation and basis. Considering the non-standardization of keywords, some studies have improved measurement methods from three perspectives: keyword frequency calculation, keyword selection, and results analysis. Regarding improvements to keyword frequency calculation, Ni Lijuan employed absolute word

frequency values to describe research status and reveal hotspot trends. To eliminate errors caused by fluctuations in annual publication counts, Qiu Junping proposed using average keyword frequency—dividing a keyword's annual frequency by the total number of publications that year—to determine its growth or decline. Gong Yongqiang explored trends based on keyword frequency proportions, i.e., the ratio of a particular keyword to the total keywords in a given year. For processing different sample datasets, Cang Hongyu proposed Z-Score standardization of keyword frequencies to mitigate effects from large disparities between domestic and international literature counts. Regarding keyword selection improvements, G. Chen et al. combined keyword popularity indices with domain relevance metrics for keyword selection. Concerning low-frequency term weighting, E. S. Atlam et al. proposed negative weight functions and negatively weighted inverse term frequency functions to improve keyword recall and precision. G. Chen et al. compared traditional Term Frequency (TF), TF-IDF, and TF-Keyword Activity Index (TF-KAI) methods, concluding that TF-KAI outperformed the others in both quality and quantity of keyword selection. For results analysis improvements, Li Shanshan et al. categorized keywords into low, medium, and high-frequency zones using quantitative methods to qualitatively analyze inherent literature patterns and research hotspots. While these methods suit different application scenarios, existing literature addressing absolute word frequency focuses on the impact of annual total word frequency and total publication counts, without considering self-proportion. Self-proportion can mitigate numerical frequency advantages while reflecting self-change rates. This study combines annual total word frequency and total keyword frequency to propose a weighted keyword model for exploring new research approaches.

1 Related Research

Keyword frequency analysis is a bibliometric method that identifies research hotspots and development directions based on the frequency of keywords or descriptors expressing core content in literature from a specific research field. Although hotspot analysis articles commonly employ various bibliometric methods, those using keyword frequency analysis account for 61% of all such literature, making it the most prevalent approach. Simultaneously, among literature applying keyword frequency analysis, studies using keywords as the statistical element dominate, due to keywords' direct availability and lack of segmentation requirements.

Most research outcomes use natural keyword frequency as their foundation. Considering keyword non-standardization, some studies have improved measurement methods in three aspects: keyword frequency calculation, keyword selection, and results analysis. For keyword frequency calculation improvements regarding absolute frequency-based statistical analysis, Ni Lijuan used absolute word frequency values to describe research status and reveal hotspot trends. To eliminate errors from annual publication count fluctuations, Qiu Junping utilized average keyword frequency—dividing each year's keyword frequency by

that year's total literature count—to judge growth or decline. Gong Yongqiang explored trends based on keyword frequency proportions, i.e., a keyword's ratio to the total annual keyword count. For different sample data processing, Cang Hongyu proposed Z-Score standardization of keyword frequencies to eliminate effects from large domestic/international literature count disparities. For keyword selection improvements, G. Chen et al. combined keyword popularity indices with domain relevance metrics. For low-frequency term weighting, E. S. Atlam et al. proposed negative weight functions and negatively weighted inverse term frequency functions to improve recall and precision. G. Chen et al. compared TF, TF-IDF, and TF-KAI methods, finding TF-KAI superior in keyword selection quality and quantity. For results analysis improvements, Li Shanshan et al. divided keywords into low, medium, and high-frequency zones to qualitatively analyze literature patterns and hotspots using quantitative methods. While applicable to different scenarios, current literature addressing absolute word frequency focuses on annual total word frequency and total publication counts' impact, without considering self-proportion. Self-proportion can mitigate numerical frequency advantages while reflecting self-change rates. This study combines annual total word frequency and keyword total frequency to propose a weighted keyword model for more accurately and objectively revealing disciplinary hotspots and trends, validated using China's library and information science research field.

2 Weighted Keyword Relative Frequency Model

Keyword annual distribution reflects yearly research priorities, while keyword growth over time reflects annual research hotspots. This study organically integrates keyword annual distribution with yearly frequency proportions. First, an annual-keyword frequency matrix is constructed. Based on horizontal and vertical dimension weighting of this matrix, a relative frequency calculation formula is derived to accurately reflect keyword annual distribution. Then, comprehensive weighted keyword ranking scores are determined to obtain more effective keyword rankings. This method is termed the Weighted Relative Keyword Frequency Model (WRKFM).

2.1 Relative Frequency Calculation

Construct an annual-keyword frequency matrix where function $f(i, j)$ represents the frequency of keyword i in year j. All keywords across all years can be represented by Matrix (1):

$$\begin{matrix} f(1,1) & \cdots & f(1,m) \\ \vdots & \ddots & \vdots \\ f(n,1) & \cdots & f(n,m) \end{matrix}$$

Matrix (1)

To reflect frequency intensity across different keywords in the same year and across different years for the same keyword, keyword frequencies undergo two-dimensional weighting:

1. **Vertical weighting**: Dividing a keyword's annual frequency by that year's total keyword frequency. Let n_j be the total keyword count in year j. Applied to Matrix (1), each element in column j is multiplied by $\frac{1}{n_j}$, represented by Matrix (2):

$$\begin{matrix} f(1,1) \times \frac{f(1,1)}{n_1} & \cdots & f(1,m) \times \frac{f(1,m)}{n_m} \\ \vdots & \ddots & \vdots \\ f(n,1) \times \frac{f(n,1)}{n_1} & \cdots & f(n,m) \times \frac{f(n,m)}{n_m} \end{matrix}$$

Matrix (2)

2. **Horizontal weighting**: Calculating a keyword's annual frequency proportion within its total statistical period frequency. Let m_i be the total frequency of keyword i. Applied to Matrix (1), each element in row i is multiplied by $\frac{f(i,j)}{m_i}$, represented by Matrix (3):

$$\begin{matrix} f(1,1) \times \frac{f(1,1)}{m_1} & \cdots & f(1,m) \times \frac{f(1,m)}{m_1} \\ \vdots & \ddots & \vdots \\ f(n,1) \times \frac{f(n,1)}{m_n} & \cdots & f(n,m) \times \frac{f(n,m)}{m_n} \end{matrix}$$

Matrix (3)

2.2 Weighted Keyword Relative Frequency Model Design

Based on Matrix (2) and Matrix (3), the Weighted Keyword Relative Frequency Model is represented by Matrix (4):

$$\begin{matrix} f(1,1) \times \frac{f(1,1)}{n_1} \times \frac{f(1,1)}{m_1} & \cdots & f(1,m) \times \frac{f(1,m)}{n_m} \times \frac{f(1,m)}{m_1} \\ \vdots & \ddots & \vdots \\ f(n,1) \times \frac{f(n,1)}{n_1} \times \frac{f(n,1)}{m_n} & \cdots & f(n,m) \times \frac{f(n,m)}{n_m} \times \frac{f(n,m)}{m_n} \end{matrix}$$

Matrix (4)

To derive more scientific, objective, and accurate data results and effectively transform them into knowledge conclusions, the WRKFM calculation procedure is designed as follows:

Step 1: Determine the time domain, count keywords and their frequencies, construct the annual-keyword frequency matrix, and calculate Matrix (4) results.
Step 2: Calculate the sum of each row's elements in Matrix (4), i.e., the relative word frequency W for n keywords, and rank them in descending order to

identify high-frequency keywords.
Step 3: Observe each row's values in Matrix (4) individually and plot their trends, representing the relative word frequency change trends for predicting development directions.
Step 4: Analyze the difference between the ranking from Step 2 and the original absolute word frequency ranking to facilitate monitoring of burst-type keywords.

The model's main characteristics are: (1) Relative word frequency incorporates the influence of annual word frequency proportions, overcoming the limitation of improving frequency solely from sample capacity. Thus, if a keyword's total absolute frequency is high and its overall change is substantial within the time domain, its relative frequency will be larger; (2) Relative word frequency highlights data where a keyword's proportion is large and absolute frequency is high in a given year, while weakening data with low absolute frequency, making it easier to detect keywords with development potential; (3) The ranking change magnitude of low-frequency keywords' relative word cumulative frequency correlates with burst theme types, enabling side-detection of burst terms and supplementing the utilization value of low-frequency words.

3 Empirical Analysis

Using the above model, comparative analysis was conducted on literature in the library and information science field. Document information from 18 core library and information science journals published between 2012-2016 was downloaded from CNKI and CSSCI. After manually removing non-academic journal literature such as communications and call-for-papers, 24,618 documents were obtained. Using Excel for statistics, 34,553 keywords were identified. After manually merging 82 groups of synonyms and removing 120 meaningless keywords, the study selected keywords with absolute frequency ≥ 5 (totaling 2,164 keywords).

3.1 Keyword Weighting Calculation

Following Matrix (4), keywords were weighted and calculated. Table 1 lists the top 50 keywords by relative word frequency value from 2012-2016.

Table 1 Partial Results of Absolute and Weighted Relative Word Frequencies

[The table shows comparative data for keywords including 高校图书馆, 公共图书馆, 数字图书馆, 社会网络分析, 图书馆服务, 移动图书馆, 图书馆联盟, 大学图书馆, 机构知识库, etc., with columns for each year (2012-2016) showing both absolute frequencies and calculated relative frequencies, plus cumulative totals.]

Based on Table 1, after manually filtering peripheral and non-core keywords, prominent relative frequency rankings include both traditional research directions and research hotspots.

Traditional research includes: "university library," "public library," "digital library," "information literacy," "information retrieval," and "knowledge management."

Research hotspots encompass: (1) **Library services**, including "information service," "reading promotion," "knowledge service," "subject service," "subject librarian," "knowledge sharing," "mobile library," and "library consortium"; (2) **Informetrics tools and applications**, including "cloud computing," "social network analysis," "competitive intelligence," "knowledge mapping," "bibliometrics," "visualization," and "data mining"; (3) **Information resources**, including "big data," "online public opinion," "linked data," "institutional repository," "social network," "e-government," and "emergency events."

3.2 Weighted Keyword Relative Frequency Change Trends

Compared with absolute word frequency trends, relative word frequency trends offer superior visual effects, being more sensitive to data with large proportions and high absolute frequencies in specific years while weakening years with smaller absolute frequencies. This facilitates capturing keywords with high change rates and detecting those with development potential. Figure 1 [Figure 1: see original paper] shows that "big data" rose sharply in 2016, with an absolute frequency of 174—approximately 40% of the five-year total frequency—making it a high-quantity-high-quality keyword. This demonstrates the model's effective capture of such keywords.

"Reading promotion" exhibited low frequency but significant growth rate during 2012-2014, followed by high frequency but slower growth during 2014-2016. Figure 2 [Figure 2: see original paper] shows a 平缓 front end and steep rear end, illustrating the model's macro-level grasp of keywords.

Combined analysis of Figures 1 and 2 reveals: (1) **Growth-type keywords** include "big data," "reading promotion," and "online public opinion." "Big data" shows rapid growth (from 9 occurrences in 2012 to 174 in 2016), while "online public opinion" demonstrates slow growth, both aligning with current research environments and societal demands; (2) **Fluctuation-type keywords** include "public library," "university library," and "knowledge management." For instance, "university library" peaked in 2013, troughed in 2015, and remained roughly stable in 2012 and 2016; (3) **Decline-type keywords** include "competitive intelligence," "knowledge service," "subject service," and "social network analysis." "Competitive intelligence," for example, dropped from 98 occurrences in 2012 to 39 in 2016, indicating a clear decline—hot before the study period but losing momentum afterward. Growth-type keywords are more likely to become future research trends. However, fluctuation-type and decline-type keywords constitute a large proportion, requiring expanded keyword scope to identify more growth-type keywords for better trend prediction.

3.3 WRKFM vs. Keyword Simulation Results

3.3.1 High-Frequency Keyword Experimental Results For the top 100 keywords by relative frequency ranking, comparing relative and absolute frequency rankings reveals the top 10 keywords with the largest absolute ranking differences in Table 2 . Negative values indicate lower relative frequency ranking than absolute frequency ranking; positive values indicate the opposite.

Table 2 High-Frequency Keyword Ranking Comparison

[Table shows keywords like 国家图书馆, 图书情报学, 公共文化服务 with ranking differences]

From Table 2, keyword pairs with similar cumulative absolute frequencies were selected for analysis (Table 3):

1. **"Database" vs. "Intelligence Analysis"**: "Database" showed prominent values in 2012 but clearly decreased in subsequent years, while "intelligence analysis" maintained consistently high levels despite smaller peak values. Although sharing identical absolute frequencies, "intelligence analysis" exhibited significantly higher relative frequency (Figure 3 [Figure 3: see original paper]).
2. **"Information Needs" vs. "Social Network"**: Despite similar totals and distributions, "information needs" showed larger differences between extreme and average values, affecting cumulative relative frequency totals (Figure 4 [Figure 4: see original paper]).
3. **"Reader Service" vs. "WeChat"**: Although "WeChat" had zero occurrences in 2012, its substantial growth trend highlighted its development potential, making its cumulative relative frequency more prominent than "reader service" (Figure 5 [Figure 5: see original paper]).
4. **"Mobile Service" vs. "User Needs"**: Both are fluctuation-type keywords, but the former's higher peak values highlight its overall advantage and larger cumulative relative frequency (Figure 6 [Figure 6: see original paper]).
5. **"Information Resources" vs. "Information Behavior"**: The former's peak advantage is evident; despite lower 2015-2016 values, its ranking still rose (Figure 7 [Figure 7: see original paper]).

For high-frequency keywords, those rising in weighted relative frequency ranking represent "high-quantity-high-quality" and "medium-quantity-high-quality" types—keywords with large growth trends, peak advantages, high frequency, and stability. Following Logistic growth patterns, concepts with rapidly increasing frequencies are emerging, while slowing growth indicates maturity. Thus, WRKFM can quickly and objectively identify high-quality keywords with development potential, revealing disciplinary hotspots and predicting trends.

3.3.2 Low-Frequency Keyword Experimental Results Theme burst refers to thematic changes where attention shifts dramatically within a short

period due to specific events. Over time, burst themes may become research hotspots or fade into ordinary themes. Monitoring burst terms is therefore significant. Based on temporal frequency changes, burst themes are categorized as rising, declining, rise-then-fall, emergent, or stable types. Low-frequency keyword ranking changes closely correlate with keyword bursts.

Partial keywords were selected from each ranking change magnitude stage to compare low-frequency keyword ranking changes (Table 4 and Figure 8 [Figure 8: see original paper]). Negative values indicate lower relative frequency ranking than absolute frequency ranking; positive values indicate the opposite.

Table 4 Low-Frequency Keyword Absolute vs. Relative Frequency Differences

[Table shows keywords like 数据素养教育, 移动社交网络, 文本相似度, libraries, 儿童图书馆, 政务信息资源, 潜在语义分析, 网络计量学 with minimal absolute frequencies but notable relative frequencies]

Figure 8 Cumulative Absolute Frequency (left) and Cumulative Relative Frequency (right) Line Charts for Partial Low-Frequency Keywords, 2012-2016

Results show that keywords with prominent ranking changes mostly exhibit sharp rises or drops. Since low-frequency keywords are more sensitive to frequency changes than high-frequency ones, sharp fluctuations cause significant ranking changes. Thus, statistical analysis of low-frequency keyword ranking change magnitude can side-detect burst terms and characterize relationships between ranking changes and burst theme types.

Keywords with significant ranking increases primarily represent stable theme bursts, such as "latent semantic analysis" and "webometrics," indicating stable or suddenly increased frequencies. This suggests these studies are stabilizing and will likely remain in stable development, or have integrated other disciplinary content to form new research themes.

Keywords with significant ranking decreases primarily represent emergent theme bursts, such as "data literacy education," "mobile social networks," and "cultural poverty alleviation." Sharp frequency fluctuations indicate high burst potential, suggesting these studies have social timeliness and gradually increasing research heat.

Conclusion

Addressing the common practice of simple statistical counting and crude analysis of word frequencies to reveal disciplinary hotspots and trends, this study proposes the Weighted Relative Keyword Frequency Model (WRKFM). By constructing an annual-keyword frequency matrix and applying horizontal and vertical weighting, the model derives a relative frequency calculation formula to obtain weighted comprehensive scores for more effective keyword ranking. This approach more accurately, objectively, and scientifically reveals disciplinary

hotspots and trends while exposing inherent knowledge patterns. Empirical evidence demonstrates that WRKFM offers the following advantages over absolute frequency analysis: (1) High-quantity-high-quality and medium-quantity-high-quality keywords rank higher, revealing disciplinary hotspots with development potential; (2) Low-frequency keywords with sharp rises/drops show more significant ranking changes, enabling side-detection of burst terms to predict research hotspots and trends, thereby facilitating burst keyword discovery.

This study achieves its intended design goals, but requires further research and refinement: (1) The experiment only used author-provided keywords; future data collection could expand to title, abstract, and full-text keywords to improve accuracy and comprehensiveness; (2) WRKFM ranking scores are affected by temporal frequency peaks—keywords with sudden frequency spikes in specific years receive higher scores, disadvantaging stable keywords; (3) The model only calculates and analyzes annual-keyword distributions and proportions without semantic weighting or achieving "core keywords floating up, auxiliary keywords sinking down," requiring additional methods for further weighted judgment and automatic core keyword identification.

References

[1] Li Wenlan, Yang Zuguo. Keyword frequency analysis of Chinese information science journal articles [J]. Information Science, 2005(1): 68-70, 143.
[2] Zhang Qin. Review of word frequency analysis application in disciplinary development dynamic research [J]. Library and Information Service, 2011(2): 95-98, 128.
[3] Tian Dan, Liu Yishan, Wang Yulin. Bibliometric analysis of hotspot analysis articles: Taking word frequency analysis as example [J]. Information Science, 2017, 35(8): 164-169.
[4] Gong Yongqiang, Liu Li. Analysis of information science research hotspots based on word frequency analysis [J]. Library Science Research, 2011(13): 9-13.
[5] An Xingru. Methodological research on word frequency analysis in China (I): Definition, classification and problems of statistical analysis elements [J]. Journal of Intelligence, 2016, 35(2): 75-80, 43.
[6] Ni Lijuan, Yu Shuli. Analysis of archival science research hotspots: Word frequency analysis of keywords in "Archival Science Research" and "Archival Science Communication" from 2004-2008 [J]. Archival Science Communication, 2010(1): 19-22.
[7] Qiu Junping, Ding Jingda. Empirical analysis of Chinese library science research from 1999-2008 (Part 2) [J]. Journal of Library Science in China, 2009, 35(6): 79-87, 118.
[8] Cang Hongyu, Tan Zongying. Analysis of information retrieval research hotspots at home and abroad: Based on Z-Score standardized word frequency [J]. Library Development, 2009(1): 93-98.
[9] Chen G, Xiao L, Zhao X G. A keyword selection method based on the combination of popularity and domain relevancy of keywords: A holistic perspective

[J]. Journal of the China Society for Scientific and Technical Information, 2014, 33(9): 959-968.
[10] Atlam E S, Fuketa M, Aoe J, et al. Similarity measurement using term negative weight and its application to word similarity [J]. Information Processing & Management, 2000, 36(5): 717-736.
[11] Chen G, Xiao L. Selecting publication keywords for domain analysis in bibliometrics: A comparison of three methods [J]. Journal of Informetrics, 2016, 10(1): 212-223.
[12] Li Shanshan, Zhang Guoqiang, Xu Guifen. Review of ERP system research hotspots based on keyword analysis [J]. Information Science, 2012, 30(8): 1272-1276.
[13] Yang Jing, Chang Chun. Research on concept word frequency change based on Logistic population growth law [J]. Information Science, 2017, 35(8): 15-18, 50.
[14] Wang Mengting. Research on theme burst detection based on change point detection [J]. Information Science, 2016, 34(12): 36-39.

Author Contributions

Feng Guohe: Proposed research ideas and designed the study;
Kong Yongxin: Conducted experiments, collected, cleaned, and analyzed data, drafted the manuscript;
Xiao Jieqiong: Responsible for final version revision.

English Abstract

A New Model for Hotspot and Trend Analysis Based on Weighted Keywords

Purpose/Significance: To overcome the limitations of absolute keyword frequency analysis, this study explores domain hotspots and trends through multi-factor weighting and ranking of keywords. **Method/Process:** The method constructs an annual-keyword frequency matrix, processes keyword frequencies through horizontal and vertical weighting, designs a relative frequency model, and calculates weighted comprehensive scores to obtain more effective keyword rankings. **Result/Conclusion:** Based on weighted keyword ranking, three keyword types can be identified—high-quantity-high-quality, low-quantity-high-quality, and burst terms—greatly benefiting hotspot mining and trend analysis.

Keywords: keyword frequency analysis; weighted keywords; hotspot research; trend analysis

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