Research on Construction of Keyword Hierarchical Structure Based on Co-occurrence Relationship Postprint

Authors: Xiong Huixiang, Chen Ziwei, Ye Jiaxin

Date: 2023-10-08T00:00:00+00:00

Abstract

[Purpose/Significance] Keywords, as the most widely applied knowledge units in literature, in-depth mining of their semantic relationships can provide underlying support for knowledge association, resource recommendation, and related tasks. [Method/Process] Mining correlations between keywords based on their direct and indirect co-occurrence relationships, analyzing keyword distribution patterns, and constructing a hierarchical structure among keywords by combining the scope of their conceptual meanings. [Results/Conclusion] Using "knowledge graph" as the root node to demonstrate the construction steps of keyword hierarchical structure; research indicates that this method possesses certain feasibility and effectiveness, and can satisfactorily construct keyword hierarchical structures.

Full Text

Abstract

[Purpose/Significance] Keywords serve as the most widely used knowledge units in scientific literature, and in-depth mining of their semantic relationships can provide foundational support for knowledge association and resource recommendation. **[Method/Process]** This study mines correlations between keywords based on their direct and indirect co-occurrence relationships, analyzes keyword distribution patterns, and constructs hierarchical structures among keywords according to their conceptual scope. **[Result/Conclusion]** Using "knowledge graph" as the root node, this paper demonstrates the construction process of a keyword hierarchical structure. The research shows that the proposed method is feasible and effective, capable of building keyword hierarchies that accurately reflect semantic relationships.

Keywords: scientific and technical literature keywords; keyword hierarchy; keyword characteristics

Introduction

Scientific and technological literature primarily comprises titles, keywords, abstracts, and full texts, among which keywords represent the most commonly used knowledge units for expressing content features. Compared to titles, keywords can capture different facets of textual content; compared to abstracts, they offer more condensed knowledge representation; and compared to full texts, they provide greater convenience and efficiency [1-3]. Consequently, keywords have become the most widely applied and closely examined knowledge unit in scientific literature.

The development and utilization of keywords primarily involve mining inter-keyword correlations to characterize texts, resources, or users, thereby establishing connections among documents, resources, and users through keyword associations to enable knowledge linking and resource recommendation. Early approaches relied heavily on dictionary resources, but due to limitations such as slow update cycles and limited coverage, research has gradually shifted toward learning keyword features from large-scale corpora, typically employing vector representations to calculate inter-keyword similarity [4]. However, keywords exhibit multiple relationships—including synonymy, hyponymy, antonymy, and homonymy—yet existing research often reduces these complex relationships to a single similarity metric. For instance, similarity mining based on keyword co-occurrence fails to distinguish among different relationship types, lacking deep semantic analysis and resulting in suboptimal performance [5-6].

From the perspective of developing the value of scientific literature, this study combines co-occurrence analysis with distributional feature analysis to construct a keyword hierarchy that reflects hyponymic relationships among research scopes, thereby enabling more effective keyword utilization and advancing related research.

Literature Review

Word Relevance Mining

Dictionary-based Approaches. Dictionary-based relevance mining relies on classification rules established during dictionary construction to identify semantic connections. WordNet is a common semantic dictionary for English that effectively mines conceptual relationships between words for document and image similarity calculations [7]; Tongyici Cilin (Synonym Forest) is a semantic dictionary organizing synonyms into a five-level hierarchical tree structure, enabling relationship mining based on this architecture [8]; HowNet is another dictionary for Chinese word relevance mining that operates based on sememes describing word concepts [9]. Additionally, combining multiple dictionaries can

expand computational coverage and improve accuracy compared to using a single dictionary [10].

Corpus-based Approaches. Corpus-based methods depend on text feature learning and representation techniques, offering substantially broader coverage than dictionary-based approaches. Current text representation primarily involves training text features into word vectors, notably through Word2vec algorithms (CBOW, Skip-gram) [11] and more recent models such as CNN, LSTM, and BERT [12-13]. Tian Xing et al. integrated Word2vec vectors into the Jaccard method for short text relevance mining, significantly improving effectiveness [14]; E. L. Pontes et al. used CNN to parse local word contexts and LSTM to analyze global sentence contexts, preserving text information to enhance relevance mining [15]; M. M. Sanjeev et al. employed BERT for semantic relevance mining between words and sentences, applying it to email retrieval [16].

While dictionary-based methods comprehensively capture word relationships with good performance, they suffer from update difficulties and limited computational scope. Corpus-based methods, though capable of automatic relationship mining across large vocabularies, require high-quality corpora and perform poorly on low-frequency or emerging terms [17].

Word Hierarchical Relationship Mining

Word hierarchical relationship mining focuses on identifying and presenting hyponymic relationships between terms—establishing superior-subordinate relationships and constructing corresponding structures. Common targets include social media tags and academic keywords. G. Tibély et al. extracted tag hierarchies from protein function and movie tags using complex network theory through network weighting and co-occurrence analysis [18]; S. Li et al. constructed keyword hierarchies based on academic keyword co-occurrence and word order in phrases [19]; Xiong Huixiang et al. established tag hierarchical relationships based on conceptual scope and co-occurrence of library tags [20-21].

Previous studies have built word hierarchies primarily on co-occurrence relationships but considered only whether words co-occur, without distinguishing semantic types or functions. This limitation makes it difficult to explain the rules governing hierarchical progression and restricts the applicability of constructed hierarchies.

Research Framework and Key Steps

To better mine word relevance, this study adopts a dictionary-inspired approach that deeply analyzes co-occurrence patterns to semi-automatically construct a word hierarchy reflecting hyponymic relationships. This hierarchy is then combined with corpus-based methods to expand coverage and improve quality. Academic keywords are selected as the research object due to their standardized, concise, and semantically clear characteristics. Based on semantic type and

function, academic keywords can be categorized into research method keywords, research topic keywords, research scope keywords, etc. [22].

Research method keywords reflect the methodologies employed in scientific literature. Mining similarities and differences in research methods across documents can effectively reveal literature connections and expand method applicability. Therefore, this study focuses on research method keywords, constructing their hierarchy by mining co-occurrence relationships with other keyword types. If a research method keyword co-occurs with multiple research topic or scope keywords, it can be inferred that the method applies broadly across topics, warranting a higher position in the hierarchy.

The proposed framework consists of three steps: data collection and preprocessing, keyword similarity calculation, and hierarchy construction, as shown in Figure 1 [Figure 1: see original paper].

Data Collection and Preprocessing

Keywords are collected from literature databases, filtered, and statistically analyzed. Following criteria in references [3] and [22], keywords are classified into research method keywords and non-research method keywords. High-frequency non-research method keywords are selected as feature terms to characterize research method keywords in subsequent analysis.

Keyword Similarity Calculation

Similarity is calculated based on keyword co-occurrence matrices. Direct co-occurrence refers to two research method keywords appearing in the same document, indicating simultaneous use in one study. Indirect co-occurrence means two research method keywords are applied within the same research topic or scope. A co-occurrence matrix of research method keywords captures direct co-occurrence, while a matrix between research method keywords and feature term keywords captures indirect co-occurrence. Cosine similarity measures the vector distance between research method keywords, yielding direct and indirect co-occurrence similarities. As this study focuses on hierarchy construction, these two similarities are weighted and integrated to obtain comprehensive co-occurrence similarity.

Establishing Keyword Hierarchy

Hierarchy construction involves five steps: (1) conceptual scope measurement, (2) root node determination, (3) conceptual scope threshold setting, (4) child node identification, and (5) level progression.

(1) Conceptual Scope Measurement. The conceptual scope of a research method keyword is measured through its co-occurrence with feature term keywords. Feature terms reflect research topics and objects; more associated feature terms indicate broader applicability and larger conceptual scope.

(2) Root Node Determination. A root node with larger conceptual scope yields a more broadly applicable hierarchy. After measuring conceptual scopes, keywords with larger scopes are selected as root nodes.

(3) Threshold Setting. To ensure keywords with similar conceptual scopes reside at the same level while maintaining distinctions between levels, scope thresholds must be established for each hierarchy level based on scope distribution analysis.

(4) Child Node Identification. After establishing the root node and thresholds, child nodes are selected based on similarity metrics. A child node must: (a) show sufficient comprehensive co-occurrence similarity with the root node, (b) demonstrate adequate direct or indirect similarity with a parent node, and (c) have a conceptual scope appropriate for its level.

(5) Level Progression. After establishing the root node as level one, child nodes are added as level two. These child nodes then become parent nodes for level three, with new nodes added based on similarity and scope thresholds. Each keyword can appear only once in the hierarchy; if a child node meets similarity thresholds with multiple parents, it attaches to the most similar parent.

Empirical Study and Results Analysis

Data Collection and Preprocessing

Considering that research methods within a discipline change little over short periods and journals maintain thematic consistency, six methodologically relevant journals were selected as data sources: *Library and Information Service*, *Information Theory and Practice*, *Journal of Intelligence*, *Information Science*, *Journal of Intelligence Studies*, and *Data Analysis and Knowledge Discovery* [23]. Fifty-five high-frequency research method keywords such as "experimental method," "empirical research," and "statistical analysis" were selected [23].

On CNKI, the search expression targeted these six core information science journals with any of the 55 keywords, covering July 2016 to June 2021, yielding 1,489 relevant documents (Table 1 shows partial data).

After collection, keywords were standardized and filtered. Synonyms were unified (e.g., "K-means," "k-means clustering," and "K-means algorithm" were standardized to "K-means"). Low-frequency keywords (frequency < 5) were removed as they hinder correlation mining. High-frequency non-research method keywords (frequency ≥ 9) were selected as feature terms. This process yielded 40 research method keywords and 48 feature term keywords (Tables 2 and 3).

Using Co-Occurrence 6.7 (COOC6.7) [24], co-occurrence matrices were constructed: Table 2 shows the matrix among research method keywords; Tables 2 and 3 together inform the matrix between research method keywords and feature term keywords.

Similarity Calculation

Direct Co-occurrence Similarity. Based on the research method keyword co-occurrence matrix, cosine similarity measures vector distances between keywords, representing direct co-occurrence similarity (Table 4).

Indirect Co-occurrence Similarity. Using the matrix between research method and feature term keywords, cosine similarity measures distances to quantify indirect co-occurrence similarity (Table 5).

Comprehensive Co-occurrence Similarity. After multiple experiments adjusting weights, equal weighting (0.5 each) yielded optimal results. The direct and indirect similarity matrices were summed and averaged to produce comprehensive co-occurrence similarity (Table 6).

Hierarchy Construction

Following the procedure in Section 3.2.3, the hierarchy was constructed. Based on the co-occurrence matrix between research method and feature term keywords, a co-occurrence count ≥ 1 indicates relevance. The number of associated feature terms represents each method' s conceptual scope (Figure 2 [Figure 2: see original paper]).

"Knowledge graph" showed the largest conceptual scope and was selected as the root node. A four-level hierarchy was constructed based on scope distribution and node counts per level. Analysis revealed significant fluctuations near scope values of 22, 15, and 8, which were set as thresholds for levels 1, 2, and 3 respectively, with level 4 set at 1.

Child node selection required comprehensive similarity ≥ 0.15 with the root node. Among 39 research method keywords, 24 met this threshold. Based on similarity results (Table 6), three child nodes were added to the root "knowledge graph" as level two. These three nodes had six child nodes (level three), which in turn had five child nodes (level four). Fourteen of the twenty-four eligible keywords were successfully integrated; ten failed to meet similarity conditions with any parent node.

The final hierarchy is shown in Figure 3 [Figure 3: see original paper].

Results Analysis

To evaluate the proposed method, hierarchies were also constructed using only direct or only indirect co-occurrence similarity (Figure 4 [Figure 4: see original paper]). Comparison reveals that single-indicator approaches produce less satisfactory results, while the comprehensive similarity approach yields richer, more extensively connected hierarchies. The constructed hierarchy effectively groups keywords with similar research scopes and clusters keywords highly relevant to the same topics into appropriate levels.

Conclusion

This study focuses on research method keywords, integrating direct and indirect co-occurrence relationships to analyze associated research scopes and construct a keyword hierarchy. Empirical results demonstrate that the proposed method builds more comprehensive and tightly connected hierarchical structures than single-indicator approaches.

Limitations and Future Work: First, indirect co-occurrence encompasses multiple scenarios, but this study only considered cases where two methods apply to the same topic or scope; future research should explore additional indirect relationships. Second, the limited dataset restricts generalizability; larger samples would better reveal keyword interrelationships and improve hierarchy construction.

References

[1] PUTRA J W G, KHODRA M L. Automatic title generation in scientific articles for authorship assistance: a summarization approach[J]. Journal of ICT research and applications, 2017, 11(3): 253-267.

[2] LUO Wei, TAN Yushan. Content-based Big Data Mining and Application of Scientific and Technical Literature[J]. Information Studies: Theory & Application, 2021, 44(6): 154-157.

[3] HU Changping, CHEN Guo. Characteristics of Keywords in Scientific Papers and Their Impact on Co-word Analysis[J]. Journal of the China Society for Scientific and Technical Information, 2014, 33(1): 23-32.

[4] HAN Pu, WANG Dongbo, ZHU Hengmin. Research on Chinese Similar Word Mining and Similarity Calculation Based on Complex Networks[J]. Journal of the China Society for Scientific and Technical Information, 2015, 34(8): 885-894.

[5] HAN Pu, WANG Dongbo, WANG Zimin. Research Progress on Word Similarity Calculation and Similar Word Mining[J]. Information Science, 2016, 34(9): 161-165.

[6] WEI Ruibin, JIANG Qianwen, ZHANG Ruili. A Comparative Study of Research Methods Based on Co-citation and Co-word Analysis: Taking Co-word Analysis and Content Analysis as Examples[J]. Journal of Intelligence, 2019, 38(2): 36-42, 4.

[7] VARELAS G, VOUTSAKIS E, RAFTOPOULOU P, et al. Semantic similarity methods in wordnet and their application to information retrieval on the web[C]//Proceedings of the 7th annual ACM international workshop on Web information and data management. New York: ACM, 2005: 10-16.

[8] TIAN Jiule, ZHAO Wei. Word Similarity Calculation Method Based on Tongyici Cilin[J]. Journal of Jilin University (Information Science Edition), 2010,

28(6): 602-608.

[9] WANG Yi, WANG Xiaolin. Word Relatedness Calculation Based on an Improved Sememe Association Algorithm[J]. Journal of the China Society for Scientific and Technical Information, 2012, 31(12): 1271-1275.

[10] ZHU Xinhua, MA Runcong, SUN Liu, et al. Word Semantic Similarity Calculation Based on HowNet and Cilin[J]. Journal of Chinese Information Processing, 2016, 30(4): 29-36.

[11] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space[EB/OL].[2022-07-31].https://doi.org/10.48550/arXiv.1301.3781.

[12] PETERS M E, NEUMANN M, IYYER M, et al. Deep contextualized word representations[EB/OL].[2022-07-31].https://doi.org/10.48550/arXiv.1802.05365.

[13] DEVLIN J, CHANG M W, LEE K, et al. Bert: pre-training of deep bidirectional transformers for language understanding[EB/OL].[2022-07-31].https://doi.org/10.48550/arXiv.1810.04805.

[14] TIAN Xing, ZHENG Jin, ZHANG Zuping. Jaccard Similarity Algorithm Based on Word Vectors[J]. Computer Science, 2018, 45(7): 186-189.

[15] PONTES E L, HUET S, LINHARES A C, et al. Predicting the semantic textual similarity with siamese CNN and LSTM[EB/OL].[2022-07-31].https://doi.org/10.48550/arXiv.1810.10641.

[16] SANJEEV M M, RAMALINGAM B, TK S K. Realtime semantic similarity analysis of bulk outlook Emails using BERT[C]//2020 International Conference on advances in computing, communication & materials (ICACCM). Piscataway: IEEE, 2020: 89-94.

[17] YAN Qiang, ZHANG Xiaoyan, ZHOU Simin. Keyword Extraction Method Based on Sememe Similarity[J]. Data Analysis and Knowledge Discovery, 2021, 5(4): 80-89.

[18] TIBELY G, POLLNER P, VICSEK T, et al. Extracting tag hierarchies[J]. PloS one, 2013, 8(12): e84133.

[19] LI S, SUN Y, SOERGEL D. A new method for automatically constructing domain-oriented term taxonomy based on weighted word co-occurrence analysis[J]. Scientometrics, 2015, 103(3): 1023-1042.

[20] XIONG Huixiang, WANG Xuedong. Research on Constructing Tag Concept Space in Folksonomy[J]. Journal of the China Society for Scientific and Technical Information, 2012, 31(9): 984-992.

[21] XIONG Huixiang, YE Jiaxin. Research on Constructing Social Tag Hierarchical Structure Based on Tongyici Cilin[J]. Journal of Intelligence, 2018, 37(1): 126-131.

[22] YE Jiaxin, XIONG Huixiang, YANG Zirong, et al. Research on the Impact of Keyword Frequency and Semantic Features on Scientific Literature Clustering[J]. Information Science, 2021, 39(8): 3-9.

[23] SUN Hongfei, HOU Wei, ZHOU Lanping, et al. Statistical Analysis of the Application of Research Methods in Chinese Information Science in the Past Five Years[J]. Information Science, 2014, 32(4): 77-84.

[24] Xueshu Diandi, Wenxian Jiliang. COOC: A New Software for Bibliometric Analysis and Knowledge Mapping[EB/OL].[2021-07-15].https://mp.weixin.qq.com/s/8RoKPLN6b1M5_{jCk1.

Author Contributions

XIONG Huixiang: Research supervision; CHEN Ziwei: Data collection, manuscript writing and revision; YE Jiaxin: Manuscript revision.

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