Dynamic Topic Identification from the Perspective of Word Co-occurrence Frequency Changes (Postprint)

Authors: Xi Chongjun, Liu Wenbin, Ding Kai

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

Abstract

[Purpose / Significance] Topic identification research is crucial for clarifying the knowledge structure and research hotspots within a domain. Conducting dynamic identification of domain topics can effectively help researchers understand and grasp the development trends and future directions of the field. [Method / Process] By utilizing the tensor data structure and incorporating the time dimension into the word co-occurrence matrix, dynamic topic identification can be performed with only a single clustering process. [Results / Conclusion] The tensor structure and non-negative tensor decomposition algorithm provide a novel method for dynamic topic identification from the perspective of word co-occurrence frequency changes. This method is simpler and more efficient compared to traditional approaches, and effectively avoids information loss.

Full Text

Abstract

Topic recognition research is crucial for clarifying the knowledge structure and research hotspots within a field. Dynamic identification of domain topics can effectively help researchers understand and grasp the development trends and future directions of a field. This study employs tensor data structures to incorporate a time dimension into word co-occurrence matrices, enabling dynamic topic identification through a single clustering process. The tensor structure and non-negative tensor decomposition algorithm provide a novel method for dynamic topic recognition from the perspective of word co-occurrence frequency changes. Compared with traditional approaches, this method is simpler, faster, and effectively avoids information loss.

Keywords: keyword co-occurrence; non-negative matrix factorization; non-negative tensor factorization; dynamic topic recognition; knowledge manage-

ment

Introduction

In the information age, the rapid growth of scientific literature has made it impossible for researchers to absorb and master tens of thousands of research outcomes within a short timeframe. Even when focusing narrowly on a specific domain through continuous reading, it remains difficult to clarify the research hotspots and directions of that field. Therefore, research on topic mining and evolution is particularly important, as it can help researchers understand domain development trends and future directions while providing an effective solution to the information overload crisis. This paper explores dynamic topic identification methods based on changes in word co-occurrence frequencies, aiming to provide better support for scientific and technological decision-making.

Topic identification and evolution research analyzes document collections using associations between document features to discover topics, revealing the content embedded in document collections and enabling understanding of current research hotspots while predicting future development trends. Scholars have conducted extensive research on topic identification and evolution analysis, with approaches ranging from shallow to deep: methods based on external citation relationships, methods based on internal word analysis, and methods based on full-text content text mining.

Methods based on citation relationships can be categorized into co-citation analysis, bibliographic coupling, and direct citation analysis. These approaches primarily utilize citation relationships between documents to determine their degree of association, thereby clustering documents to achieve topic identification. For instance, Zhu Qingsong et al. proposed a topic evolution analysis method based on co-citation of main path literature, revealing disciplinary topic evolution through co-citation analysis of key documents on citation main paths. Huang Fu et al. constructed research frontiers through coupling analysis between core literature and their cited documents, followed by co-citation analysis between core literature and their citing documents. Song Yanhui et al. used author bibliographic coupling analysis to explore the knowledge structure of informatics since the new century, based on data from seven informatics journals indexed in SCI and SSCI between 2000-2010.

Word-based analysis methods primarily include word frequency analysis and word co-occurrence analysis. Word frequency analysis identifies research priorities and hotspots by counting keyword frequency changes in literature, while word co-occurrence analysis examines relationships between words by counting their joint occurrences, clustering words to obtain topics. For example, Feng Guohe et al. conducted objective dynamic tracking and analysis of disciplinary development based on lifecycle theory and word frequency analysis. Chu Jiewang et al. analyzed research hotspots, application fields, and research methods in knowledge management over the past decade through keyword frequency

statistics. Jiang Xin et al. performed evolutionary analysis of research themes in China' s scientific data field from 2005-2016 using word frequency analysis combined with co-word analysis. Zhao Limei et al. revealed mainstream research paradigms in digital library research under the big data context using co-word analysis as the basic framework. Tang Guoyuan et al. identified five steps in the analytical process of disciplinary topic evolution research based on co-word analysis through manual interpretation, summarizing strategies, analytical methods, and tools used by researchers at each step.

Text mining methods extract topics through text mining techniques and classify them using relevant evaluation criteria. For example, Hu Jiming et al. constructed an LDA model suitable for dynamic text content topic mining. Yang Chao et al. developed an LDA topic model based on "subject-action-object" (SAO) structure to identify and analyze patent literature topic structures. J. Kim et al. conducted technology forecasting through text mining and decision tree methods, extracting features representing technology topic domains from fields such as paper authors, journals, disciplinary categories, patent assignees, and patent categories.

Research Framework

The specific research framework of this study is shown in [Figure 1: see original paper]. To incorporate a time dimension into word co-occurrence matrices and enable dynamic topic identification from the perspective of word co-occurrence frequency changes, this paper first discusses the construction methods of word co-occurrence matrices, data processing methods, and clustering approaches.

First, regarding the construction of word co-occurrence matrices: literature serves as the carrier of keywords, while authors are the main body of scientific research. The keyword sets used by these two entities reflect the knowledge structure of a field from different perspectives. Therefore, this study considers constructing keyword co-occurrence matrices from both literature and author perspectives, then fuses these two perspective-based matrices to compare differences in topic identification results obtained from the three types of keyword co-occurrence matrices.

Second, regarding the processing of word co-occurrence matrices: when conducting research based on co-occurrence data, some scholars argue that analysis can be performed directly on raw data, while others believe raw data should be standardized before analysis. In previous keyword co-occurrence based topic identification research, there is no consensus on whether and how to standardize co-occurrence matrices. Therefore, this paper standardizes keyword co-occurrence matrices from both symmetric and asymmetric perspectives, comparing the impact of performing or not performing standardization operations and different standardization approaches on topic identification results.

Third, regarding clustering methods for word co-occurrence matrices: compared with traditional clustering algorithms (hierarchical clustering, principal compo-

nent analysis, singular value decomposition, etc.), non-negative matrix factorization can effectively avoid shortcomings such as single-attribute keywords and negative weight values. Non-negative tensor decomposition represents an extension of non-negative matrix factorization into high-dimensional space. Therefore, this study first confirms the effectiveness of non-negative matrix factorization relative to traditional clustering algorithms, then compares the advantages and disadvantages of non-negative decomposition algorithms versus non-negative tensor decomposition algorithms in dynamic topic identification.

Dataset and Research Methods

Dataset Construction

This study retrieved relevant literature on knowledge management from the Web of Science database using "knowledge management" as the topic keyword, limiting document type to "article" and publication years to 2017-2021. The search yielded 4,898 documents containing 11,343 keyword fields and 12,178 author fields. After cleaning the data fields and removing the native term "knowledge management," keywords with frequency greater than 1 were selected for study. Keyword co-occurrence matrices were constructed in three ways:

(1) Literature-perspective keyword co-occurrence matrix construction. Let $KT_{m\times p}$ be the keyword-document co-occurrence matrix, where m is the number of keywords and p is the number of documents. Matrix elements represent the frequency of keywords appearing in documents. Clearly, $KT_{m\times p}$ is a 0-1 value matrix. The literature-based keyword co-occurrence matrix $AT_{m\times m}$ can be defined as:

$$AT_{m\times m} = KT_{m\times p} \times (KT_{m\times p})^T \quad \text{(Formula 1)}$$

(2) Author-perspective keyword co-occurrence matrix construction. Similarly, let $KR_{m\times q}$ be the keyword-author co-occurrence matrix, where m is the number of keywords and q is the number of authors. Matrix elements represent the frequency of authors using keywords. The author-based keyword co-occurrence matrix $AR_{m\times m}$ can be defined as:

$$AR_{m\times m} = KR_{m\times q} \times (KR_{m\times q})^T \quad \text{(Formula 2)}$$

(3) Fused literature and author dual-perspective keyword co-occurrence matrix construction. Considering that both literature-based and author-based keyword co-occurrence essentially calculate the number of times keywords co-occur, differing only in perspective—one from literature, one from authors. For a given domain, the research outcomes within a certain time period are fixed. Since scientific literature is the carrier of research outcomes and authors are the main body of scientific research, they complement each other and partition the research situation within a domain from different perspectives. Therefore, this study combines these two perspectives. The fused

literature and author keyword co-occurrence matrix $ATR_{m\times m}$ can be defined as:

$$ATR_{m\times m} = AT_{m\times m} + AR_{m\times m} \quad \text{(Formula 3)}$$

Data Processing

(1) Symmetric perspective standardization. In 2009, N. J. van Eck et al. noted that similarity measures should be used to standardize co-occurrence data, comparing several common similarity measures (association strength, cosine similarity, inclusion index, Jaccard index) and found that probability-based similarity measures (association strength) outperformed set-theory-based measures (cosine similarity, inclusion index, Jaccard index). Therefore, this study uses the association strength formula to standardize keyword co-occurrence matrices.

Taking the fused literature and author keyword co-occurrence matrix $ATR_{m\times m}$ as an example, let the element in row i and column j of matrix $ATR_{m\times m}$ be $atrij$. After similarity processing using Formula (4), matrix $ATR'_{m\times m}$ is obtained:

$$ATR'_{m\times m} = \frac{atrij}{\sum_{i=1}^{m} atrij \times \sum_{j=1}^{m} atrij} \quad \text{Formula (4)}$$

(2) Asymmetric perspective standardization. The above method standardizes the keyword co-occurrence matrix from a symmetric perspective. Although the co-occurrence frequency of two keywords is unique, the influence of individual keyword frequency means that high-frequency keywords are associated with many terms while low-frequency words are only associated with few terms. Therefore, the association degree calculated from the high-frequency word perspective differs from that calculated from the low-frequency word perspective. This study uses Formula (5) to perform asymmetric perspective similarity measurement on matrix $ATR_{m\times m}$ to obtain matrix $ATR''_{m\times m}$:

$$ATR''_{m\times m} = \frac{atrij}{\sum_{i=1}^{m} atrij} \quad \text{Formula (5)}$$

Non-Negative Matrix Factorization

Non-negative matrix factorization originated from principal component analysis and was first proposed by P. Paatero et al., called positive matrix factorization. Its basic idea is to decompose a non-negative matrix into the product of two non-negative matrices. For a keyword co-occurrence matrix $A_{R\times m}$, where m represents the number of keywords, the non-negative matrix factorization algorithm decomposes it into $m \times m^*$, where matrix $V_{r\times m}$ can be interpreted as r topics, with each row element representing the non-negative weight of m keywords in that topic. Therefore, each row of the vocabulary can be sorted by weight value to obtain the keyword types contained in each topic, and topics can be named according to keyword weight values.

Non-Negative Tensor Decomposition

A tensor is a multidimensional array. The most commonly used tensor decomposition methods are CP decomposition and Tucker decomposition. CP decomposition decomposes an n-order tensor into a sum of multiple rank-1 tensors, while Tucker decomposition decomposes it into the product of a core tensor and several factor matrices. The core tensor can be seen as a condensed form of the original tensor. When the core tensor is a diagonal tensor, Tucker decomposition degenerates into CP decomposition (see [Figure 2: see original paper]). Non-negative tensor decomposition is an extension of non-negative matrix factorization into high-dimensional space. It retains the advantages of tensors while avoiding negative elements and is widely applied in image processing, audio classification, and text mining.

When using non-negative tensor decomposition for topic identification, an appropriate tensor must first be constructed. Taking a three-order tensor as an example, since this study conducts dynamic topic identification based on changes in keyword co-occurrence frequency, we construct a three-order tensor $X_{I\times I\times K}$ of <keyword, keyword, year>. As shown in [Figure 3: see original paper], the black circles in the keyword co-occurrence matrix represent the co-occurrence strength between keywords. Performing non-negative tensor decomposition on this tensor yields factor matrices $A_{I\times R}$, $B_{R\times I}$, $C_{K\times R}$, and core tensor $\Lambda_{R\times R\times R}$, where I represents the number of keyword types, K represents the number of years, and R represents the number of clusters. Similar to non-negative matrix factorization results, factor matrices $A_{I\times R}$ and $B_{R\times I}$ in non-negative tensor decomposition can be interpreted as R topics and the keyword types and weight values contained in each topic, with consistent clustering results in both factor matrices. Additionally, factor matrix $C_{K\times R}$ can be interpreted as the weight values of R topics in each year, i.e., topic research intensity. The core tensor $\Lambda_{R\times R\times R}$ can be interpreted as the comprehensive strength of R topics. Thus, the three-order tensor <keyword, keyword, year> is reduced to a two-order matrix <topic, year>, enabling dynamic topic identification. As shown in [Figure 3: see original paper], the size of black circles in the topic boxes represents the intensity of topics appearing in that year.

Experimental Results

Dataset Construction Group Experiment Results

Through multiple experiments, when the number of clusters exceeds 5, some clusters exhibit highly overlapping keywords. Therefore, this study sets the number of clusters to 5. The non-negative matrix factorization clustering results for the three keyword co-occurrence matrices are shown in . The results show that in non-negative matrix factorization clustering, the weight values of keywords in each cluster are non-negative, overcoming the limitation in principal component analysis where weight values can be positive or negative. Additionally, keyword types are repeated across clusters, overcoming the limitation in

hierarchical clustering where a keyword belongs to only one cluster, which aligns with real-world situations. Specifically, the clustering results under the three keyword co-occurrence matrices show both similarities and differences.

First, the dominant words (keywords with the highest weight values) in each cluster are basically consistent across the three keyword co-occurrence matrices. These dominant words can assist in cluster naming, indicating that whether from a literature or author perspective, the research hotspots in the foreign knowledge management field over the past five years are basically the same, mainly including Knowledge Sharing, Innovation, Intellectual Capital, Knowledge, Organizational Performance, and SEMs. The differences lie in the research directions under each major topic (i.e., the types of keywords with low weight values differ). For example, in the Innovation theme from the literature perspective, keywords sorted by weight value are SMEs, Performance, Dynamic Capabilities, Entrepreneurship, etc., while from the author perspective they are SMEs, Dynamic Capabilities, Organizational Performance, Information Technology, etc. Both perspectives focus on enterprise innovation, but the literature perspective emphasizes entrepreneurship while the author perspective emphasizes information technology.

Furthermore, using the Jaccard similarity algorithm to calculate the association degree between topics in each clustering result yields statistical data including mean, range, and standard deviation (see [Figure 4: see original paper]-[Figure 6: see original paper]). The results show that the literature-perspective clustering results have the highest mean association degree between each topic and other topics in the same clustering result, with the smallest range and standard deviation. The author-perspective clustering results have relatively low mean association degrees with larger ranges and standard deviations. The fused literature and author perspective clustering results fall between the single-perspective results. This indicates that topic differentiation is more pronounced in author-perspective clustering results than in literature-perspective results. Since the number of documents far exceeds the number of authors, literature-perspective clustering results can conduct in-depth mining of domain topics, while author-perspective clustering results can provide comprehensive identification of domain topics. Combining the number of keywords contained in each topic under the three clustering results (see [Figure 7: see original paper]), the literature perspective includes more keyword types per topic than the author perspective, indicating more in-depth and detailed topic content mining. Therefore, the fused literature and author keyword co-occurrence matrix can both comprehensively and deeply reflect domain research situations compared to single-perspective matrices.

Dataset Processing Group Experiment Results

The first group of experiments indicates that the fused literature and author dual-perspective keyword co-occurrence matrix better reflects domain research situations. Therefore, this study continues analysis using this matrix. First,

the fused literature and author dual-perspective keyword co-occurrence matrix is standardized from both symmetric and asymmetric perspectives, then non-negative matrix factorization is applied to cluster the matrices before and after standardization. The clustering results are shown in .

The results show that clustering results from the raw co-occurrence matrix and the asymmetrically standardized matrix share some dominant words (such as Knowledge Sharing, Innovation, Knowledge, etc.), while the symmetrically standardized matrix shows significant differences. Examining the raw data reveals that dominant words in clustering results from raw and asymmetrically standardized matrices are generally high-frequency keywords with obvious weight value differences within clusters. In contrast, keywords in symmetrically standardized clustering results have relatively low frequencies with small weight differences within clusters. This occurs because symmetric perspective standardization eliminates the influence of high-frequency keywords. Additionally, asymmetrically standardized clustering results not only cluster high-frequency keywords but also aggregate some low-frequency keywords, as some keywords, although low in frequency, consistently co-occur with other terms, indicating high association and thus being clustered together—a feature absent in the other two clustering results.

These results indicate that using raw keyword co-occurrence matrices or performing asymmetric perspective standardization can analyze hotspot research topics within a domain, as high-frequency keywords often represent research priorities and hotspots. Asymmetrically standardized clustering results can more comprehensively analyze domain research situations by including low-frequency keyword aggregation. Symmetric perspective standardization can analyze the latest frontier research trends by eliminating high-frequency keyword influence while preserving keyword associations.

Dynamic Topic Recognition Results Analysis

Based on previous experimental results, the third group of experiments continues using the fused literature and author dual-perspective keyword co-occurrence matrix with asymmetric perspective standardization, then compares the advantages and disadvantages of non-negative matrix factorization and non-negative tensor decomposition in dynamic topic identification. Since non-negative matrix factorization processes data in matrix form, the keyword co-occurrence matrices for 2017-2021 must be time-sliced by year, requiring 5 separate clustering operations where each year' s dataset consists of the co-occurrence matrix of all keywords appearing that year. Non-negative tensor decomposition can process high-dimensional data, enabling direct clustering of all keywords from 2017-2021 by first constructing a three-order tensor divided into 5 slices along the year dimension, with each slice representing the co-occurrence matrix of all keywords appearing in 2017-2021 for a particular year. The clustering results of both algorithms are shown in .

The results show that under non-negative matrix factorization, the main research hotspots from 2017-2021 are roughly the same each year (with similar dominant keywords in each cluster), though research directions and granularity differ slightly (with differences in keyword quantities and types per cluster). Non-negative tensor decomposition performs clustering only once for 2017-2021, with results generally consistent with non-negative matrix factorization (dominant words in non-negative tensor decomposition results are those appearing frequently across the 5 years in non-negative matrix factorization results). However, since non-negative tensor decomposition performs only one clustering, the research content of identical topics across years remains consistent and relatively comprehensive, while non-negative matrix factorization clusters each year separately, potentially resulting in similar topics with different content across years —thus providing more detailed characterization of topic research content.

Furthermore, by calculating topic similarity using the Jaccard similarity algorithm on yearly clustering results from non-negative matrix factorization, a topic evolution trajectory map is obtained (see [Figure 8: see original paper]). In contrast, non-negative tensor decomposition results can use the core tensor to generate yearly topic research intensity diagrams (see [Figure 9: see original paper]), where research intensity is not measured by keyword quantity or frequency but by the evolution of keyword co-occurrence relationships across years —a capability difficult for non-negative matrix factorization to achieve.

These results indicate that non-negative tensor decomposition can simply and quickly obtain domain research topics and their yearly research intensity with only one clustering, greatly reducing algorithmic complexity and information loss. For detailed analysis of yearly research situations, non-negative matrix factorization can be used for year-by-year analysis to obtain specific research content and changes for each year, as well as topic evolution across years, though this requires multiple clustering operations and data processing and makes it difficult to observe topic evolution driven by keyword co-occurrence changes.

Conclusion

In summary, when using keyword co-occurrence data for domain topic identification, the fused literature and author dual-perspective keyword co-occurrence matrix better reflects domain research situations. When using co-occurrence data for topic identification, similarity measurement standardization is necessary. Symmetric perspective standardization can eliminate high-frequency keyword influence for analyzing frontier trends, while asymmetric perspective standardization can study hotspot issues. In dynamic topic identification, non-negative tensor decomposition can simply and quickly obtain domain research topics and their yearly research intensity, while non-negative matrix factorization can more detailedly characterize topics and their evolution trajectories, though requiring multiple operations.

This study addresses the limitation of traditional dynamic topic identification

based on word co-occurrence matrices requiring multiple clustering operations by proposing a new data construction and processing method. The tensor structure incorporates time dimensions into word co-occurrence matrices, preserving original data information as much as possible. Non-negative tensor decomposition for dynamic topic identification requires only one clustering to obtain yearly topic situations, effectively avoiding information loss. Additionally, this paper discusses several word co-occurrence matrix construction methods and processing approaches: in dataset construction, keyword co-occurrence matrices were built from literature perspective, author perspective, and fused dual perspectives; in data processing, similarity measures were applied from symmetric and asymmetric perspectives to standardize co-occurrence matrices, comparing standardization impacts on topic identification results. Experimental results demonstrate that the fused literature and author dual-perspective keyword co-occurrence matrix can more comprehensively reflect domain knowledge structure, with symmetric and asymmetric perspective standardization each having advantages in analyzing research hotspots and frontiers. This study aims to provide methodological and procedural references for keyword co-occurrence based topic identification research, improving topic identification accuracy and providing better support for scientific and technological decision-making.

References

[1] BUSH V. As we may think[J]. The Atlantic monthly, 1945 (7): 1-2 . [2] Liu Xiang, Ma Feicheng, Chen Xiaojun, et al. Structure and evolution of knowledge networks—Conceptual and theoretical progress[J]. Information Science, 2011, 29(6): 801-809. [3] Ba Zhichao, Yang Zijiang, Zhu Shiwei, et al. Research on domain topic evolution analysis method based on keyword semantic network[J]. Information Studies: Theory & Application, 2016, 39(3): 67-72. [4] Wang Liya. Research progress on topic evolution[J]. Information Research, 2014(4): 29- [5] Shao Zuoyun, Li Xiuxia. Review of literature knowledge discovery methods combining citation analysis and content analysis[J]. Information Studies: Theory & Application, 2020, 43(3): [6] Zou Lixue, Wang Li, Liu Xiwen. Research progress on topic models constructed using citations[J]. Library and Information Service, 2019, 63(23): 131-138. [7] Zhu Qingsong, Leng Fuhai. Topic evolution analysis based on co-citation of main path literature[J]. Journal of the China Society for Scientific and Technical Information, 2014, 33(5): 498-506. [8] Huang Fu, Hou Haiyan, Ren Peili, et al. Selection of research front detection methods based on co-citation and bibliographic coupling[J]. Journal of Intelligence, 2018, 37(12): 13-19, 35. [9] Song Yanhui, Wu Yishan. Research on the knowledge structure of informatics based on author bibliographic coupling analysis[J]. Library and Information Service, 2014, 58(1): 117-123. [10] Zhang Jie, Wang Hong. Comparative analysis of research hotspots in mobile learning at home and abroad based on word frequency analysis and visualized co-word network diagrams[J]. Modern Distance Education, 2014(2): 76-83. [11] Ye Chunlei, Leng Fuhai. Research on improvement of disciplinary topic evolution method based on co-word analysis[J]. Information Studies: Theory & Application, 2012, 35(3): 79-82.

[12] Feng Guohe, Kong Yongxin. Research on disciplinary hotspots based on time-weighted keyword frequency analysis[J]. Journal of the China Society for Scientific and Technical Information, 2020, 39(1): 100-110. [13] Chu Jiewang, Qian Qian. Research hotspots and research methods of knowledge management in the past 10 years based on word frequency analysis[J]. Information Science, 2014, 32(10): 156- [14] Jiang Xin, Wang Dezhuang, Ma Haiqun. Evolutionary analysis of research themes in China' s "scientific data" field from the perspective of keyword frequency changes[J]. Modern Intelligence, 2018, 38(1): 141-146, 161. [15] Zhao Limei, Zhang Hua. Analysis of research frontiers of digital libraries in China' s big data era—From the perspective of co-word analysis[J]. Information Science, 2019, 37(3): 97-104. [16] Tang Guoyuan, Zhang Wei. Research progress and analysis of disciplinary topic evolution based on co-word analysis method[J]. Library and Information Service, 2015, 59(5): 128-136. [17] Hu Jiming, Chen Guo. Content topic mining and evolution based on dynamic LDA topic model[J]. Library and Information Service, 2014, 58(2): 138-142. [18] Yang Chao, Zhu Donghua, Wang Xuefeng, et al. Patent technology topic analysis: LDA topic model method based on SAO structure[J]. Library and Information Service, 2017, 61(3): 86-96. [19] KIM J, HWANG M, JEONG D H, et al. Technology trends analysis and forecasting application based on decision tree and statistical feature analysis[J]. Expert systems with applications, 2012, 39(16): 12618-12625. [20] WALTMAN L, VANECK N J. Some comments on the question whether co-occurrence data should be normalized[J]. Journal of the American Society for Information Science and Technology, 2007, 58(11): 1701- [21] LEYDESDORFF L. Should co-occurrence data be normalized? a rejoinder[J]. Journal of the American Society for Information Science and Technology, 2007, 58(14): 2411-2413. [22] van ECK N J, WALTMAN L. How to normalize cooccurrence data? an analysis of some well-known similarity measures[J].Journal of the American Society for Information Science and Technology, 2009, 60(8): 1635- [23] PAATERO P, TAPPER U. Positive matrix factorization: a nonnegative factor model with optimal utilization of error estimates of data values[J]. Environmetrics, 1994, 5(2): [24] Zhang Xiangsun, Zhang Zhongyuan. Non-negative matrix factorization: Models, algorithms and applications[J]. Journal of Chongqing Normal University (Natural Science Edition), 2013, 30(6): 1-8. [25] Wu Jibing, Huang Hongbin, Deng Su. Tensor decomposition clustering method for network heterogeneous information[J]. Journal of National University of Defense Technology, 2018, 40(5): 146-152, 170. [26] Xiong Liyan, He Xiong, Huang Xiaohui, et al. Review of tensor decomposition algorithms research and application[J]. Journal of East China Jiaotong University, 2018, 35(2): 120-128. [28] LUO J, GWUN O. A comparison of sift PCA-SIFT and SURF[J]. International journal of image processing, 2009, 3(4): 143-152. [29] C I C H O C K I A , Z D U N E K R , P H A N A H , e t a l . Nonnegative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation[M]. Hoboken: Wiley Publishing, 2009.

Author Contributions: Fang Jie proposed the research idea, provided guidance, and revised the paper; Cui Lanlan collected data, designed the research

approach, performed data analysis, and wrote and revised the paper.

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