Research on Automatic Identification of Domain Groups in Scientific Structure Maps (Postprint)

Authors: Wang Xiaomei, Deng Qiping

Date: 2017-10-11T00:00:00+00:00

Abstract

【Purpose】To explore automatic identification methods for research field clusters in science structure maps, rapidly outline the overall landscape of science structure, and enhance timeliness. 【Methods】 Thematic similarity among research fields is measured using characteristic words, while simultaneously considering their relative positional relationships to group positionally adjacent and thematically similar fields into clusters. Effectiveness evaluation metrics are designed to compare optimal parameter combinations across different methods and recommend the optimal approach. 【Results】 The method can effectively identify field clusters in science structure maps across different periods. 【Limitations】 The method' s effectiveness is based on experimental results from "science structure map" data; whether the parameter combinations are applicable to other datasets requires further validation. 【Conclusion】 This provides an effective method for automatic identification of field clusters in science structure maps.

Full Text

Auto-Identifying Research Area Groups in Science Map

Wang Xiaomei[1], Deng Qiping[1,2]

[1] National Science Library, Chinese Academy of Sciences, Beijing 100190, China
[2] University of Chinese Academy of Sciences, Beijing 100049, China

This work was supported by the National Natural Science Foundation of China project "Analysis Methods and Applications for Characteristics and Evolution Dynamics of Science Structure" (Project No. 71173211).

Abstract

[**Objective**] This study explores automatic identification methods for research area groups in science maps to rapidly outline the overall structure of science

and improve timeliness. [**Methods**] We measured thematic similarity among research areas using feature words while considering their relative spatial relationships, grouping adjacent areas with similar topics into domain clusters. An effectiveness evaluation index was designed to compare optimal parameter combinations across different methods and recommend the best approach. [**Results**] The proposed method effectively identified research area groups in science maps across different periods. [**Limitations**] The method' s effectiveness was validated based on experimental results from "Mapping Science Structure" data; whether the parameter combinations are applicable to other datasets requires further verification. [**Conclusions**] This study provides an effective method for the automatic identification of research area groups in science maps.

Keywords: Science structure, Research area, Research area groups, Automatic detection
Classification: TP393, G35

The scientific knowledge structure system has long been a research focus. Science maps visualize highly abstract scientific fields, particularly the macrostructure of natural science and basic research, through intuitive 图谱 forms that reveal relationships and developmental processes among scientific hotspots and frontiers, enabling researchers to grasp the overall scientific landscape quickly, comprehensively, and vividly. As early as 1965, De Solla Price hypothesized that scientific structure systems were implicit in literature databases and proposed that integrating citation relationships among journals could reveal disciplinary structures and even delineate research directions in greater detail [1]. Subsequently, Carpenter et al. mapped disciplinary subfields by clustering journals from the SCI database [2], while Small et al. pioneered automatic detection of scientific structure by using computer technology to cluster highly cited scientific literature and identify overall specialty structures and their interrelationships [3]. In recent years, science mapping research has continued to evolve alongside developments in bibliometric methods.

A critical challenge in science map research is interpretation. Current analytical units for science map construction primarily include citations, keywords, authors, and journals, each with corresponding co-occurrence analyses such as co-citation, co-word, co-author, and journal citation analyses. Co-citation analysis particularly enhances understanding of how scientific discoveries contribute to thematic development and provides more detailed characterization of disciplinary structures. However, in practical applications, science maps generated through direct co-citation clustering—using research areas as basic units—often lack readability. Presenting hundreds of research areas simultaneously makes it difficult for readers to extract useful information intuitively. Consequently, researchers typically conduct content analysis of each research area and manually group them into domain clusters (research categories) based on content similarity, naming these clusters to create an overview of the entire scientific structure that reflects relationships among major research categories. Therefore, domain cluster identification holds significant importance in science map research. Yet,

as science evolves with emerging research areas appearing and existing ones disappearing, science maps continuously change. Rapidly and effectively constructing science maps and identifying their domain clusters represents both a key requirement and a major challenge. Previous studies have primarily relied on manual interpretation; this research aims to explore an automatic method for identifying domain clusters in science maps.

2.1 Drawing Steps of Science Structure Map

The process involves three main steps: selecting analytical units such as citations, keywords, authors, or journals, each with corresponding co-occurrence analyses; establishing relationships among these units; and visualizing them in low-dimensional space (typically two-dimensional) to display both the units and their interconnections. The science map discussed in this paper was developed by our research team; specific drawing methods are detailed in references [5-8]. The analytical unit is citations, forming "research areas" containing several research papers through co-citation clustering of highly cited papers from Thomson Reuters' Essential Science Indicators (ESI) database. A gravity model algorithm determines each research area' s layout position in two-dimensional space, with fixed initial positions for the first period and subsequent periods using parallel mapping from the previous period to ensure stable and comparable layouts. Relative positions among research areas reflect their degree of association, with closer distances indicating stronger correlations.

2.2 Research Area Group Identification Methods

Science maps drawn through the above process contain numerous research areas, making it impractical to label each area's name in the diagram and thus limiting direct information provision. To enhance readability, researchers must partition research areas into domain clusters, also called research categories or quasi-disciplinary structures, which represent higher-level scientific organization. In science maps generated through citation clustering, domain cluster identification typically employs manual interpretation, with automatic identification methods still in exploratory stages.

Manual interpretation represents the conventional approach, where researchers provide domain experts with paper lists from citation clusters and extracted keyword information. Experts name each research area and assign it to appropriate research categories based on the provided keywords and paper information. Results are then used to draw irregular regions in the science map, grouping research areas belonging to the same category to generate domain clusters. This method was used in the Mapping Science Structure series [Figure 2009: see original paper] [6], [Figure 2012: see original paper] [7], [Figure 2015: see original paper] [8] and in similar studies on science structure evolution by Japan' s National Institute of Science and Technology Policy (NISTEP) [9]. While manual interpretation yields the most accurate results, its workflow is cumbersome, demands high expertise, and delays publication timelines, necessitating

an effective automatic alternative.

Research on automatic identification methods remains limited. Our team attempted three-level clustering in Mapping Science Structure 2015 [8], constructing citation relationships among research areas for tertiary clustering to automatically identify domain clusters, but results were unsatisfactory. Analysis revealed two main limitations: first, citation relationships at the tertiary level represent aggregations of paper-level citations, creating excessive abstraction that amplifies and distorts relationships, thereby reducing clustering accuracy; second, relative positions among research areas reflect their degree of association, making spatial relationships crucial for identification, yet this method ignored positional information, yielding low accuracy.

To improve automatic identification accuracy, researchers have attempted to utilize spatial relationships among research areas. Boyack et al. defined quasi-disciplinary structures and employed a semi-automatic method to partition research areas into groups. Their approach divided the science map into grids, selected specific grids as disciplinary seeds, and used the number of shared documents in overlapping circumscribed circles as a connection mechanism to join adjacent grids or grid groups with the most shared documents, repeating until all grids were connected into a quasi-disciplinary structure [10]. Although this method considered spatial relationships, it depended heavily on the number of selected seeds, with different seed quantities producing substantially different domain clusters. NISTEP' s 2014 Science Map 2010 & 2012 proposed a method combining spatial relationships with thematic similarity, similarly dividing the map into grids and sorting them by paper count. It calculated the number of shared feature words within a certain range, considering grids with values exceeding a threshold as belonging to the same candidate domain cluster, repeating until all grids were assigned, then deleting and merging candidate clusters according to specific rules [11]. While NISTEP' s method combined spatial and thematic similarity for relatively accurate identification, our experiments revealed low discriminative power in dense regions and high parameter sensitivity, making practical application difficult due to multiple involved parameters.

Research area positions are obtained through layout algorithms, where closer distances indicate stronger citation relationships and higher likelihood of belonging to the same domain cluster, making spatial relationships essential for identification. Spatial relationships reflect co-citation proximity, and incorporating text analysis to assess thematic similarity can further improve partitioning accuracy. Addressing the aforementioned limitations, this study draws upon NISTEP' s 2014 method, experimenting with two thematic similarity measurement approaches to improve candidate cluster identification, aiming to enhance discriminative power and precision in dense regions. We evaluate method effectiveness across multiple dimensions including F-values, cluster counts, and overlaps to identify the most effective method and optimal parameter combination.

3.1 Method Introduction

Association Similarity Measurement: Based on layout principles, research areas in closer proximity exhibit stronger associations. This study divides the science map into grids according to research area coordinate ranges, with areas within the same grid defaulting to the same domain cluster. Thematic similarity establishes inter-grid connection mechanisms, linking grids with similar content within specified ranges into domain clusters. Grid division must balance density: overly sparse grids contain too many research areas, hindering discrimination of adjacent clusters, while overly dense grids contain too few areas, failing to establish effective connection mechanisms and producing numerous small clusters.

Thematic Similarity Measurement: We employed two text similarity-based approaches to measure grid thematic similarity. First, feature word count measurement uses the Alchemy API [12] to extract feature words describing research area themes from paper titles and abstracts. For any two grids requiring similarity measurement, we select the n most frequent feature words from their constituent research areas and use the count of shared feature words as similarity, considering grids with similarity above a threshold as belonging to the same domain cluster. Second, feature vector measurement extracts all words from research area paper titles and abstracts, converts them to feature vectors after stopword removal, and uses cosine similarity between vectors to represent thematic similarity. When the average similarity between research areas in two grids exceeds a threshold, the grids are considered thematically similar and belonging to the same domain cluster.

Candidate Domain Cluster Identification: We propose a dynamic identification method that combines positional association similarity and thematic similarity to improve candidate cluster identification precision. The static feature word method, based on NISTEP' s approach, uses fixed feature word sets for similarity measurement, treating all grids above threshold as equally similar and assigning them to the same cluster. However, research area distribution in science maps is uneven, with dense regions potentially containing multiple domain clusters exhibiting substantial thematic similarity variation. Sparse regions contain fewer shared feature words, making it difficult for static methods to simultaneously distinguish clusters in dense regions while identifying those in sparse regions.

To address this limitation, we propose dynamic feature word extraction, where candidate cluster feature words evolve during identification. The process gradually incorporates the most similar grids into candidate clusters, re-extracting feature words after each iteration to measure similarity dynamically. This better identifies domain clusters within dense regions. As shown in [Figure 2: see original paper], grids surrounding a candidate grid are organized in layers to limit cluster scale. Research areas in closer proximity have higher probability of belonging to the same cluster, so identification begins from the first layer.

Starting with candidate grid A as the center, the most thematically similar grid in the first layer (B6 in the figure) is incorporated into candidate cluster A. Feature words for cluster A are then re-extracted to measure similarity with remaining grids, and B4 is incorporated next. This process repeats until all grids in the first layer exceeding the threshold are included, after which they are removed from candidate grids. The same method processes other specified layers, ultimately producing a candidate domain cluster centered on grid A.

3.2 Specific Process

The implementation involves six steps. First, the science map is divided into grids, typically 20×20 or 30×30, adjustable according to research area distribution. Second, grid density (paper count) is calculated, and grids are sorted descending by density as candidate grids. Third, candidate domain clusters are identified sequentially using the aforementioned methods until all candidate grids are assigned; clusters containing only one research area are removed. Fourth, the maximum and minimum X and Y coordinates of research areas in each candidate cluster are determined along with the center point, and an ellipse is drawn centered at the center point with (Xmax–Xmin) as X-axis length and (Ymax–Ymin) as Y-axis length to identify the cluster. Fifth, overlapping clusters are deleted: clusters completely contained within others are removed; for ellipses defined by $X^2/A^2+Y^2/B^2=1$, when another ellipse' s center (x1, y1) satisfies $x1^2/A^2+y1^2/B^2<0.5$, the smaller cluster is deleted; after grid refinement, clusters with over 80% of grids contained in other clusters are deleted. Sixth, clusters with substantial overlap are merged: when overlap similarity exceeds the merge threshold (30 shared feature words for feature word methods, 0.15 for feature vector methods), the two clusters are merged. Smaller overlaps are permitted as they reflect interdisciplinary content.

3.3 Effectiveness Evaluation

Domain cluster identification essentially groups thematically similar research areas. Our clustering is fuzzy, allowing a research area to belong to multiple clusters. We validated results using a modified F-value metric [13] for clustering effectiveness. In practice, automatic identification should approximate manual labeling to reflect the main structure of science, so we compared automatic results with manual interpretations. For each manually labeled domain cluster Pj, we assume a corresponding automatically identified cluster Ai exists (unknown initially). To find Ai, we traverse all clustering results, calculating precision, recall, and F-value, selecting the cluster with optimal F-value. The F-value for Pj is calculated as the maximum F(Pj, Ai) across all clusters. The overall result F-value is computed as the weighted average: $F = \Sigma(wj \times F(Pj))$, where $wj = |Pj| / \Sigma|Pj|$.

The original F-value assumes a one-to-one correspondence between manual and automatic clusters, but automatic identification typically produces more clusters than manual labeling, with some automatic clusters representing finer granular-

ity. Therefore, we merge automatic clusters where 70% of research areas belong to the same manual cluster before calculating F-values. Since some research areas lack corresponding clusters, we further consider discriminative power using a modified Fdi to represent method effectiveness: Fdi = (Total research areas - Unidentified research areas) / Total research areas.

4 Method Comparison and Analysis

We conducted comparative analysis using data from Mapping Science Structure 2015 [8], programming the described methods and processes to generate domain cluster visualizations across different periods and parameters. Using the modified Fdi metric, we compared the effectiveness of improved methods against original approaches and examined parameter impacts to identify optimal combinations.

Experiments used two periods: Science Map 2006 [Figure 2006: see original paper]-2011 and Science Map 2008 [Figure 2008: see original paper]-2013, shown in [Figure 3: see original paper]. Each circle represents a research area, with size proportional to paper count and numbers indicating area IDs. The 2006-2011 map contains 149 research areas, while 2008-2013 contains 212, manually divided into 10 major categories (domain clusters with 2 research areas) shown as irregular colored regions. Category names are listed in .

4.2 Effectiveness Analysis

We validated three methods across multiple dimensions including Fdi values, cluster counts, and overlap situations, comparing different methods and parameters to identify optimal combinations. The science map size is 360×480 pixels; research area coordinates were translated to fit the canvas. Adjustable parameters include grid size, similarity threshold, and cluster scale. We tested 20×20 and 30×30 grids; feature word methods used thresholds of 2, 3, and 4 shared words; feature vector methods used thresholds of 0.07 and 0.12. Cluster scale in static methods was set to 70 pixels (absolute distance) based on experimentation, while dynamic methods used relative distance measured in grid layers (set to 2 in tests). and present overall Fdi values for both periods.

Results demonstrate that our proposed dynamic identification method generally outperforms others across both periods. Dynamic feature word methods achieve higher Fdi values with smaller variation, indicating greater precision and lower parameter sensitivity. For denser maps (2008-2013), 30×30 grids perform better, with Fdi decreasing as shared feature word thresholds increase. For sparser maps (2006-2011), 20×20 grids yield higher Fdi values, with optimal performance at a threshold of 3.

Cluster counts increase and sizes decrease with higher grid density and similarity thresholds. Variation across methods and parameters is modest: 13-16 clusters for 2008-2013 and 14-19 for 2006-2011. Dynamic feature word methods produce

more clusters than static methods, as they more reliably distinguish strongly related fields like condensed matter physics, nanotechnology, and synthetic chemistry. Overlap patterns mirror cluster count trends, increasing with grid density and thresholds. Feature vector methods show lowest overlap, followed by static feature word methods, with dynamic feature word methods showing highest overlap—primarily among strongly related and interdisciplinary fields, which we consider reasonable.

Overall, static feature word methods exhibit low discriminative power in dense regions, oversimplified thresholding, fewer identified clusters, lower overlap, minimal Fdi values, and high parameter sensitivity—particularly with 30×30 grids where cluster counts increase rapidly and effectiveness drops significantly. Dynamic feature word methods effectively identify strongly related disciplinary areas, producing more clusters with higher overlap and larger Fdi values. Feature vector methods, while achieving high Fdi values in some cases, show instability across datasets, poor performance on small networks, and high computational complexity for large datasets without manual references for parameter selection.

This study explored automatic identification methods for domain clusters in science maps using research areas as basic units, identifying optimal parameter combinations through experimental comparison. By combining spatial relationships with thematic relevance, dividing maps into grids, using feature words to measure thematic similarity, and establishing connection mechanisms, we employed three identification methods and evaluated effectiveness using modified F-values based on manual labels. The improved dynamic identification method accurately identifies domain clusters, particularly in dense regions, with lower parameter sensitivity and greater stability. While feature vector methods may achieve high Fdi values, their instability across datasets and computational demands suggest using dynamic feature word methods or hybrid approaches for practical applications. All conclusions are based on experiments with Mapping Science Structure 2015 data; future work should validate these methods on other datasets.

References

[1] De Solla Price D J. Networks of Scientific Papers [J]. Science, 1965, 149(3683): 510-515.
[2] Carpenter M P, Narin F. Clustering of Scientific Journals [J]. Journal of the American Society for Information Science, 1973, 24(6): 425-436.
[3] Small H, Griffith B C. The Structure of Scientific Literatures I: Identifying and Graphing Specialties [J]. Science Studies, 1974, 4(1): 17-40.
[4] Chen Chaomei. Mapping Scientific Frontiers: The Quest for Knowledge Visualization [M]. Beijing: Science Press, 2014.
[5] Wang Xiaomei, Han Tao, Wang Jun, et al. Mapping Science Based on Co-citation Analysis [J]. Science Focus, 2009, 4(4): 1-15.
[6] Pan Jiaofeng, Zhang Xiaolin, Wang Xiaomei, et al. Mapping Science Structure 2009[M]. Beijing: Science Press, 2010: 12-18.

[7] Pan Jiaofeng, Zhang Xiaolin, Wang Xiaomei, et al. Mapping Science Structure 2012[M]. Beijing: Science Press, 2013: 13-18.
[8] Wang Xiaomei, Han Tao, Wang Jun, et al. Mapping Science Structure 2015[M]. Beijing: Science Press, 2015: 10-34.
[9] NISTEP. Science Map 2008[R/OL]. [2010-05-11]. http://data.nistep.go.jp/dspace/bitstream/11035/686/1/N
NR139-FullJ.pdf.
[10] Boyack K W, Klavans R. Creation of a Highly Detailed, Dynamic, Global Model and Map of Science[J]. Journal of the Association for Information Science and Technology, 2014, 65(4): 670-685.
[11] NISTEP. Science Map 2010 & 2012 [R/OL]. [2014-07-11]. http://data.nistep.go.jp/dspace/bitstream/11035
NR159-FullJ.pdf.
[12] Alchemy API. AlchemyLanguage Features [EB/OL]. [2014-10-15]. http://www.alchemyapi.com/products/alchemylanguage.
[13] Zhou Zhaotao. Quality Evaluation of Text Clustering Result and Investing on Text Representation [D]. Beijing: The Graduate School of Chinese Academy of Sciences, 2005.

Conflict of Interest Statement: All authors declare no conflict of interest.

Supporting Data: Supporting data is self-archived by the authors, E-mail: wangxm@mail.las.ac.cn.
[1] Wang Xiaomei, Deng Qiping. RACoordinate2006_2011.csv. Research area coordinates for "Science Map 2006~2011".
[2] Wang Xiaomei, Deng Qiping. RASize2006_2011.csv. Research area paper counts for "Science Map 2006~2011".
[3] Wang Xiaomei, Deng Qiping. RACategories2006_2011.csv. Research area disciplinary classifications for "Science Map 2006~2011".
[4] Wang Xiaomei, Deng Qiping. FeatureWords2006_2011.json. Research area feature words for "Science Map 2006~2011".
[5] Wang Xiaomei, Deng Qiping. RAGroups2006_2011.csv. Research group identification results for "Science Map 2006~2011".
[6] Wang Xiaomei, Deng Qiping. RACoordinate2008_2013.csv. Research area coordinates for "Science Map 2008~2013".
[7] Wang Xiaomei, Deng Qiping. RASize2008_2013.csv. Research area paper counts for "Science Map 2008~2013".
[8] Wang Xiaomei, Deng Qiping. RACategories2008_2013.csv. Research area disciplinary classifications for "Science Map 2008~2013".
[9] Wang Xiaomei, Deng Qiping. FeatureWords2008_2013.json. Research area feature words for "Science Map 2008~2013".
[10] Wang Xiaomei, Deng Qiping. RAGroups2008_2013.csv. Research group identification results for "Science Map 2008~2013".

Author Contributions:
Wang Xiaomei: Conceived research ideas, designed research methodology, revised final manuscript.
Deng Qiping: Collected and cleaned data, implemented algorithms, drafted manuscript.

Received: 2015-11-12
Revised: 2016-02-22

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