Interdisciplinary Literature Discovery Based on the Rao-Stirling Index: A Case Study of Nanoscience and Nanotechnology (Postprint)

Authors: Han Zhengqi, Liu Xiaoping, Xu Han

Date: 2023-08-26T00:00:00+00:00

Abstract

[Purpose/Significance] The discovery of interdisciplinary literature constitutes an important prerequisite for conducting interdisciplinary research. Rapidly and accurately identifying domain-relevant interdisciplinary literature from vast amounts of publications enables researchers to comprehend interdisciplinary dynamics within the field and identify interdisciplinary research hotspots and frontiers. This study proposes a method for discovering interdisciplinary literature in a field based on the Rao-Stirling index, and explores the feasibility of this method using nanoscience and nanotechnology as a case study. [Method/Process] Literature in the field of nanoscience and nanotechnology was retrieved from the Web of Science database. A mapping table of journal abbreviations, full names, and subject categories was constructed. Python programming was employed to construct a subject distribution matrix of references for the literature. R programming was utilized to calculate the Rao-Stirling index for each document to measure the interdisciplinarity of the literature. Based on these measurements, literature in the field of nanoscience and nanotechnology was classified into three levels according to the degree of interdisciplinarity, thereby discovering interdisciplinary literature in the field. [Results/Conclusion] The method for discovering interdisciplinary literature based on the Rao-Stirling index enables measurement of interdisciplinarity at the literature level within a field and facilitates the discovery of interdisciplinary literature. This research method can also be extended to other research domains.

Full Text

Preamble

Interdisciplinary Literature Discovery Based on Rao-Stirling Indices: A Case Study in Nanoscience and Nanotechnology

Han Zhengqi[1],[2], Liu Xiaoping[2], Xu Han[1],[2]
[1] University of Chinese Academy of Sciences, Beijing 100049
[2] National Science Library, Chinese Academy of Sciences, Beijing 100190

Abstract

[**Purpose/Significance**] Discovering interdisciplinary literature is a crucial prerequisite for interdisciplinary research. Rapidly and accurately identifying domain-related interdisciplinary publications from massive literature collections helps researchers grasp interdisciplinary dynamics and recognize research hotspots and frontiers. This study proposes a method for discovering interdisciplinary literature based on the Rao-Stirling index and uses nanoscience and nanotechnology as a case study to explore the feasibility of this approach. [**Method/Process**] We downloaded nanoscience and nanotechnology publications from the Web of Science database, constructed a mapping table of journal abbreviations to full names and subject categories, used Python to build a reference subject distribution matrix for the literature, and employed R programming to calculate the Rao-Stirling index for each document to measure its interdisciplinary degree. Based on these measurements, we classified nanoscience and nanotechnology literature into three levels according to their degree of interdisciplinarity. [**Result/Conclusion**] The proposed method can effectively measure the interdisciplinary degree at the document level and identify interdisciplinary literature. This approach is also extensible to other research fields.

Keywords: Interdisciplinarity; Rao-Stirling; Interdisciplinary Literature; Nanoscience and Nanotechnology

---

1. Introduction

As scientific research continues to deepen, disciplines have undergone continuous differentiation, subdividing into numerous sub-disciplines and developing toward greater specialization. Simultaneously, however, disciplines are gradually moving toward integration, developing in the direction of high-level synthesis, holism, and cross-fertilization. Absorbing knowledge from different disciplinary fields is considered a key factor in scientific development. Major social problems and complex scientific issues often require collaboration across multiple disciplines through intersection, penetration, and fusion to achieve solutions.

Among the 365 natural science awards granted between 1900 and 2011, 198 (54.2%) were in interdisciplinary fields, showing a steady upward trend. Most of the "Top Ten Scientific Breakthroughs" selected annually by *Science* magazine also represent achievements from interdisciplinary research. Many important scientific discoveries, such as DNA' s double helix structure, genetic recombination technology, and magnetic resonance imaging, have emerged from breaking disciplinary boundaries. As interdisciplinary research plays an increasingly important role in scientific and technological progress, examining it from different perspectives becomes essential.

Interdisciplinary research (IDR), also known as cross-disciplinary research, involves integrating knowledge across disciplines. It often signifies innovation and progress, with interdisciplinary intersections frequently becoming new research hotspots and frontiers. As physicist Werner Heisenberg noted, "The most important scientific discoveries often occur at the intersection of two different lines of thought." Identifying relevant interdisciplinary intersections within a field, revealing interdisciplinary relationships and characteristics, and understanding hotspot directions and trends have become important branches of interdisciplinary research. Such work can provide researchers with theoretical insights and hypotheses to help them choose appropriate research directions, support university curriculum adjustments and the establishment of interdisciplinary research centers, and inform scientific managers' project evaluations and policy-making by government agencies.

Interdisciplinary research can address questions in unknown domains that traditional disciplines cannot reach. These unknown areas often become sources of major academic breakthroughs when traditional disciplines have matured. Interdisciplinary research represents an effective approach to solving complex and challenging scientific problems.

---

2. Literature Review

Research on interdisciplinarity includes theoretical and empirical studies. Theoretical research primarily explores the origins, evolution, formation, and development mechanisms of interdisciplinary phenomena, while empirical research mainly uses bibliometric or text mining methods to discover interdisciplinary characteristics of specific journals, institutions, or disciplines. This paper focuses on the empirical research component.

Based on different research objects, interdisciplinary literature discovery studies can be divided into four categories: keyword-based, document-based, author-based, and citation-based approaches.

Keyword-based approaches treat paper keywords as analysis units, counting keyword co-occurrence frequencies. Higher co-occurrence frequencies indicate stronger associations between the disciplinary fields represented by the

keywords and greater interdisciplinarity. For example, Rip et al. used co-word analysis to study dynamic changes in the biotechnology field over a decade, exploring how connections between research areas evolved over time. Wei Jianxiang used CNKI database records from core journals in information science and computer science, employing fuzzy C-means clustering (FCM) to discover interdisciplinary literature and IS-PSO algorithms for knowledge discovery and visualization. Li Changling et al. used time series analysis to dynamically display changes in interdisciplinary research themes between information science and computer science. However, keyword-based approaches typically use text clustering methods, whose algorithmic complexity increases dramatically with dimensions, making them unsuitable for large-scale datasets and limiting their ability to identify semantic relationships between terms.

Document-based approaches rely on co-classification analysis, assigning documents or journals to categories according to existing classification standards (Dewey Decimal Classification, Web of Science subject categories, etc.). Interdisciplinary characteristics are reflected by counting the number of different subject categories to which documents or journals belong. For instance, Wu Lei et al. used Web of Science subject classifications to identify interdisciplinary literature in animal resources and breeding. Allan used Dewey Decimal Classification of research outputs to determine interdisciplinary activities at a university. Tijssen used co-classification analysis to examine interdisciplinary structures in energy research. While suitable for large-scale, multi-domain studies, this approach suffers from database management lags and cannot be applied to recent research. Additionally, co-classification analysis works well for fields with clear disciplinary boundaries but less effectively for those with ambiguous boundaries.

Author-based approaches assess interdisciplinarity by counting the number of disciplines represented by authors' publications or co-authors. Qin et al. analyzed research collaboration by examining the number and types of authors, institutions, and disciplines. Schummer identified interdisciplinary literature by determining researchers' disciplines through their institutional affiliations. While this approach overcomes classification limitations and reveals more comprehensive cognitive and social phenomena, it relies heavily on expert judgment and manual work to assign authors to disciplines, making it practical only for small-scale studies.

Citation-based approaches are the most common method for measuring interdisciplinarity, operating on the principle that the number of different disciplines represented in a paper' s citations indicates its interdisciplinary degree. A paper' s cited disciplines can be represented by the subject categories of the journals in which cited articles appear. Chubin et al. used citation category analysis to distinguish interdisciplinary from traditional single-discipline research. Porter et al. proposed the "out-category citation" method to measure interdisciplinarity in demography, operations research/management, and toxicology. Choi examined communication patterns within and across disciplines in astronomy. Wei Haiyan et al. used citation analysis to study information

science and related disciplines. Citation analysis can fundamentally identify interdisciplinary literature by revealing which disciplines inspire researchers, thus better capturing interdisciplinary development trajectories.

Current empirical studies either examine interdisciplinarity at the document or journal level, or investigate cross-disciplinary collaboration at the institutional level. For field-level interdisciplinary research, existing studies have only examined macro-level cross-domain relationships or identified interdisciplinary themes between two specific fields, without identifying all interdisciplinary topics related to a given field. The key challenge is how to discover all interdisciplinary literature related to a field from massive document collections and filter highly interdisciplinary documents for subsequent theme identification.

---

3. Research Methods

This study employs citation analysis for interdisciplinary literature discovery. Metrics used to measure interdisciplinary diversity typically originate from other fields, particularly ecology. Table 1 summarizes common indicators.

Table 1. Common Indicators for Interdisciplinary Research Based on Citation Analysis

Indicator	Definition/Description
Variety	Quantifies the number of disciplinary categories in references: v_{max} = maximum number of WoS subject categories among references
Balance	Describes the evenness of subject distribution: $Balance = \frac{-\sum p_i \ln p_i}{\ln(v)}$
Disparity	Measures the distinctiveness of element distribution: $Disparity = \frac{\sum_{i,j} d_{ij}}{v(v-1)}$
Rao-Stirling Diversity	Integrates three dimensions: $Rao{-}Stirling\ Diversity = \sum_{i,j} p_i p_j d_{ij}$

Where v_{max} is the maximum number of subject categories among references in the dataset, p_i is the proportion of references belonging to subject category i, and d_{ij} is the distance between subject categories i and j.

Variety represents the number of subject categories to which references can be assigned—more categories indicate stronger interdisciplinarity. **Balance** describes the evenness of reference distribution across subjects, using Shannon' s evenness index. If a paper cites equal numbers of references from each category, its distribution is perfectly balanced, indicating stronger interdisciplinarity. **Disparity** measures the distinctiveness of cited disciplines—if references

belong to similar subjects (e.g., physical chemistry and applied physics), disparity is low and interdisciplinarity is weak; if subjects are dissimilar (e.g., zoology and applied physics), disparity is high and interdisciplinarity is strong.

The first three formulas represent three different aspects of diversity. The Rao-Stirling index integrates these dimensions—variety, balance, and disparity—into a single measure.

First-generation biodiversity metrics (distribution-sensitive) include the simplest diversity measure (counting different disciplines) and more sophisticated indices like entropy and Gini-Simpson index, which measure uncertainty and probability of belonging to different categories. However, these metrics assume all subject categories are completely different, ignoring similarities between them. If two papers cite the same number of subject categories, they are considered equally interdisciplinary regardless of how similar those subjects are.

Second-generation biodiversity metrics (distribution- and similarity-sensitive) address this limitation. The Rao-Stirling index, proposed by Andy Stirling, considers both similarity and dissimilarity between disciplines. The formula is:

$$Rao-Stirling\ Diversity = \sum_{i,j} p_i p_j d_{ij}$$

This study uses nanoscience and nanotechnology as a case study, calculating the Rao-Stirling index based on the subject distribution of cited journals to measure document-level interdisciplinarity and identify interdisciplinary literature in the field.

---

4. Data Sources and Processing

4.1 Sample Data Sources

We selected nanoscience and nanotechnology publications from Thomson Reuters' Web of Science (WoS) SCI-EXPANDED database as our research object. The search query was WC=("Nanoscience & Nanotechnology"), limited to document types ARTICLE, PROCEEDINGS PAPER, and REVIEW, published in 2016, retrieved on March 31, 2017, yielding 40,618 records. We used Leydesdorff' s ISI.exe program to extract bibliographic information and generate a relational database.

4.2 Constructing Journal Abbreviation-Full Name-Subject Category Mapping Table

Since downloaded citation information only included journal names without subject categories, we needed to construct a journal-subject category mapping table. WoS staff assign subject categories to each journal based on multiple indicators.

The Journal Citation Reports (JCR) provides subject category information for each journal. We downloaded JCR' s journal and subject category data and built a mapping table using Python.

Because JCR uses full journal names while WoS reference data uses abbreviated names, we also needed a full name-abbreviation mapping table. The ESI Journal List provides both formats. By linking the JCR full name-subject table with the ESI full name-abbreviation table, we constructed a comprehensive mapping table covering 11,375 journals, yielding 17,961 journal-subject mappings (average 1.58 subjects per journal).

4.3 Mapping References to Subjects

After cleaning the data by removing records with incomplete information, we obtained 40,491 documents with 1,649,353 references (average 40.73 references per paper). Using MySQL and our mapping table, we assigned each reference to one or more WoS subject categories based on its journal. This successfully mapped 1,461,341 references (88.60% success rate), with each reference journal corresponding to an average of 2.24 subject categories.

To ensure reliable measurements, we excluded documents with fewer than 3 subject categories in their references, leaving 38,598 documents (95.32% of the downloaded collection) for analysis.

---

5. Interdisciplinarity in Nanoscience and Nanotechnology

5.1 Constructing Reference Subject Category Matrix

Using Python, we converted the reference subject data into a matrix format (Table 4), where each row represents a document' s reference subject distribution and each column shows the number of references belonging to a particular subject category.

Table 4. Reference Subject Category Matrix (Partial)

Document ID	Physics, Atomic, Molecular & Chemical	Electrical & Electronic Engineering	···
1	15	8	···
2	25	12	···

5.2 Identifying Interdisciplinary Literature

In the Rao-Stirling formula, d_{ij} represents the distance between subject categories i and j. For reliability and comparability, we used the WoS subject similarity matrix from Chavarro et al., based on Rafols et al.' s method using

60,947,519 documents from the 2007 JCR to construct a co-citation matrix and calculate cosine similarities between subjects.

We wrote R programs to measure each document' s variety, balance, disparity, and Rao-Stirling diversity index. Table 5 shows partial results sorted by Rao-Stirling index (higher values indicate greater interdisciplinarity).

Table 5. Interdisciplinarity Measurement Results for Nanoscience and Nanotechnology Literature

Document ID	Title	Rao-Stirling	Variety	Balance	Disparity
29215	Qualitative Data Analysis in Product Development···	0.99	15	0.87	0.96
8050	Settling into the mainstream? Lessons for governance ···	0.92	25	0.84	0.93
...	...	...	...	...	...

Figure 1 [Figure 1: see original paper] shows the frequency distribution histogram of Rao-Stirling indices. Most values cluster between 0.2 and 0.5, allowing us to classify literature into three levels: low interdisciplinarity (<0.2), moderate (0.2-0.5), and high (>0.5). Researchers can set different thresholds based on their needs to filter interdisciplinary literature for further theme identification.

6. Discussion and Conclusion

Interdisciplinary literature discovery is essential for interdisciplinary research. Our citation analysis approach using the Rao-Stirling index offers several advantages:

1. **Comprehensive discovery**: It can identify all interdisciplinary literature related to a field with relatively low computational complexity, making it suitable for large datasets.
2. **Source-level analysis**: By measuring the subject distribution of citations, it reveals which disciplines inspire researchers, better capturing interdisciplinary development pathways.
3. **Quantitative measurement**: It enables quantitative assessment of interdisciplinarity, allowing selection of highly interdisciplinary documents. Highly interdisciplinary papers often correlate with high citation counts, representing current research hotspots and frontiers.

The method and tools developed (mapping tables, Python and R programs) are transferable to other fields. The identified highly interdisciplinary nanoscience literature can serve as a foundation for subsequent theme identification research.

Limitations and future work: 1. The study uses WoS subject categories; different classification systems may yield different results. 2. The epistemological validity of citation-based interdisciplinary identification requires further verification. 3. No standard threshold exists for defining interdisciplinary or highly interdisciplinary literature. While documents with Rao-Stirling $ $0.5 appear highly interdisciplinary upon review, expert validation is needed.

Future research will compare multiple classification systems, apply expert judgment to nanoscience results for theme identification, and extend the method to other disciplines.

References

[1] QIN J, LANCASTER F W, ALLEN B. Types and levels of collaboration in interdisciplinary research in the sciences[J]. Journal of the American Society for Information Science, 1997, 48(10): 893-916. [2] Lu Yongxiang. The significance of interdisciplinary and cross-disciplinary science[J]. Bulletin of Chinese Academy of Sciences, 2005, 20(1): 58-60. [3] Hui Sen. Research on interdisciplinary phenomena in Nobel natural science awards[D]. Zhengzhou: Zhengzhou University, 2014. [4] RIP A, COURIAL J P. Co-word maps of biotechnology: an example of cognitive scientometrics[J]. Scientometrics, 1984, 6(6): 381-400. [5] Wei Jianxiang. Research on interdisciplinary knowledge discovery and visualization[D]. Nanjing: Nanjing University, 2010. [6] Li Changling, Guo Fengjiao, Wei Xuqiu. Analysis of interdisciplinary research themes based on temporal keywords: a case study of information science and computer science[J]. Information and Documentation Services, 2014(6): 44-48. [7] Wu Lei, Sun Wei. Research on methods for discovering and evolving interdisciplinary hotspot themes: a case study of animal resources and breeding[J]. Digital Library Forum, 2015(12): 15-20. [8] ALLAN A. A method for determining interdisciplinary activities within a university[J]. Library research, 1980, 2(1): 83-94. [9] TIJSSEN R J W. A quantitative assessment of interdisciplinary structures

in science and technology: co-classification analysis of energy research[J]. Research policy, 1992, 21(1): 27-44. [10] SCHUMMER J. Multidisciplinarity, interdisciplinarity, and patterns of research collaboration in nanoscience and nanotechnology[J]. Scientometrics, 2004, 59(3): 425-465. [11] CHUBIN D E, PORTER A L, ROSSINI F A. "Citation classics" analysis: an approach to characterizing interdisciplinary research[J]. Journal of the American Society for Information Science, 1984, 35(6): 360-368. [12] PORTER A L, CHUBIN D E. An indicator of cross-disciplinary research[J]. Scientometrics, 1985, 8(3/4): 161-176. [13] CHOI J M. Citation analysis of intra- and interdisciplinary communication patterns of anthropology in the U.S.A.[J]. Behavioral & social sciences librarian, 1988, 6(3-4): 65-84. [14] Wei Haiyan, Yin Huaiqiong, Liu Li. Research on information science and related disciplines based on citation analysis[J]. Journal of Intelligence, 2010, 29(2): 38-43. [15] Yang Liangbin, Zhou Qiuju, Jin Bihui. Interdisciplinary measurement based on bibliometrics and empirical research[J]. Library and Information Service, 2009, 53(10): 87-90. [16] Guo Fengjiao, Li Changling, Wang Xiaodi, et al. Interdisciplinary literature discovery based on vector space model: a case study of information science and computer science[J]. Information Studies: Theory & Application, 2014(3): 171-175. [17] MUGABUSHAKA A M, KYRIAKOU A, PAPAZOGLOU T. Bibliometric indicators of interdisciplinarity: the potential of the Leinster-Cobbold diversity indices to study disciplinary diversity[J]. Scientometrics, 2016, 107(2): 593-607. [18] STIRLING A. A general framework for analysing diversity in science, technology and society[J]. Journal of the Royal Society Interface, 2007, 4(15): 707-719. [19] LEYDESDORFF L. isi.exe[EB/OL]. [2017-07-04]. http://www.leydesdorff.net/software/isi/index.htm. [20] Incites Journal Citation Reports[EB/OL]. [2017-07-04]. https://jcr.incites.thomsonreuters.com/JCRJournalHomeAction.action. [21] Incites Essential Science Indicators help[EB/OL]. [2017-07-04]. http://ipscience-help.thomsonreuters.com/incitesLiveESI/ESIGroup/overviewESI/esiJournalsList.html. [22] CHAVARRO D, TANG P, RAFOLS I. Interdisciplinarity and research on local issues: evidence from a developing country[J]. Research evaluation, 2014, 23(3): 195-209. [23] RAFOLS I, PORTER A L, LEYDESDORFF L. Science overlay maps: a new tool for research policy and library management[J]. Journal of the Association for Information Science & Technology, 2010, 61(9): 1871-1887. [24] CHEN S, ARSENAULT C, LARIVIÈRE V. Are top-cited papers more interdisciplinary?[J]. Journal of informetrics, 2015, 9(4): 1034-1046.

Author Contributions

Liu Xiaoping: Conceived the research direction; provided guidance and revision for paper writing.
Han Zhengqi: Performed data processing and analysis; drafted the manuscript.
Xu Han: Collected data.