https://doi.org/10.51408/issi2025_101

Research on the Measurement Method of Disciplinary Diversity Based on Lexical Semantic Analysis

Guo Chen[1], Yifan Yang[2]

[1] *delphi1987@qq.com*, [2]*1005104368@qq.com*
NJUST Nanjing University of Science and Technology, No. 200 Xiao Ling Wei, Nanjing, Jiangsu (China)

Abstract

Existing methods for measuring disciplinary diversity mainly focus on literature (such as citations) as the unit of analysis. This paper proposes a new approach to measuring disciplinary diversity at a fine-grained level based on lexical semantics. Taking articles from the OpenAlex dataset between 2014 and 2023 as an example, the breadth of concept distribution is calculated within the semantic space of given disciplinary vocabulary to measure disciplinary richness; the external word frequency ratio and similarity of high-frequency disciplinary vocabulary are integrated to calculate the concept overflow degree, thereby measuring the degree of disciplinary intersection. Based on this, a two-dimensional matrix is constructed to locate types of disciplinary diversity and further analyze the temporal trends and causes of diversity in various disciplines. According to disciplinary richness and intersection, 19 first-level disciplines are categorized into four major types: Diverse Integration, Deep Specialization, Broad Interaction, and Single Cohesion, and the classification results are analyzed. Additionally, the trends and causes of changes in richness and intersection at both macro and micro levels are analyzed for each discipline. This study proposes a more fine-grained disciplinary diversity measurement method at the lexical semantic level, providing a new and broader perspective for the study of disciplinary diversity.

Introduction

Traditional methods for measuring disciplinary diversity primarily use literature as the basic unit of analysis, and there is still room for refinement from a fundamental granularity perspective. Words are the fundamental units of knowledge expression, and using their semantics can provide a deeper understanding of the structure and differences in human knowledge content. In psychology, researchers have begun to use word semantics to conduct cognitive experiments. For example, Olson et al. (2021) used cosine distance to calculate the pairwise semantic distances between 10 nouns to measure human divergent thinking, finding it more effective than traditional alternative uses tasks and bridging associative gap tasks. Their findings, published in Nature, have garnered widespread attention. This has inspired many scholars to conduct related work. Hubert et al. (2024) also used the same method to measure the degree of human thinking divergence. Similarly, in addition to measuring human creativity and divergent thinking, word semantics can also be used to measure differences in knowledge. Hur (2024) introduced semantic heterogeneity based on word embedding techniques in content analysis when calculating the diversity of patent entities, representing diversity through the semantic distance between patent entities. Lix et al. (2022) used word semantics to calculate the diversity of team discourse, a concept of fine-grained knowledge participation that is difficult to track

with previous text analysis methods. Thus, it appears possible to use word semantics to reveal disciplinary diversity, but similar research has not yet been conducted.

Words are the most basic units for representing semantics, and in the process of inheriting, communicating, and diffusing scientific knowledge, the finest granularity unit is the conceptual knowledge described by words. Therefore, measuring disciplinary diversity from the perspective of the aggregation and intersection of word semantics is both a natural and inevitable requirement. Based on this, this paper takes word semantics as the starting point, utilizes word semantic representation and deep learning techniques, and analyzes disciplinary diversity from a finer-grained lexical level. It comprehensively considers word frequency and semantic relationships between words, quantifying disciplinary diversity from two dimensions: disciplinary richness and disciplinary intersection. The two dimensions are combined to classify types of disciplinary diversity. In the experimental section, a semantic space for 19 first-level disciplines is constructed using the open-source OpenAlex data, and the proposed method is applied to classify diversity types and analyze time series trends. The empirical results demonstrate that this method can effectively analyze the development characteristics and changes in the degree of intersection of different disciplines, providing a novel perspective and approach for disciplinary evaluation and prediction research.

Data and methods

We used the paper data from OpenAlex between 2014 and 2023 as the experimental subjects, obtaining a total of 72 million records. First, we classified the major disciplines based on the fos (field of study) field in the paper data. If a paper's fos field contains multiple disciplines, it is included in multiple major disciplines. According to the Microsoft discipline classification, there are 19 first-level disciplines. Each discipline is divided into subsets based on the year, resulting in a total of 190 subsets.

The text content undergoes stemming and keyword matching, and the Word2Vec model is trained using incremental learning. The frequency of a word's appearance in different disciplines is used to determine whether it is a discipline-specific term. In this paper, words that appear fewer than nine times are designated as discipline-specific terms for use in subsequent metric calculations.

Currently, the measurement of disciplinary diversity is typically focused at the literature level, resulting in a relatively coarse research granularity that fails to capture subtle semantic changes. However, more fine-grained lexical semantic analysis has been successfully applied to measure the degree of individual divergent thinking and team diversity, indicating that lexical semantic analysis has a solid foundation for representing diversity. Lexical items are the most basic units for representing disciplinary knowledge, and semantic changes can directly explain the development and evolution of disciplinary knowledge. The broader the distribution of vocabulary in a semantic space within a discipline, the richer the disciplinary knowledge is. Therefore, this study employs lexical semantics to measure disciplinary diversity from two key dimensions: the richness within disciplines and the intersection between disciplines. From a semantic perspective, disciplinary

richness can be represented by the average distance between high-frequency words; the greater the average distance, the higher the internal richness of the discipline. Intersection can be represented by the degree of overlap in semantic space; the greater the overlap, the higher the external intersection between disciplines.

Measurement of disciplinary richness

We can measure the average distance between each word and other words to obtain the average distance of elements within the semantic space (or the distance between each word and the document centroid), which can be used to measure the conceptual breadth within that semantic space. Let N be the total number of high-frequency words, v_i and v_j be the word vectors obtained through word embedding, and f_i and $f_j$be the word frequencies.

$$2 \times \frac{\sum_{k=1}^{N} \sum_{i \neq j} \frac{distance\,(v_i \times f_i, v_j \times f_j)}{f_i + f_j}}{N(N-1)}$$

Measurement of interdisciplinary

Combining the word similarity calculation metric and the True Diversity metric (Zhang, L et al., 2016), we have proposed a disciplinary intersection metric based on high-frequency word calculations.For the calculation of disciplinary intersection, let n be the total number of fields, and N be the total number of high-frequency words,. For two different fields i and j, w_{ki} and w_{kj}represent the same words appearing in different fields. p_{ki} and $p_{kj}$are the proportions of the words w_{ki} and w_{kj}in the high-frequency word set N of fields i and j, respectively.

$$\frac{\sum_{k=1}^{N} \sum_{j \neq i}^{n} \mathrm{Cos}\left(w_{ki}, w_{kj}\right) p_{ki}\, p_{kj}}{n-1}$$

Building on the aforementioned approach, lexical semantic calculations can be used to determine the richness within disciplines and the intersection between disciplines. These two metrics can be employed for both two-dimensional matrix analysis and time-series trend analysis. First, a two-dimensional matrix can be used to categorize disciplines into four types, and the possible reasons for these classifications can be analyzed. On the other hand, time-series trend analysis can be conducted to examine the changes in disciplinary richness and intersection over time, and further analysis can be performed from a lexical perspective to understand the reasons for these changes.

Results

Analysis of Lexical Semantic Dimensionality Reduction Visualization Results

To more intuitively observe the distribution of word vectors, this paper employs the UMAP dimensionality reduction algorithm to visualize the distribution of vocabulary from various disciplines, as shown in Figure 1 below.

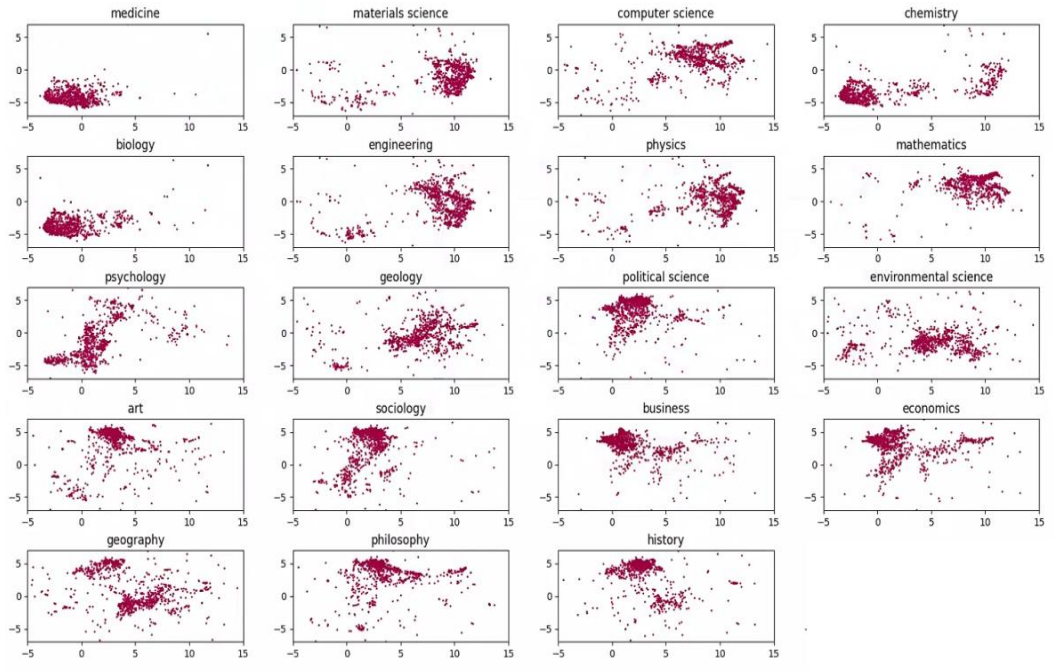

Figure 1. Semantic distribution map of various disciplines.

The vocabulary from the disciplines of medicine, chemistry, and biology exhibit a clear clustering trend in the semantic space, primarily concentrating in the lower left corner of the space. Computer science and mathematics form another concentrated area in the upper right corner. The close connection between these two disciplines may stem from their shared reliance on algorithmic thinking, logical reasoning, and theoretical modeling. Materials science, engineering, and physics are concentrated in the lower right corner of the space. This phenomenon is related to the technical and engineering methods these disciplines employ in solving practical problems.

On the other hand, the vocabulary from political science, art, sociology, business, economics, philosophy, and history is concentrated in the upper left corner of the space. These disciplines focus more on human society, culture, economy, and political phenomena, and they may have more intersections in research methods and theoretical frameworks, such as qualitative analysis, historical comparison, and critical thinking, leading to the formation of a relatively independent cluster in the semantic space.

The vocabulary from psychology, geography, environmental science, and geology is concentrated in the central region of the space. These disciplines all focus to some extent on the interaction between human activities and the natural environment. They may share common research methods and focal points in data collection, spatial analysis, and environmental monitoring, thus forming a central interdisciplinary cluster in the semantic space.

Identification of Disciplinary Diversity Types by Integrating Richness and Intersection

The results of the metrics for 19 disciplines over a 10-year period were combined and analyzed. The mean values of disciplinary richness and intersection were used as the origin, with different point shapes representing different disciplines. The horizontal axis represents disciplinary richness, and the vertical axis represents disciplinary intersection, as shown in Figure 2 below.

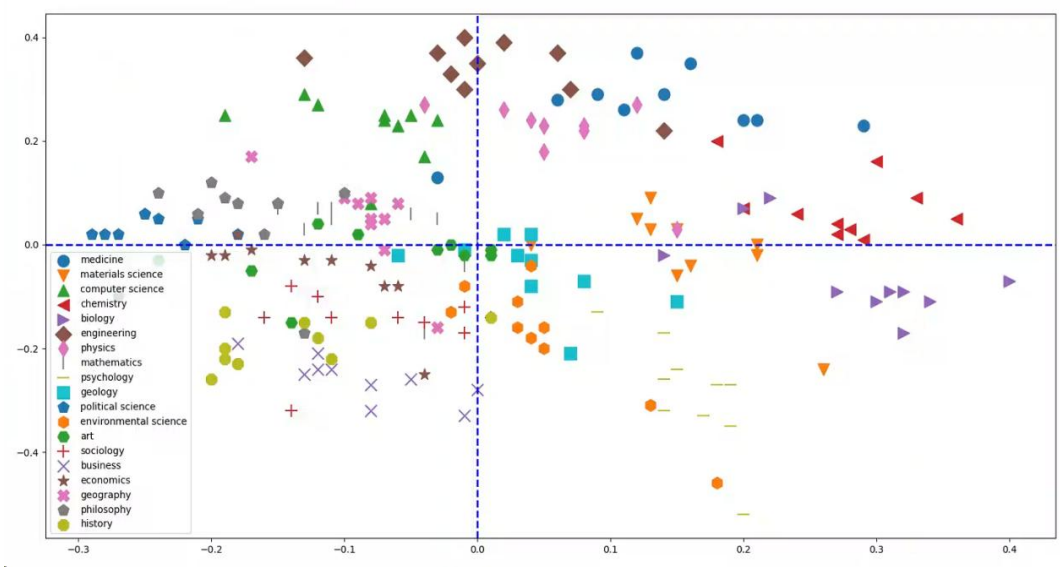

Figure 2. Classification of disciplinary diversity types.

Based on the situation in Figure 2, all points were divided into four regions according to the natural boundaries where disciplinary richness equals zero and disciplinary intersection equals zero. The division results are presented in Table 1.

Table 1. Classification results of disciplinary diversity types that integrate richness and intersectionality.

Table	*Low disciplinary richness*	*High disciplinary richness*
High interdisciplinary degree	Computer science Engineering Geography Mathematics Philosophy	Medicine Material science Chemistry Physics
Low interdisciplinary degree	History Business Political science Art Sociology Economics	Environment science Geology Psychology Biology

Analysis of Temporal Trends in Disciplinary Richness

The trends in disciplinary richness are shown in Figure 3, with high richness and high intersection in red, high richness and low intersection in green, low richness and high intersection in yellow, and low richness and low intersection in blue. Overall, the richness of most disciplines is declining, such as computer science, chemistry, and biology, while a few disciplines are experiencing an increase in richness, such as art, history, and sociology. The decline in richness for most disciplines reflects a trend towards specialization and concentration.

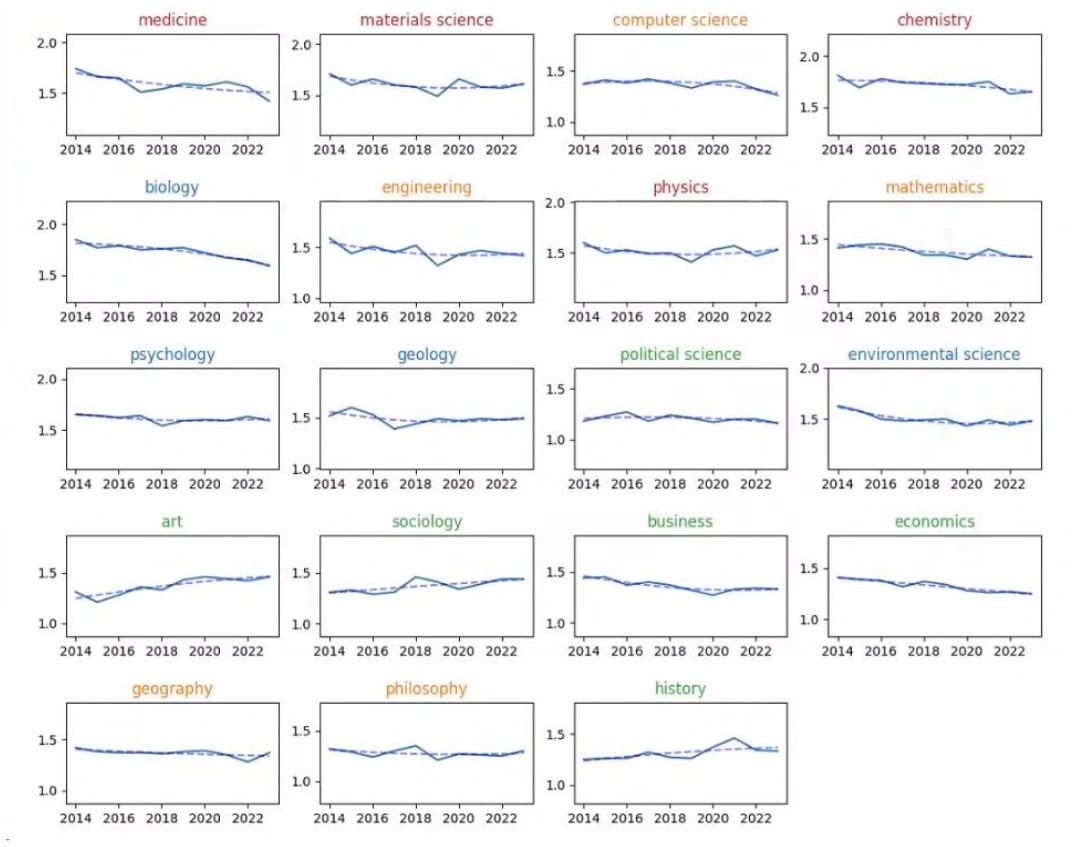

Figure 3. Time series trend chart of richness changes in various disciplines.

Analysis of Temporal Trends in Disciplinary Intersection

The trends in the intersection of various disciplines are shown in Figure 4. There is an increase in the degree of intersection for all disciplines to varying extents, reflecting a growing trend of interdisciplinary integration. As complex problems emerge, different fields begin to collaborate, sharing knowledge and technology to promote innovation and solve practical issues. This trend also reflects an increased demand for comprehensive research, leading to the gradual blurring of disciplinary boundaries and fostering the emergence of new research methods and fields.

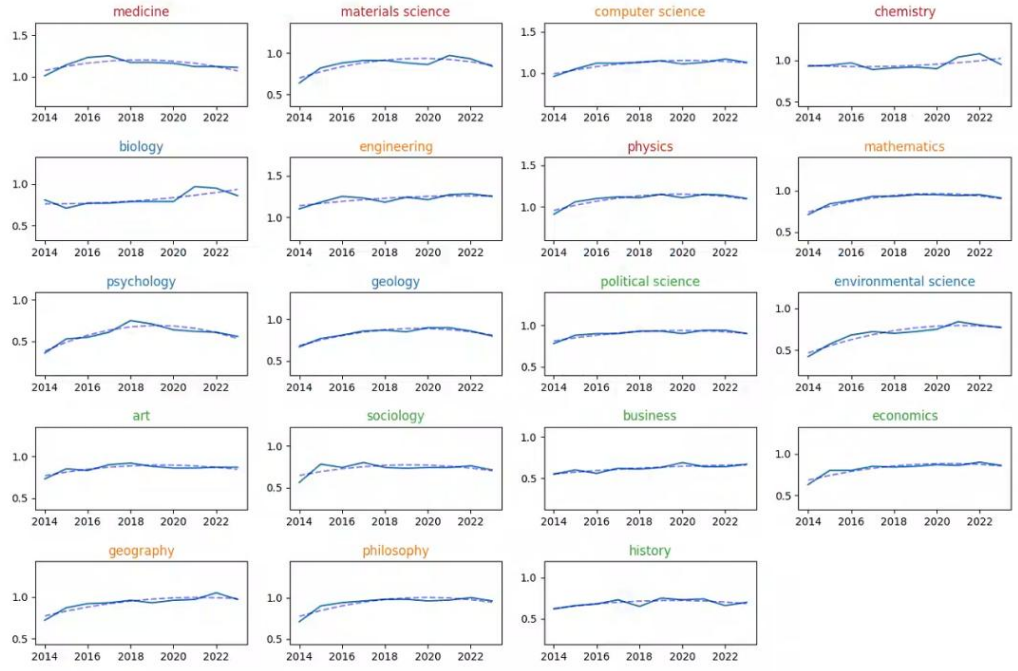

Figure 4. Trend Chart of Temporal Changes in Interdisciplinary Intersectionality among Various Disciplines.

Discussion

This paper proposes a new method for measuring disciplinary diversity based on lexical semantic analysis. Through an empirical study of articles from the OpenAlex dataset between 2014 and 2023, the effectiveness and feasibility of this method have been validated. The results indicate that this method can accurately quantify disciplinary richness and intersection from a finer-grained lexical semantic perspective, providing a new perspective for the classification and temporal change analysis of disciplinary diversity.

Despite the achievements of this study, there are some limitations. First, lexical semantic analysis relies on the quality of word embedding models and the comprehensiveness of the corpus. Imbalances in corpora across different disciplines may affect the accuracy of the measurement results. Second, this paper primarily focuses on two dimensions: disciplinary richness and intersection. Future research could consider incorporating additional dimensions, such as disciplinary balance and innovativeness, to more comprehensively reflect disciplinary diversity.

References

Hubert, K. F., Awa, K. N., & Zabelina, D. L. (2024). The current state of artificial intelligence generative language models is more creative than humans on divergent thinking tasks. *Scientific Reports*, *14*(1), 3440.

Hur, W. (2024). Entropy, heterogeneity, and their impact on technology progress. *Journal of Informetrics*, *18*(2), 101506.

Lix, K., Goldberg, A., Srivastava, S. B., & Valentine, M. A. (2022). Aligning differences: Discursive diversity and team performance. *Management Science*, *68*(11), 8430-8448.

Olson, J. A., Nahas, J., Chmoulevitch, D., Cropper, S. J., & Webb, M. E. (2021). Naming unrelated words predicts creativity. *Proceedings of the National Academy of Sciences*, *118*(25), e2022340118.

Zhang, L., Rousseau, R., & Glänzel, W. (2016). Diversity of references as an indicator of the interdisciplinarity of journals: Taking similarity between subject fields into account. *Journal of the association for information science and technology*, *67*(5), 1257-1265.