DSI DATA SCIENCE AND INFORMETRICS
VOL.2, NO.3, AUG. 2022

Detecting the research diversity of researchers in library and information science: An exploratory study

Yuehua Zhao[a,b], Sicheng Zhu[b], Jie Wu[b], Hao Wang[a,b], Sanhong Deng[a,b]

a. Jiangsu Key Laboratory of Data Engineering and Knowledge Service, Nanjing University, Nanjing, China

b. School of Information Management, Nanjing University, Nanjing, China

ABSTRACT

Metrics have emerged as an important tool for quantitatively evaluating researchers from a variety of perspectives, including research impact, research quality, interdisciplinarity, and cross-disciplinarity. Especially in the field of library and information science, many previous studies have highlighted the characteristics of researchers in this field. However, only a minority of the studies address the aspect of diversity in research topics. The purpose of this study is to (1) evaluate the topic diversity of researchers in library and information science and (2) examine the relationships between the researcher topic diversity and research impact. We propose an indicator to quantify author topic diversity, which we refer to as author topic diversity (ATD). Latent Dirichlet Allocation (LDA) is used to detect topics in the field, while cosine similarity is used to calculate the diversity of research topics in a given researcher's publications. The results show that topic diversity in the field of library and information science varies greatly from author to author. In addition, weak positive correlations are found between the ATD and citation indicators, suggesting that engaging in diversified topics may lead to higher research impact.

KEYWORDS

Research diversity; Research evaluation; Library and information science

1 Introduction

Advancing professional development requires a steady stream of publications, often achieved through continuous additions to an existing research agenda (Jia et al., 2017). Despite widespread criticism of the usefulness of evaluating researchers based on research outputs, publication production and the number of citations remain the most commonly used metrics for career advancement and funding applications. However, creative publications tend to have a greater impact than conservative ones (Foster et al., 2015). Sociologists have suggested that the mechanisms that govern scientists' selection of research topics may be the result of a trade-off between conservative production and risky innovation (Bourdieu, 1975). Given the contemporary understanding that interdisciplinary communication and collaboration are necessary not only for the curiosity-driven pursuit of fundamental knowledge but also for addressing complicated socioeconomic challenges (Okamura, 2019). It would be better to clarify whether participation in diversified research carries the risk of failure or the opportunity for a breakthrough.

Diversity was first proposed by Stirling (2007) and Rafols and Meyer (2010). Early diversity is an ecological concept that mainly considers the number of categories, the relative number of elements in each category, and the difference or similarity between categories. Efforts to quantitatively measure the diversity of research have been made by assessing the interdisciplinary degree of research. Several attempts have been made to develop methods to define the interdisciplinarity index. For a more sophisticated quantitative approach to interdisciplinarity conceptualized as disciplinary diversity, the following three characteristics are required: 'variety' (number of disciplines included), 'balance' (evenness of distribution among disciplines), and 'dissimilarity' (degree of dissimilarity between the disciplines) (Okamura, 2019; Yegros-Yegros et al., 2015).

Despite extensive research on interdisciplinarity which has been popular in scientometrics, quantitative assessments of the research diversity of individual scientists within a given discipline remain limited. This study aims to bridge this potential research gap by developing a measure of the diversity of scholar's research topics. Indicators capturing the degree of diversity in interdisciplinary studies (i.e., disciplinary diversity) rely on established disciplinary classifications, where diversity refers to the number of disciplinary categories. Balance refers to the evenness of the distribution of disciplines, and disparity refers to the extent to which these disciplines differ or resemble each other from a cognitive perspective (Yegros-Yegros et al., 2015). Here, we focus on measuring the diversity of topics within a given discipline rather than across disciplines. Therefore, we propose to classify the topics in a given discipline using topic modeling techniques. As part of an exploratory study, we examined researchers in the field of library and information science. The research questions we aim to answer in this study are as follows: (1) What is the topic diversity of researchers in library and information science?, and (2) How is the topic diversity of researchers related to their influence on research?

2 Related work

2.1 Research evaluation

Currently, the evaluation of scholars is mainly based on publications and citations. Citation-based evaluation can be divided into three aspects: the traditional evaluation index, the h-index and its derivative index, and the journal evaluation index.

Publication-based evaluation of scholars is mainly based on the number of published articles, which is the most traditional method of evaluation by scholars. The number of published articles refers to the total number of articles published by a given author in a given period. The productivity of scientific research has been measured by the number of articles published by scholars. It is the most basic link in the evaluation system of scholars, but the manner of weighting and quality makes it useless for evaluating the scientific influence of scholars (Gao & Zhang, 2016). Scholars are always developing new indicators based on the number of published articles to make a more accurate and fair evaluation of scholars.

The citation-based evaluation system of scholars is huge. In 1955, Eugene Garfield proposed the use of citations to evaluate scholars' scientific research. He believed that the frequency of citations represented the degree of recognition in the literature. This study opened the door to quality rather than quantity. However, the simple calculation of citation frequency leads to exaggerating the author's influence due to the small number of co-authored papers (Gao & Zhang, 2016). For this reason, Schubert and Glänzel (2006) formally proposed the c/p index, which combines the number of papers with the number of citations, i.e., the citation frequency of each article. This research provides a method for evaluating the influence of scholars in different eras. However, it is difficult to fairly evaluate authors who have published many articles, but only a few outstanding articles that have made a significant contribution to the science. To reduce the influence of the number of publications, the number of important papers and the number of citations of important papers became new evaluation indicators. Both of them prevent the accumulation of the number of publications from affecting the results. However, the number of important publications is subjectively defined by the researcher and lacks a certain degree of objectivity (Gao & Zhang, 2016).

Since the H-index in 2005, it has become a new kind of academic evaluation index, and a series of extended indices have been derived. Hirsch (2005) combined the number of scholars' publications with the frequency of citations, which means that the author has at most h papers that have been cited at least h times. Ball (2005) believes that the h-index is fair in evaluating scholars because it highlights scholars who have made lasting and significant contributions but have not gained a reputation. However, Moed et al. (2006) point out that the h-index is unfair to scientists who are just beginning their careers and to scientists who have a small number of publications but are highly cited. On this basis, Egghe (2006) proposed the g-index, which removed the limitation on the total number of documents and solved the problem of unfair evaluation of the h-index for scholars with short academic careers and few publications. In addition, Jin et al. (2007) proposed the extension of the h-index, R-index, and AR-index. The former corrects the flaw that scholars with the same h-index cannot be further evaluated, and the latter introduces the length of the paper for the first time and solves the problem of h, where the index value only increases but does not decrease. Li et al. (2015) proposed the v-index based on the h-index and considered the influence of the author's signature rank as a new addition for the first time.

The journal evaluation index of journals has been gradually applied to the evaluation of scholars. The grade of the paper published in the journal has become a common method for evaluating scholars. In 1955, American intelligence scientist Eugene Garfield proposed the journal impact factor (JIF), that is, the impact factor of a journal in the current year is the number of citations of papers published in the previous two years but cited in the same year divided by the number of papers published in the previous two years (Garfield, 2006). This method takes into account the number of articles and the number of citations. The research findings of Bornmann and Williams (2017) suggest that the journal impact factor has an impact on the evaluation of junior scholars. However, Shi et al. (2017) showed that the journal impact factor has an inverted U-shaped relationship with a time lag that can be manipulated. In 2008, the characteristic factor was proposed (Bergstrom et al., 2008), and applied to the evaluation of scholars by West et al. (2013). The characteristic factor considers the citation relationship between scholars, but the calculation is too complicated, and its application is difficult. In 2016, an article proposed the use of the median instead of the average to evaluate the journal impact ("Time to remodel the journal impact factor," 2016), but it did not solve the problem of the number of citations. At the same time, Scopus released a new type of journal evaluation index named CiteScore (Zijlstra & McCullough, 2016), which challenged the JIF. CiteScore is calculated based on a larger database, but the long-term lack of quantitative evaluation indicators means that the evaluation function of CiteScore is not widely accepted. In this regard, the journal impact factor still holds an unshakable position.

2.2 Research diversity

The original calculation of diversity (like Simpson diversity) did not take into account the differences between categories. On this basis, Stirling (2007) introduced the concept of diversity into the field of measurement science through an adjustable weight distribution to balance the consideration of differences. However, Jost (2009) pointed out that Stirling's diversity measurement method cannot simultaneously satisfy the principles of symmetry, output independence, transmission principle, homogeneity, replication principle, and normalization principle. Leinster et al. (2012) combined existing research and proposed a more complex formula for calculating diversity, which can simultaneously account for the three aspects of diversity and satisfy the six principles proposed by Jost. Scientists have gradually expanded the concept of diversity in the field of ecology. Currently, diversity research can be divided into four main subjects: disciplines, journals, authors, and institutions.

Interdisciplinary research is an important branch of research on diversity. In 1962, American psychologist R.S. Woodworth first proposed that interdisciplinary activities refer to research that breaks the boundaries of known disciplines and integrates two or more activities. Wagner et al. (2010) expanded the definition and pointed out that interdisciplinary activities are the integration of data, methods, tools, concepts, and theories from each discipline to solve more complex problems and develop a more holistic and comprehensive understanding of the problem. Steele and Stier (2000) used the Brillouin index (Brillouin & Hellwarth, 1956) proposed in 1956 to measure the degree of interdisciplinarity in environmental science. Inspired by this, Ma and Chen (2015) used the Brillouin index to investigate the diversity of 23 humanities and social science disciplines by combining the common characteristics of the diversity of disciplines and biodiversity and found that the Brillouin index can well reflect the differences between disciplines. In addition, Huang and Chang (2011) used direct citation and author's collaborative analysis for the first time to examine the intersectionality of graphic science. The study considers changes over time and experiences the degree of increase in diversity between different disciplines. Inspired by biodiversity indicators in ecology, Zhang et al. (2016) proposed a new diversity indicator, 2DS, which measures the diversity of knowledge across sciences and has been widely used. Leydesdorff et al. (2019) used diversity as an indicator to quantify the degree of interdisciplinary collaboration. Diversity has become an important indicator in interdisciplinary research.

Another important branch of diversity research is the measurement of journal diversity. Rafols and Meyer (2010) introduced diversity systematically and quantitatively into the field of information science for the first time. This study constructed an easy-to-understand conceptual framework for interdisciplinary research and introduced two indicators of topic diversity and consistency to measure the degree of differences and similarity among research units. Based on this, Liu et al. (2012) constructed a collection-based new knowledge framework that is more suitable for measuring journal diversity and provides a foundation for follow-up research.

Skupin et al. (2013) visually represent journal topic diversity by visualizing the topic distribution of journals in the field of medicine. Leydesdorff et al. (2018) evaluated the degree of interdisciplinarity in journals by analyzing the impact of diversity and centrality of journals at different levels in the academic network. Zhang et al. (2016) used the method proposed by Leinster and Cobbold (2012), combined with subject classification models, citation analysis, and other methods to measure the diversity of journals. This study adds to the measurement of journal diversity and examines the relationship between the degree of journal interdisciplinarity and the impact of citations. There is an "optimal value" between the degree of citation influence and interdisciplinarity. From the perspective of citations, diversity is not as high as possible, but the conclusion does not apply to all disciplines and journals (Zhang et al., 2016). These studies are mostly based on the degree of interdisciplinarity of journals and evaluate and rank journals based on the calculation of diversity. However, the "disciplines" and "topics" in the proposed indicators are mainly determined by people. Waltman (2016) pointed out that this artificial classification method is highly subjective, so it is difficult to reach a consensus. Based on this, Bu et al. (2021) made improvements by extracting technical terms from abstracts of articles published in journals, clustering them in a network of co-occurrence relations to obtain more fine-grained topics, and then calculating the diversity. This avoids the subjectivity of manual topic-level classification.

The study of author diversity can be divided into the study of the diversity of the author group itself and the influence of author interdisciplinarity. In the former, more attention is paid to the objective differences of the author group, such as race, gender, nationality, etc. Ghiasi et al. (2015) studied 680,000 authoritative journal articles in engineering and 970,000 authors and found that female scientists accounted for a very small proportion of the number, citations, and scientific research collaborations, but they occupied more prominent positions at the center of the collaboration network. Lerback et al. (2020) found that articles published by research teams with greater country and gender diversity have higher citation rates. The greater the ethnic diversity in the US author team, the lower the team's acceptance rate and the article's citation rate. This indicates that on a global level, the greater the diversity of members of a scientific research team, the more beneficial it is. However, in individual countries, residual racial discrimination may cause diversity to have a negative impact.

2.3 Diversity of research interests

The research interests of early scientists tended to be very broad, but as science developed and specialization deepened, most scientists became more willing to perform in a specific field, which also led to a simplification of research interests. At present, interdisciplinary research is increasing day by day, which is conductive to the change of research interests of scientists, and encourages more scientists to expand their research fields and conduct more extensive research. In their study, De Domenico et al. (2016) pointed out that a fundamental driver of scientific research is the evolution of scientists' research interests, which is particularly manifested in the change of research topics. This study shows the importance of change in the research interests of scientists. Jia et al. (2017) analyzed the three characteristics of transfer of scientists' research interests: heterogeneity, topicality hypothesis, and topic similarity, and used the seaside walk model to explain the reasons for the change in scientists' research interests.

In addition, the influence of interdisciplinarity on authors is also an important part of exploring the diversity of research interests of scientists. Steele and Stier (2000) confirmed in their research that the more frequently the literature is cited, the greater the influence, which also reflects that the more research interests the scholars have, the greater their influence. However, Li et al. (2018) found that there is no compelling relationship between scientists' interdisciplinary citation preference and their academic influence, so it needs to be analyzed separately in different fields. The degree to which academics are inclined to engage in interdisciplinary research activities also varies. Hirsch (2005) found that female scientists are more willing to conduct interdisciplinary research activities than male scientists. In other words, the overall research interests of female scientists are diverse. Rhoten and Pfirman (2007) also confirmed that female scientists are more attracted to interdisciplinary research and are more likely to be successful in conducting interdisciplinary research.

3 Methodology

3.1 Data collection

In this exploratory study, we selected 21 journals ranked as Quarter 1 under the discipline of Information Science & Library Science by the Journal Citation Report (https://jcr.clarivate.com/jcr/home). We retrieved all publications of each journal through the Web of Science (http://apps.webofknowledge.com/) and downloaded the data with all fields on December 18th, 2021. The main fields used in the follow-up study included authors, publication year, addresses, abstracts, and citation frequency. The retrieved dataset contained 24,400 papers, ranging from 1975 to 2021.

We removed the data where any of the fields for authors, publication year, addresses, abstract, and citation frequency were empty. Records containing empty data fields were mainly found in the early literature. After purging the records with empty fields, the dataset contained 22,626 papers, ranging from 1979 to 2021.

3.2 Data preparation

3.2.1 Field matching

The format of the author fields and address fields in the original data is shown in Table 1 (a). We have converted them to the various types in Table 1 (b) using string matching. With a simple match, we can separate each author and the corresponding address from the authors and addresses. In cases where an author matches more than one address in a publication, as in Example B in Table 1 (a), we keep only the first matched address.

Table 1 (a) The format of the Authors fields and Addresses fields in the original data

Example	Author	Address
Example A	Author1; Author2; Author3	[Author1; Author2] Address1; [Author3] Address2
Example B	Author1; Author2	[Author1] Address1; [Author1, Author2] Address2

Table 1 (b) The format of the Authors fields and Addresses fields in the converted data

Example	Author	Address
Example A	Author1	Address1
Example A	Author2	Address1
Example A	Author3	Address2
Example B	Author1	Address1
Example B	Author2	Address2

For the disambiguation of author names, we separate the list of authors and list of addresses and keep the author-address-abstract-paper ID association. However, there is no connection between the author list and the institution list in data before 2008 in WoS. As a result, we could not match the author and the institution, as shown in Table1. In addition, considering the time lag of citation, we restrict the data ranging from 2008 to 2018. In total, 34,531 distinct authors with 10,620 papers are matched.

3.2.2 Author name disambiguation

In most cases, the author's main affiliation, such as the name of the affiliated university or institute, appears at the beginning of the associated address. We extract the main affiliation by taking the address before the first comma. Given the large volume of authors in the original dataset, we roughly assume they are the same person if the combination of truncated address and author name is the same.

To have enough publications to recognize the research diversity of the author, we kept the authors who published more than five papers in the selected journals. The remaining authors were manually disambiguated by matching the corresponding full addresses to ensure an exact match. Thus, after manual disambiguation, we found 581 authors with at least five publications and 3,386 papers written by these authors. The following analysis was performed based on this dataset.

3.3 Topic detection

In this study, we identified the research topics of the researchers based on the abstracts of their publications. In addition, the topics were uncovered through the topic modeling method.

3.3.1 Text preparation

Before topic detection, the text is processed in the following steps: Tokenizing the text, replacing all whitespaces with single spaces, removing all punctuation and numbers, removing stop words, lemmatizing words, stemming words, and removing frequent and rare words. Most of this work is done using the Python package Spacy. For most of the steps, we just used the default method, except for stemming words and removing frequent and rare words. There are three main algorithms for extracting word stems: Porter, Snowball, and Lancaster. The Porter algorithm was developed in the 1980s, and its main focus was on removing common word endings to parse them into generic forms. It is a good basic word stem parser but is not recommended for complex applications. In general, it is used in research as a good basic stemming algorithm that guarantees repeatability. It is also a very mild stemming algorithm compared to other algorithms. In contrast, Lancaster's algorithm is the most aggressive stemming method that sometimes transforms the input into some rather strange words. With several rounds of experiments on our dataset, we decided to use the Snowball algorithm. The Snowball algorithm is also known as the Porter2 stemming algorithm. It is considered better than Porter because it extends Porter with many optimizations. The difference in accuracy between Snowball and Porter was approximately 5%.

Note that the most frequent and least frequent words can affect the generation of distinguishable topics using topic modeling methods. Therefore, we tested the removal of a certain number of terms based on term frequency. Previous studies have used two types of word frequency statistics: pure word frequency statistics and statistics based on the number of occurrences of a document (document-word frequency). In the latter, the documents in which the word occurs are counted. After testing with different settings, we defined words with a document-word frequency in the top 20% as frequent words and words with a pure word frequency of less than 20% as rare words and then removed them. Finally, we create the vectorized representation of the documents to feed the LDA model with them by computing the bag-of-words.

3.3.2 Latent Dirichlet Allocation (LDA)

The Latent Dirichlet Allocation (LDA) model was originally proposed by Blei et al. (2003) and soon became popular in topic detection (Tan et al., 2021). It is an unsupervised machine learning algorithm that eliminates the prior probability of the probabilistic latent semantic analysis (PLSA) model and the problem of easy over-fitting. LDA's core premise is that each document in a corpus is formed from the distribution of topics, and each word in the document is generated based on the word distribution per topic. The LDA model takes input from the document set and outputs the topic-document matrix. Each document is represented by a topic distribution vector. Therefore, we can use the output of the LDA model to calculate the similarity of the documents.

After the word extraction and topic detection step, we extracted 1,532 unique keywords (words or phrases with the same stems were combined into one keyword). By implementing the LDA model with abstracts from the 3,386 papers, we plotted the curve of the number of topics versus coherence based on normalized pointwise mutual information as shown in Figure 2 (Röder et al., 2015). According to Figure 2, we chose topic numbers 8, 13, and 17 to train the LDA model with more iterations to achieve the best result in topic detection.

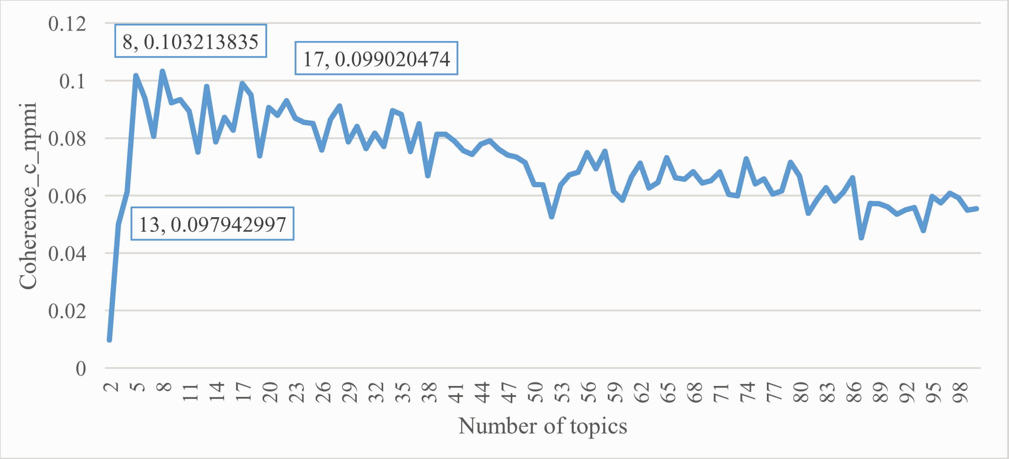

Figure 2 The Distribution of The Coherence of Topics

3.4 Diversity calculation

The first important question is which proximity metric to use to characterize document similarity. In different real-world applications, different proximity metrics may be required depending on the individual conditions. For example, Euclidean distance can adequately capture the differences between quantitative data in most circumstances. However, Kriegel et al. (2008) found that the angle variances between high-dimensional feature vectors are more sensitive than Euclidean distance. Cosine similarity produces better results in this scenario.

3.4.1 Cosine similarity

Cosine similarity measures the similarity between two non-zero vectors in an inner product space. It is equal to the cosine of the angle between them, which is equal to the inner product of the same vectors normalized to the same length. In the LDA model, each document is represented by what is called a document-topic vector. Usually, the cosine of the angle between two vectors is a useful measure of how similar two documents are in terms of topic matter (Singhal, 2001). Pennacchiotti and Gurumurthy (2011) presented a user recommendation system that recommends new friends with similar interests as a user and found that the best configurations are with LDA and cosine similarity, gaining +0.2 AUC on the baseline, outperforming existing strategies based on graph analysis.

The Euclidean dot product formula (3-1) can be used to calculate the cosine of two non-zero vectors.

$$\vec{\mathrm{A}}\cdot\vec{\mathrm{B}}= \|\vec{\mathrm{A}}\|\|\vec{\mathrm{B}}\|\cos(\theta) \quad (3\text{-}1)$$

The cosine similarity, cos(θ), is represented using a dot product and magnitude as formula (3-2) given two vectors of attributes, $\vec{A}$ and $\vec{B}$. A_i are B_i components of vectors A and B, respectively.

$$D_{cosine}(\mathrm{P,Q})=1 - \text{similarity} = 1 - \cos(\theta) = 1 - \frac{\vec{\mathrm{A}}\cdot\vec{\mathrm{B}}}{\|\vec{\mathrm{A}}\|\|\vec{\mathrm{B}}\|} = 1 - \frac{\sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{A_i B_i}}{\sqrt{\sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{A_i^2}}\sqrt{\sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{B_i^2}}} \quad (3\text{-}2)$$

The resulting similarity ranges from 0 to 1, with 0 representing orthogonality or decorrelation and values in-between representing moderate similarity or dissimilarity. Since the topic distribution probabilities cannot be negative, the cosine similarity of two documents ranges from 0 to 1 when analyzing text with the LDA model. We then define the cosine distance of two documents as the difference between their cosine similarity and 1.

3.4.2 Author topic diversity

Most of the work in topic diversity research has focused on topics in specific disciplines. Hall et al. (2008) used Latent Dirichlet Allocation and examined the strength of each topic over time to compare the diversity of ideas at different conferences. Bu et al. (2021) applied word-topic networks for topic detection to extract fine-grained topics and fitted certain diversity indicators to calculate journal topic diversity. Zeng et al. (2019) used a co-citing network of scientists to quantify the dynamics of their topic changes. In this study, author topic diversity (ATD) is defined as the average distance of the topics of all papers published by an author, which is expressed in formula (3-3). P_i *and* P_j are two different articles, and n is the number of papers of an author.

$$\text{Author topic diversity} = \frac{\sum_{i=1}^{n}\sum_{j=i+1}^{n} D(P_i, P_j)}{C_n^2} \quad (3\text{-}3)$$

The ATD ranges from 0 to 1 because the cosine distance ranges from 0 to 1, as formula (3-2) shows. A higher ATD for an author means a greater diversity of topics in the author's papers. Authors who have written papers on completely different topics have an ATD approaching 1, while the authors who have written all papers on exactly the same topic have an ATD approaching 0.

4 Results and discussion

After completing the data preparation steps, 581 authors with at least five papers after manual disambiguation and 3,386 papers written by these authors were included in the following analysis. We calculated the topic diversity of each author and the other indicators for comparison.

4.1 Data and indicators

We calculated the ATD based on the cosine similarity (ATD) with different LDA models for comparison. Table 2 summarizes the basic statistical indicators for ATD with 8 topics, 13 topics, and 17 topics. The result is that the statistical indicators for ATD increase with the number of topics. With 8 topics, 581 authors in library and information science have an average ATDtopic8 of 0.286113 (with maximum and minimum ATDtopic8 of 0.670264 and 0.007787, respectively), while the standard deviation of ATDtopic8 is 0.154723 It is quite interesting to see that topic diversity in the field of library and information science is rather different from author to author. As can be seen in Table 2, some of the descriptive statistics (including the mean, median, maximum, and minimum) of ATD increase with the number of topics. This is to be expected since the topic assignment for each paper increases with an increase in the availability of topics. A higher number of topics indicates that papers have more opportunities to include more topics, which means that authors are more likely to write papers on a variety of topics. However, the standard deviations stay stable with different numbers of topics, which means that the proposed indicator is robust to a great extent.

Table 2 The descriptive statistics of ATD with 8 topics, 13 topics, and 17 topics

Indicator	Mean	Median	Max.	Min.	SD.	N
ATDtopic8	0.286113	0.288442	0.670264	0.007787	0.154723	581
ATDtopic13	0.337261	0.340230	0.727103	0.018821	0.158894	581
ATDtopic17	0. 388788	0.401093	0.761461	0.030583	0.150907	581

In Figure 3, the x-axis represents the total number of papers published by an author, while the y-axis represents the mean and standard deviation of authors who published the corresponding number of papers, in the circumstances of 8 topics, 13 topics, and 17 topics, respectively. It shows that the mean and standard deviation of ATD vary among authors with a different number of publications, but there is no obvious upward or downward trend as

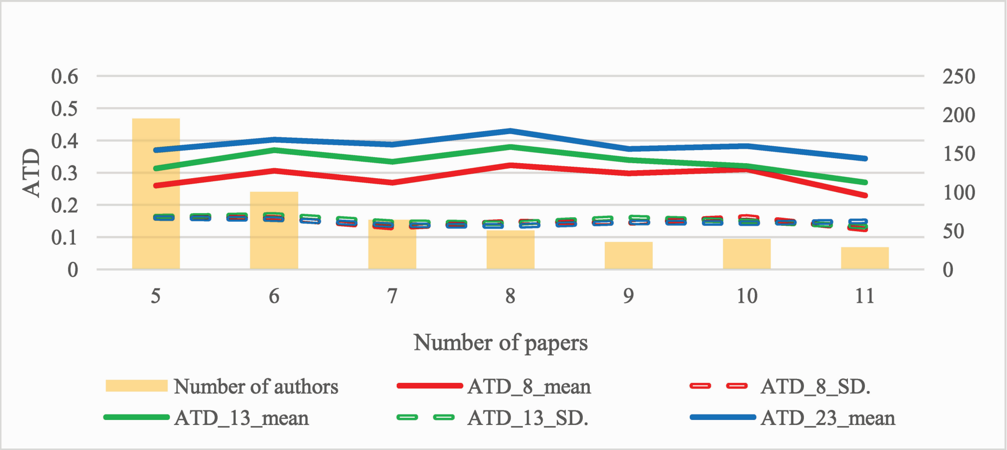

Figure 3 The mean and standard deviation of ATD vary among authors with different number of papers

the number of publications increase. This suggests that it is not necessary for authors who publish more papers to have a greater topic diversity. Thus, ATD can eliminate the influence of the total number of papers per author on the calculation of the index and reduce possible bias. When setting a different number of topics, the curves of the mean and standard deviation of the ATD fluctuate more sharply as the number of topics increases.

4.2 Indicators for comparison

We further ask how the topic diversity of researchers is related to their influence on research. To this end, we calculated the correlations between the ATD and citation indicators generated (as shown in Figure 4). We adopt Spearman and Kendall's correlation instead of Pearson's correlation because the data does not obey the typical normal distribution. Spearman correlation analysis and Kendall correlation analysis showed weak positive correlations between the ATD and all three citation indicators (i.e., total citations, average citations, and maximum citations) with high statistical significance ($p < 0.01$). This interesting finding suggests that authors who engage in more diverse topics may have a higher impact. Moreover, the resulting correlation coefficients are all lower than 0.19 and decrease as the number of topics increases. In general, the correlation coefficients between the ATD and total citations are stronger than the correlation coefficients between the ATD and citations per paper, as well as the maximum number of citations.

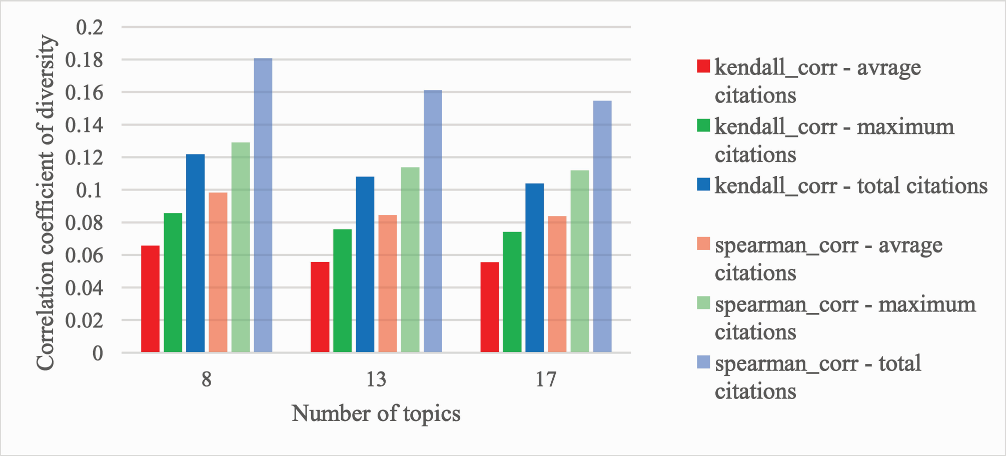

Figure 4 Comparison of different correlation coefficients and correlation variables (p-value < 0.01)

Compared to the correlation coefficients obtained in previous research, the correlation coefficients between the author-level diverse indicator and the research impact indicators produced in this study are much lower. Bu et al. (2021) calculated the correlations of six journal-level diversity indicators and impact factors for journals, and the results ranged from 0.457 to 0.714. Zeng et al. (2019) calculated the correlations between scientists' topic switching probability and research performance and found that they were significant and strongly correlated.

In this study, the results show that ATD has a weak correlation with both the number of papers and the number of citations. However, there are differences in ATD between authors. The following analysis is based on the results of the model that set the topic number as 8. In the group of authors with a high ATD, there are authors with both a high number of papers and citations, such as Pinsonneault, Alain (0.59), Agarwal, Ritu (0.54), and Pavlou, Paul A. (0.50). Likewise, we examined authors who have a low number of papers and citations but perform well in ATD. As a result, it is observed that authors with high ATD publish papers covering various topics. The examples are listed in Table 3.

Table 3 Examples of the authors with the high topic diversity

Author	**ATD**
Ghinea, Gheorghita	0.67
Title	**Citation**
What do you wish to see? A summarization system for movies based on user preferences	20
Why do commercial companies contribute to open source software?	31
On the motivating impact of price and online recommendations at the point of online purchase	30
Web 2.0 and folksonomies in a library context	19
User perceptions of online public library catalogues	16
Author	**ATD**
Jones, Donald R.	0.61
Title	**Citation**
Contained nomadic information environments: Technology, organization, and environment influences on adoption of hospital RFID patient tracking	59
Volunteers´ involvement in online community–based software development	59
The cognitive selection framework for knowledge acquisition strategies in virtual communities	22
Closing the loop: Empirical evidence for a positive feedback model of IT business value creation	6
Conceptualizing the Dynamic Strategic Alignment Competency	47
Using Visual Representations of Data to Enhance Sensemaking in Data Exploration Tasks	38
Author	**ATD**
Armstrong, Deborah J.	0.60
Title	**Citation**
Exploring neuroticism and extraversion in flow and user generated content consumption	21
Factors impacting the perceived organizational support of IT employees	48
The advancement and persistence of women in the information technology profession: An extension of Ahuja´s gendered theory of IT career stages	19
The impact of relational leadership and social alignment on information security system effectiveness in Korean governmental organizations	13
Patterns of Transition: The Shift from Traditional to Object–Oriented Development	10
Exhaustion from information system career experience: implications for turn–away intention	29

In the group of authors with a low ATD, it is not surprising that authors with a low number of publications and citations tend to focus on fewer topics. On the contrary, some authors publish a large number of papers and receive many citations, but tend to focus intently on specific topics, such as the authors listed in Table 4. These authors have focused on one or two topics in their studies. The three authors with relatively low ATD values and high publication counts and average citations as listed in Table 4 have been focused on informetric studies. When we set the number of topics as 8, the average coherence of the 8 topics is 0.1, while the coherence of the topic related to informetric studies reaches 0.15, which means that papers in this field tend to be more focused. Therefore, we observe that although some authors have been concentrated on a few topics, they may still achieve considerable research impact.

Table 4 Examples of the authors with the low topic diversity

Author	ATD	Total papers	Average citations
D´Angelo, Ciriaco Andrea	0.055141	22	29
Thelwall, Mike	0.060284	18	15
Bornmann, Lutz	0.095913	34	46

5 Conclusion

Despite ongoing efforts to better assess scientists' research performance in the field of library and information science, little is considered about the coverage of research topics throughout their careers. In this study, we propose an indicator to quantify the authors' topic diversity, namely author topic diversity (ATD). ATD does not depend on predefined classification schemes. Instead, Latent Dirichlet Allocation (LDA) is implemented to detect topics based on the abstracts of papers in a given domain. The average distance between the LDA-based representations of all papers by an author was considered when calculating the ATD. One of the most important findings of this study is that weak positive correlations occur between the ATD and all three citation indicators (i.e., total citations, average citations, and maximum citations) with high statistical significance ($p < 0.01$). Therefore, our results suggest that engaging with more topics can lead to a higher research impact.

Our work has made a first contribution to the knowledge of author-level research evaluation by providing a more advanced understanding of the underlying mechanism of researchers' topic choice and research variation. However, our sample was drawn from 21 journals in the category of library and information science. A larger sample would provide a more comprehensive look at how different patterns may occur in diverse disciplines. This indicates obvious directions for further research in comparing topic diversity for authors in different fields.

Acknowledgment

This study is supported in part by the National Natural Science Foundation of China (No. 72004091), and the Humanity and Social Science Foundation of Ministry of Education of China (#20YJC870014).

Reference

Ball, P. (2005). Index aims for fair ranking of scientists. *Nature, 436,* 900. https://doi.org/10.1038/436900a

Bergstrom, C. T., West, J. D., & Wiseman, M. A. (2008). The Eigenfactor metrics. *The Journal of neuroscience: the official journal of the Society for Neuroscience, 28* (45), 11433–11434.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *the Journal of Machine Learning research, 3*, 993–1022.

Bornmann, L., & Williams, R. (2017). Can the journal impact factor be used as a criterion for the selection of junior researchers? A large–scale empirical study based on ResearcherID data. *Journal of Informetrics, 11* (3), 788–799. https://doi.org/10.1016/j.joi.2017.06.001

Bourdieu, P. (1975). The specificity of the scientific field and the social conditions of the progress of reason. *Social Science Information, 14* (6), 19–47.

Brillouin, L., & Hellwarth, R. W. (1956). Science and information theory. *Physics Today, 9* (12), 39–40.

Bu, Y., Li, M., Gu, W., & Huang, W.–b. (2021). Topic diversity: A discipline scheme–free diversity measurement for journals. *Journal of the Association for Information Science and Technology, 72* (5), 523–539. https://doi.org/10.1002/asi.24433

De Domenico, M., Omodei, E., & Arenas, A. (2016). Quantifying the diaspora of knowledge in the last century. *Applied Network Science, 1* (1), 15–15. https://doi.org/10.1007/s41109–016–0017–9

Egghe, L. (2006). Theory and practise of the g–index. *Scientometrics, 69* (1), 131–152. https://doi.org/10.1007/s11192–006–0144–7

Foster, J. G., Rzhetsky, A., & Evans, J. A. (2015). Tradition and innovation in scientists´ research strategies. *American Sociological Review, 80* (5), 875–908.

Garfield, E. (2006). Citation indexes for science. A new dimension in documentation through association of ideas. *International Journal of Epidemiology, 35* (5), 1123–1128. https://doi.org/10.1093/ije/dyl189

Gao, Z., & Zhang, Z. (2016). Review of quantitative evaluation method for individual academic influence. I*nformation Studies:Theory & Application, 39* (1), 133–138.

Ghiasi, G., Lariviere, V., & Sugimoto, C. R. (2015). On the compliance of women engineers with a gendered scientific system. *PloS one, 10* (12), e0145931. https://doi.org/10.1371/journal.pone.0145931

Hall, D., Jurafsky, D., & Manning, C. D. (2008). Studying the history of ideas using topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP ´08)*, 363–371.

Hirsch, J. E. (2005). An index to quantify an individual´s scientific research output. In *Proceedings of the National Academy of Sciences of the United States of America, 102* (46).

Huang, M., & Chang, Y. (2011). A study of interdisciplinarity in information science: using direct citation and co–authorship analysis. *Journal of Information Science, 37*(4), 369–378. https://doi.org/10.1177/0165551511407141

Jia, T., Wang, D., & Szymanski, B. K. (2017). Quantifying patterns of research–interest evolution. *Nature Human Behaviour, 1* (4), Article 0078. https://doi.org/10.1038/s41562–017–0078

Jin, B., Liang, L., Rousseau, R., & Egghe, L. (2007). The R– and AR–indices: Complementing the h–index. *Chinese Science Bulletin, 52* (6), 855–863. https://doi.org/10.1007/s11434–007–0145–9Jost, L. (2009).

Mismeasuring biological diversity: Response to Hoffmann and Hoffmann (2008). *Ecological Economics, 68* (4), 925–928. https://doi.org/10.1016/j.ecolecon.2008.10.015

Kriegel, H.–P., Schubert, M., & Zimek, A. (2008). Angle–based outlier detection in high–dimensional data. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining,* 444–452. https://doi.org/10.1145/1401890.1401946

Leinster, T., & Cobbold, C. A. (2012). Measuring diversity: The importance of species similarity. *Ecology, 93* (3), 477–489.

Lerback, J. C., Hanson, B., & Wooden, P. (2020). Association between author diversity and acceptance rates and citations in peer–reviewed earth science manuscripts. *Earth and Space Science, 7* (5), e2019EA000946. https://doi.org/10.1029/2019ea000946

Leydesdorff, L., Wagner, C. S., & Bornmann, L. (2018). Betweenness and diversity in journal citation networks as measures of interdisciplinarity–A tribute to Eugene Garfield. *Scientometrics, 114* (2), 567–592. https://doi.org/10.1007/s11192–017–2528–2

Leydesdorff, L., Wagner, C. S., & Bornmann, L. (2019). Interdisciplinarity as diversity in citation patterns among journals: Rao–Stirling diversity, relative variety, and the Gini coefficient. *Journal of Informetrics, 13* (1), 255–269. https://doi.org/10.1016/j.joi.2018.12.006

Li, D., Tong, S., & Li, J. (2018). Analyzing interdisciplinarity and scientists´ academic impacts. *Data Analysis and Knowledge Discovery, 2* (12), 1–11.

Li, H., Xu, Q., & Li, E.(2015). A new improved indicator for the influence evaluation of the field——v–index. *Journal of Intelligence, 34* (12), 38–43.

Liu, Y., Rafols, I., & Rousseau, R.(2012). A framework for knowledge integration and diffusion. *Journal of Documentation, 68* (1), 31–44.

Ma, F., & Chen, B.(2015). Study of discipline diversity in the field of Chinese humanities and social sciences. *Information Science, 33* (4), 3–8+63.

Moed, F.H., Liu, J., & Jin, B.(2006). The h index is constructed creatively and used in evaluation with caution. *Science Focus, 1* (1), 15. (In Chinese).

Okamura, K. (2019). Interdisciplinarity revisited: Evidence for research impact and dynamism. *Palgrave Communications, 5* (1), 1–9.

Pennacchiotti, M., & Gurumurthy, S.(2011). Investigating topic models for social media user recommendation. In *Proceedings of the 20th international conference companion on World wide web*, 101 –102. https://doi.org/10.1145/1963192.1963244

Rafols, I., & Meyer, M.(2010). Diversity and network coherence as indicators of interdisciplinarity: Case studies in bionanoscience. *Scientometrics, 82* (2), 263–287 (2010). https://doi.org/10.1007/s11192–009–0041–y

Rhoten, D., & Pfirman, S. (2007). Women in interdisciplinary science: Exploring preferences and consequences. *Research Policy, 36* (1), 56–75. https://doi.org/10.1016/j.respol.2006.08.001

Röder, M., Both, A., & Hinneburg, A.(2015, February). Exploring the space of topic coherence measures. In Proceedings of the eighth ACM international conference on Web search and data mining (pp. 399–408).

Schubert, A., & Glänzel, W.(2006). A systematic analysis of Hirsch –type indices for journals. *Journal of Informetrics, 1* (3), 179–184. https://doi.org/10.1016/j.joi.2006.12.002

Shi, D., Rousseau, R., Yang, L., & Li, J. (2017). A journal´s impact factor is influenced by changes in publication delays of citing journals. *Journal of the Association for Information Science and Technology, 68* (3), 780–789. https://doi.org/10.1002/asi.23706

Singhal, A. (2001). Modern information retrieval: A brief overview. *IEEE Data Eng. Bull., 24* (4), 35–43.

Skupin, A., Biberstine, J. R., & B?rner, K.(2013). Visualizing the topical structure of the medical sciences: A self–organizing map approach. *PloS one, 8* (3), e58779. https://doi.org/10.1371/journal.pone.0058779

Steele, T. W., & Stier, J. C.(2000). The impact of interdisciplinary research in the environmental sciences: A forestry case study. *Journal of the American Society for Information Science, 51* (5), 476–484.

Stirling, A.. (2007). A general framework for analysing diversity in science, technology and society. *Journal of The Royal Society Interface, 4* (15), 707–719. https://doi.org/10.1098/rsif.2007.0213

Tan, C. & Xiong, M.(2021). Contrastive analysis in China and abroad on the Evolution of hot topics in the field of digital library based on LDA model. *Data Science and Informetrics, 1* (2), 110–130.

Time to remodel the journal impact factor. (2016). *Nature, 535* (7613).

Wagner, C. S., Roessner, J. D., Bobb, K., Klein, J. T., Boyack, K. W., Keyton, J., . . . B?rner, K. (2010). Approaches to understanding and measuring interdisciplinary scientific research (IDR): A review of the literature. *Journal of Informetrics, 5* (1), 14–26.

Waltman, L.(2016). A review of the literature on citation impact indicators. *Journal of Informetrics, 10* (2), 365–391. https://doi.org/10.1016/j.joi.2016.02.007

West, J. D., Jensen, M. C., Dandrea, R. J., Gordon, G. J., & Bergstrom, C. T.(2013). Author–level Eigenfactor metrics: Evaluating the influence of authors, institutions, and countries within the social science research network community. *Journal of the American Society for Information Science and Technology, 64* (4), 787–801. https://doi.org/10.1002/asi.22790

Yegros–Yegros, A., Rafols, I., & D´este, P.(2015). Does interdisciplinary research lead to higher citation impact? The different effect of proximal and distal interdisciplinarity. *PLoS One, 10* (8), e0135095.

Zeng, A., Shen, Z., Zhou, J., Fan, Y., Di, Z., Wang, Y., . . . Havlin, S.(2019). Increasing trend of scientists to switch between topics. *Nature communications, 10* (1), 1–11.

Zhang, L., Rousseau, R., & Gl?nzel, 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.

Zijlstra, H., & McCullough, R.(2016). CiteScore: *A new metric to help you track journal performance and make decisions.* Elsevier. https://www.elsevier.com/editors–update/story/