# Analyzing Research Diversity of Scholars Based on Multi-dimensional Calculation of Entities

Chuhan Wang
School of Information Management
Sun Yat-sen University
Guangzhou Guangdong China
wangchh33@mail2.sysu.edu.cn

Tongyang Zhang
School of Information Management
Sun Yat-sen University
Guangzhou Guangdong China
tzhang39@163.com

Yi Bu
Department of Information Management
Peking University
Beijing China
buyi@pku.edu.cn

Jian Xu *
School of Information Management
Sun Yat-sen University
Guangzhou Guangdong China
issxj@mail.sysu.edu.cn

## Abstract

Understanding diversity of research contents is essential for facilitating analysis of scholars' research characteristics. The multi-dimensional character of research diversity makes its analysis a challenging issue. Based on the three primary attributes of research diversity: variety, evenness and disparity, we apply three multi-dimensional calculation methods to analyze the research diversity of scholars as well as discuss scenarios to which the methods apply by comparing calculation results between them. Three categories of research diversity calculated include: one- dimensional diversity (measuring variety, evenness, disparity), two-dimensional diversity (measuring variety/evenness, variety/disparity), and three-dimensional diversity (measuring variety/evenness/disparity). Preliminary results of the three methods show evident differences. The one-dimensional diversity is feasible and can directly demonstrate research diversity in a certain aspect; the two-dimensional diversity can guarantee a more complicated demonstration of the diversity characteristics; the three-dimensional diversity fully reflects different diversity characteristics and is highly adjustable for highlighting certain aspect of diversity. The approach we apply offers a foundation for further studies on applying diversity calculation to evaluating academic performance of scholars in information science.

***Keywords:*** Research Diversity, Multi-dimensional Calculation, Entitymetrics

## 1 Introduction

Originally defined as an index to measure the variety of animal species in biology [1], diversity has now been widely applied to capture multiculturalism in politics, diverse customers in business, multiple transmit channels in technology, etc. The increasing wide range of contents that research studies not only enrich the original system of knowledge, but also reflect the dynamic research characteristics of scholars.

In previous work, research diversity analysis is usually associated with paper-level bibliometric studies for capturing interdisciplinarity among fields. However, detailed knowledge-level studies on the diversity of research contents are insufficient, which can offer intricate feature analysis and performance assessment of scholars. Meanwhile, there exists challenges that the use of traditional single-dimension index (e.g., species variety, disparity) is short of systematically comprehensive view angle, which impedes the full-featured diversity analysis of research. Through the reasonably combination of different indicators, research diversity of scholars can be grasped more comprehensively.

In this study, diversity analyses on research contents respectively in three dimensional structures are performed by taking account into primary attributes of diversity in biology--variety, evenness, and disparity [2]. Fixed combinations of attribute dimension ignore the changing special needs of assessment. For instance, if we require putting the emphasis only on the entity variety of the research, a multiple-dimension measurement schema is not needed. This inspires us to calculate different combinations of diversity attributes, respectively. Furthermore, we explore the feasibility of different diversity calculation methods through using the allotaxonograph for comparing diversity rankings of scholars across different dimension combinations.

Through applying multiple diversity indexes used in biology to measure research diversity reflected in bio-entities, precise and comprehensive understanding can be drawn regarding various aspects of research diversity. At the same time, the applicability of different multi-dimensional calculation methods discussed provide a reference for method selection, ultimately leading to enriching the current system for evaluating scientific scholars and promoting the development of scientific progress.

## 2 Methodology

All experimental data are obtained from PubMed Knowledge Graph (PKG) [3], a dataset extracting biomedical entities from all PubMed article abstracts and disambiguating author names with an F1 score of 98.09% according to report. The study adopts extracted entities on the types of Gene/Protein and Drug/Chemical, which are

* Corresponding author

used as proxies as research contents, to measure research diversity of scholars. We identify authors studying gene/drugs based on the author, entity and literature data from PKG in 1988-2017. Then we study the diversity of scholars' research contents from three necessary dimensions, variety, evenness and disparity individually or in combination.

## 2.1 Calculating the One-dimensional Diversity

*2.1.1Variety.* The variety of research content is represented by the number of distinct bio-entities (N) covered in all articles of an author as shown in Formular 1:

$$Div_N = \mathrm{N} \tag{1}$$

When the number of distinct entities increases, the diversity of the research content is also improved.

*2.1.2 Evenness.* Using Pielou's [4] measure of species evenness for reference, evenness in this study reflects the degree of balance in the distribution of various research content proportion $p_i$:

$$Div_S = (-\sum_i p_i \ln p_i)/\ln N \tag{2}$$

where $p_i$ presents the proportion that the quantity of the entity $i$ takes up in all entities. Given a certain number of varieties, the more evenness the entity variety, and thus the higher equality degree of the research content distribution.

*2.1.3 Disparity.* Disparity refers to the distance between research contents, which is indicated by the cosine distance $d_{ij}$ between any two entity vectors $x_i$ and $y_j$:

$$d_{ij} = 1 - \cos\theta \tag{3}$$

$$\cos\theta = \frac{X \cdot \mathrm{Y}}{\|X\| \cdot \|Y\|} = \frac{\sum_{i=1}^{n} x_i y_j}{\sqrt{\sum_{i=1}^{n} x_i} \cdot \sqrt{\sum_{j=1}^{n} y_j}} \tag{4}$$

where $x_i$ / $y_i$ respectively represents the $i_{th}$ / $j_{th}$ vector in the corresponding vector group $X$ and $Y$, and $n$ is the size of the vector dimension. In Formular 3, as the $\cos\theta$ gets closer to 1, the further between two entities.

## 2.2 Calculating the Two-dimensional Diversity

Since the variety of entities is the basic expressional form that reflects the diversity of research content, estimating the study scale without it would be difficult to carry out. We use the index variety in combine with evenness and disparity to discuss the two-dimensional diversity.

*2.2.1 Variety and Evenness.* The combination of variety and evenness expresses the concentrated level of research content. We use the diversity index Gini-Simpson [5] to quantitatively measure research diversity here:

$$Div_{GS} = 1 - \sum_{i=1}^{N} p_i^2 \tag{5}$$

*2.2.2 Variety and Disparity.* The combination of variety and disparity represents the uniqueness level of research content. Compared with the cosine distance that mainly determines differences in the direction of vectors, the Euclidean metric we use is more focused on the value difference between two vectors:

$$D_{ij}(X,Y) = \sqrt{\sum_{i=1}^{N} (x_i - y_j)^2} \tag{6}$$

## 2.3 Calculating the Three-dimensional Diversity

Based on the calculation model of Rao-Strirling (RS) diversity [6], we integrate the three primary dimensions into a single expression:

$$Div_\Delta = \sum_{ij(i\neq j)} (d_{ij})^{\alpha 1} \cdot (p_i \cdot p_j)^{\beta 1} \quad (\alpha 1=1,\ \beta 1=1) \tag{7}$$

where $d_{ij}$ denotes the cosine distance between entity $i$ and $j$. According to the RS formular definition, we can better measure the overall research diversity when setting the values of α1 and β1 to 1.

## 2.4 Comparing Calculation Methods based on Rank-turbulence Divergence

Since there is currently a lack of benchmark datasets for validating the performance of the authors' research diversity algorithms, it is difficult to directly evaluate the pros and cons of each measurement schema. Referring to validation methods used in previous similar works [7] [8], we measure the level of ranking divergence between calculation results of different schemas or a schema in different component combinations. It would better to observe how different dimensions are of differences and mutual supplement with each other, which will affect their suitability in various particular scenarios. Following previous studies, the rank-turbulence divergence, a tunable instrument for comparing any two ranked lists, is used to quantitatively compare calculation methods of research diversity in different dimensional structures [9]. Given that a scholar τ has a rank $r_{(\tau,1)}$ in the calculation method $R_1$ and $r_{(\tau,2)}$ in the calculation method $R_2$, the divergence between $R_1$ and $R_2$ is calculated as follows:

$$\begin{aligned} \boldsymbol{D}_{\boldsymbol{\alpha}}^{\mathbf{R}}(\boldsymbol{R_1} \parallel \boldsymbol{R_2}) &= \sum_{\boldsymbol{\tau} \in \boldsymbol{R}_{1,2;\alpha}} \boldsymbol{\delta D}_{\boldsymbol{\alpha},\boldsymbol{\tau}}^{\mathbf{R}}(\boldsymbol{R_1} \parallel \boldsymbol{R_2}) \\ &= \frac{\mathbf{1}}{\boldsymbol{\mathcal{N}}_{\mathbf{1},\mathbf{2};\boldsymbol{\alpha}}} \frac{\boldsymbol{\alpha}+\mathbf{1}}{\boldsymbol{\alpha}} \sum_{\boldsymbol{\tau} \in \boldsymbol{R}_{1,2;\alpha}} \left| \frac{\mathbf{1}}{\left[r_{\tau,1}\right]^{\alpha}} - \frac{\mathbf{1}}{\left[r_{\tau,2}\right]^{\alpha}} \right|^{1/(\alpha+1)} \end{aligned} \tag{8}$$

$$\begin{aligned} \boldsymbol{\mathcal{N}}_{\mathbf{1},\mathbf{2};\boldsymbol{\alpha}} = &\frac{\boldsymbol{\alpha}+\mathbf{1}}{\boldsymbol{\alpha}} \sum_{\boldsymbol{\tau} \in \boldsymbol{R}_1} \left| \frac{\mathbf{1}}{\left[r_{\tau,1}\right]^{\alpha}} - \frac{\mathbf{1}}{\left[N_1 + \frac{1}{2}N_2\right]^{\alpha}} \right|^{1/(\alpha+1)} \\ &+ \frac{\boldsymbol{\alpha}+\mathbf{1}}{\boldsymbol{\alpha}} \sum_{\boldsymbol{\tau} \in \boldsymbol{R}_1} \left| \frac{\mathbf{1}}{\left[N_2 + \frac{1}{2}N_1\right]^{\alpha}} - \frac{\mathbf{1}}{\left[r_{\tau,2}\right]^{\alpha}} \right|^{1/(\alpha+1)} \end{aligned} \tag{9}$$

where $N_1$ and $N_2$ are the number of distinct scholars in each ranked list of calculation, $\alpha$ ($\mathbf{0} \leq \boldsymbol{\alpha} \leq \infty$) is a parameter to adjust the weights of the highly/lowly ranked scholars on the compound semantic system. We tune α to make the high-ranked scholars account for as much divergence contributions of the total $\boldsymbol{D}_{\boldsymbol{\alpha}}^{\mathbf{R}}$ ($\mathbf{0} \leq \boldsymbol{D}_{\boldsymbol{\alpha}}^{\mathbf{R}} \leq \mathbf{1}$) as possible in the ranking system. The higher the divergence, the greater the ranking results between two calculation methods is.

# 3 Preliminary Results

## 3.1 Comparison of the One-dimensional Diversity results

Figure1 shows the contrast of scholar ranked lists in aspects of variety and evenness. In the graph on the left side, squares of scholars on either side of the central axis represent ones with higher ranking in the corresponding comparing aspect. Deviate from the central axis horizontally, the ranking difference of a scholar between the two lists gets bigger. In general, the overall distribution pattern of the squares is dispersive, which means there are obvious research diversity differences in the two dimensions. Upper-central area of the left graph shows scholars (e.g., Zhang P., Li N., Nakamura Y.), get high rankings in both research variety and evenness. Among scholars ranking significantly differently in the two dimensions, researches of those (e.g., De Clercq E.) with high variety and low evenness entities cover a wide scope and pay particularly attention to the in-depth study of a certain special area, while those (e.g., Zhang P.) with low variety and high evenness entities are limited in scope and allocate relatively equal attention to different studied areas.

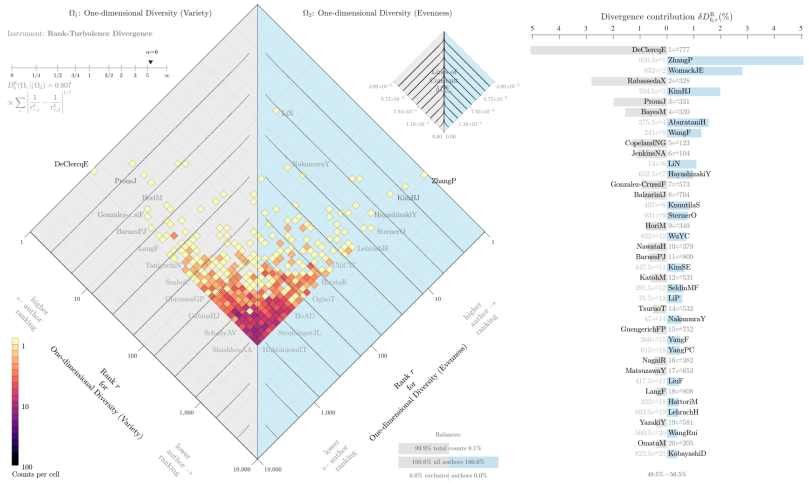

**Figure 1: Allotaxonograph Comparing Ranked Lists of Scholars in Variety and Evenness of Research.**

## 3.2 Comparison of the Two-dimensional Diversity Results

As mentioned in Section 2.1, given the two measurements of the two-dimensional diversity lay emphases on different aspects, the overall distribution pattern of squares in the allotaxonograph shown in Figure 2 is disperse widely. From squares far away from the central axis, we can find scholars whose studied entities are of high evenness and low disparity or low evenness and high disparity. For example, Rabasseda X. specializes in research of clinical trials, studying the quality of illness treatment outcomes in emphasis. The stability of variables affecting her research direction and the large number of illness cases put her remain at a high level in the ranked list of evenness/variety diversity. From the right graph of divergence contribution, we can see, affected by the relatively small quantity of entities, Zhang P., who ranks the first in evenness of the one-dimensional diversity, is reduced to rank 12th in the evenness/variety diversity.

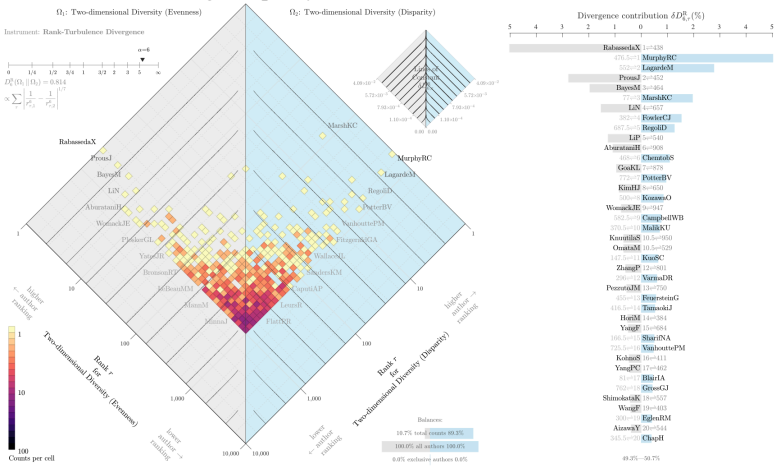

**Figure 2: Allotaxonograph Comparing Ranked Lists of Scholars in Evenness/Variety and Disparity/Variety of Research.**

### 3.3 Comparison of the Two-dimensional and Three-dimensional Diversity Results

Though from Figure 3 we see the distribution of squares shows a converging pattern as a whole, some squares representing high-ranking scholars are not distributed near the central axis and the point of converging mainly concentrate in the middle of axis, which reflect in part results of the two diversity calculation methods vary. Scholars concentrating in the middle of central axis (e.g., Chap H., Blair IA., Busse R.) have similar levels of variety, evenness and disparity in research diversity.

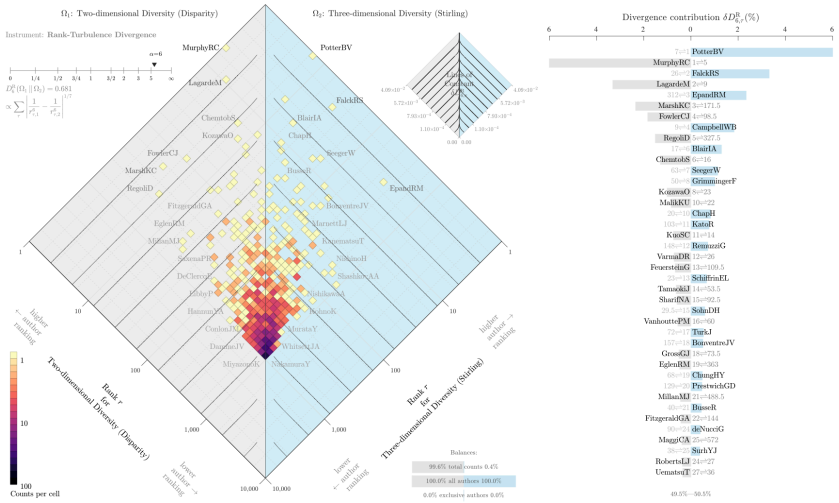


**Figure 3: Allotaxonograph Comparing Ranked Lists of Scholars in Disparity/Variety and Disparity/Variety /Evenness (Stirling) of Research.**

## 4 Conclusion and Future Work

This study analyzes the research characteristics of scholars by using entities as a proxy. It highlights three critical metrics, including variety, evenness and divergence, for assessing the diversity of research content. Different combinations of these metrics form evaluating systems in three dimensional structures.

Through comparing the three calculation methods of research diversity, we find the results of them show differences. The one-dimensional diversity is simple to calculate and feasible, but inapplicable to meet the requirements of diversity analysis on various aspects; the two-dimensional diversity is much more appropriate to diversity analysis in multiple dimensions and the emphasize in metrics can vary to distinguish any of the diversity characteristics; the three-dimensional diversity includes multiple metrics in which scholars differ, encompassing different diversity characteristics that make one scholar different from another, and it allows us to adjust the parameters for flexible testing of the research diversity. In applications of scholar evaluation, the selection of the appropriate calculation can be facilitated based on the features of each diversity calculation schema. Firstly, the one-dimensional diversity is suitable when the anticipated need is a relatively simple and easy schema that reflects diversity in a direct way. It can also be used to deeply analyze one point of diversity characteristics of an author. For instance, in the dimension of variety, through comparing the number of distinct entities for a certain author, we can identify the particular research focus of him or her. Secondly, the two-dimensional help put different emphasis on the importance of evenness/variety or disparity/variety, other than a singular focus, which gives us alternative solutions of diversity measurement and multiple perspectives in higher dimensions. It avoids limitations in the one-sidedness of the one-dimensional diversity to some extent. Thirdly, the three-dimensional diversity provides a comprehensive view of research diversity properties. The weights of different properties can be controlled to achieve the requirement for its flexible adjustment. It requires more complex calculations and data amount.

Applying research diversity calculation methods in the characteristics analysis of scholars is an appealing area worthy of further exploration. However, our preliminary results are limited in the biomedical field and their applicability in other areas needs further verification. With the development of large-scale scholarly datasets aiming to support various scientific disciplines, we will go forward to combine multidisciplinary data and investigate background diversity characteristics of scholars in the future work.

## REFERENCES

[1] Fisher A., Corbet S. A., Williams C. B, 1943. The relation between the number of species and the number of individuals in a random sample of an animal population, *The Journal of Animal Ecology*, 12 (1) :42-58.

[2] Ismael Rafols, Martin Meyer, 2010. Diversity and network coherence as indicators of interdisciplinarity: case studies in bionanoscience. *Scientometrics*, 82 (2): 263–287. DOI: 10.1007/s11192-009-0041-y.

[3] Jian Xu, Sunkyu Kim, Min Song, et al., 2020. Building a PubMed knowledge graph, *Scientific Data,* 7(1): 1-15. DOI:10.1038/s41597-020-0543-2.

[4] Pielou, E.C., 1966. Shannon's formula as a measure of specific diversity: Its use and misuse. *American Naturalist*, 100 (914): 463-465.

[5] Simpson, E. H., 1997. Measurement of diversity. *Journal of Cardiothoracic and Vascular Anesthesia*, 11 (6): 812-812. DOI: 10.1136/thx.27.2.261.

[6] Andy Stirling. 2007. A general framework for analyzing diversity in science, technology and society. *Journal of The Royal Society Interface*, 4 (15): 707-719.

[7] 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.

[8] Leidesdorff, 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.

[9] Peter S., Dodds, J. R., Minot, M. V., Arnold, et al., 2020. Allotaxonometry and rank-turbulence divergence: a universal instrument for comparing complex systems. *arXiv:2002.09770*. Retrieved from https://arxiv.org/abs/2002.09770.